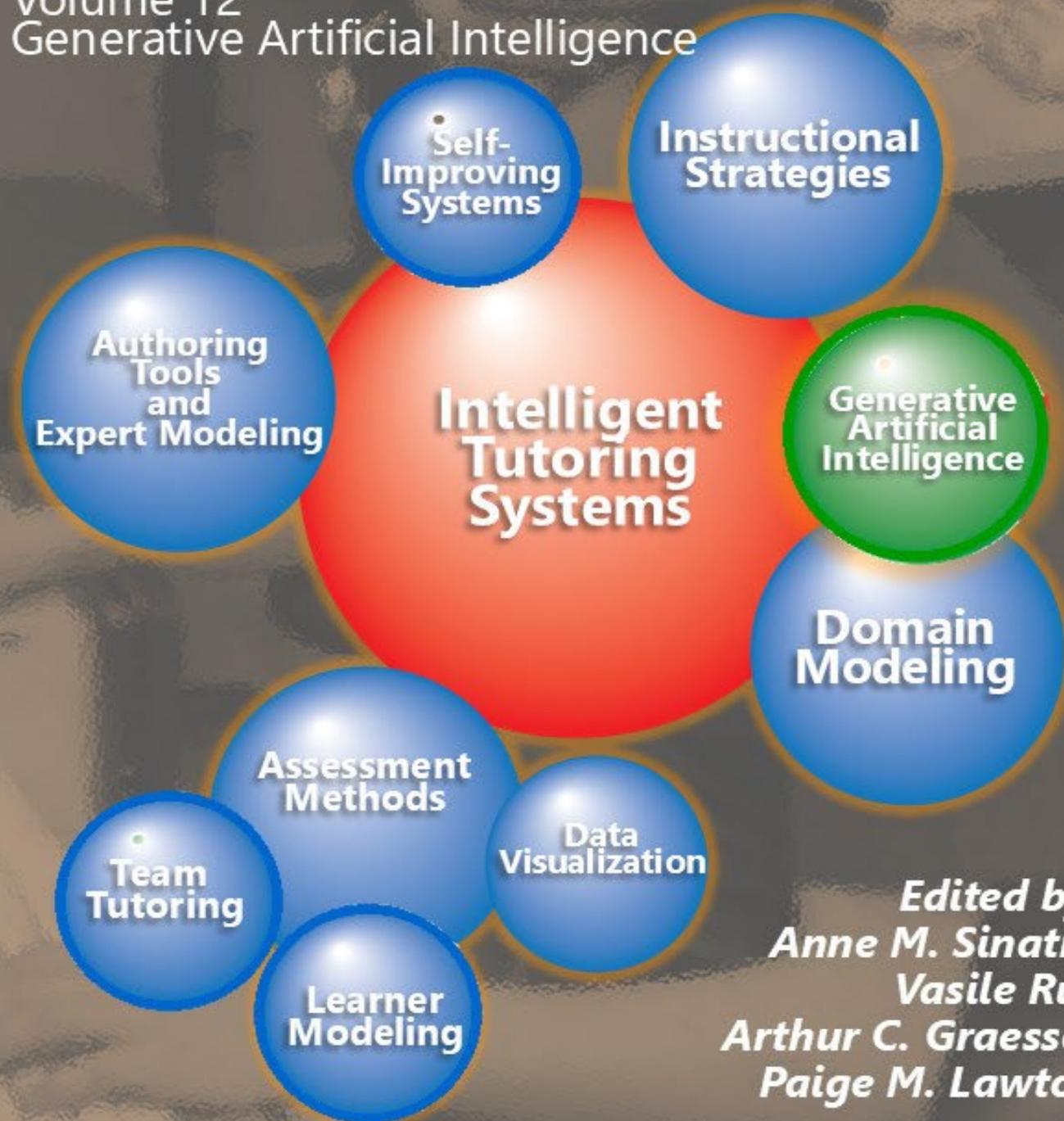


Design Recommendations for Intelligent Tutoring Systems

Volume 12
Generative Artificial Intelligence



Edited by:
Anne M. Sinatra
Vasile Rus
Arthur C. Graesser
Paige M. Lawton

A Book in the Adaptive Tutoring Series

Design Recommendations for Intelligent Tutoring Systems

Volume 12

Generative Artificial Intelligence

*Edited by:
Anne M. Sinatra
Vasile Rus
Arthur C. Graesser
Paige M. Lawton*

A Book in the Adaptive Tutoring Series

Copyright © 2026 by the US Army Combat Capabilities Development Command – Soldier Center

Copyright not claimed on material written by an employee of the US Government.

All rights reserved.

No part of this book may be reproduced in any manner, print or electronic, without written permission of the copyright holder.

The views expressed herein are those of the authors and do not necessarily reflect the views of the US Army Combat Capabilities Development Command - Soldier Center.

Use of trade names or names of commercial sources is for information only and does not imply endorsement by the US Army Combat Capabilities Development Command - Soldier Center.

This publication is intended to provide accurate information regarding the subject matter addressed herein. The information in this publication is subject to change at any time without notice. The US Army Combat Capabilities Development Command - Soldier Center, nor the authors of the publication, makes any guarantees or warranties concerning the information contained herein.

Printed in the United States of America
First Printing, March 2026

*US Army Combat Capabilities Development Command - Soldier Center
Simulation and Training Technology Center
Orlando, Florida*

International Standard Book Number:
978-0-9977258-7-2

Special thanks to Kyra Kurapati Williams, The University of Memphis, for her efforts in coordinating the workshop that led to this volume.

Dedicated to current and future scientists and developers of adaptive learning technologies

APPROVED FOR PUBLIC RELEASE

CONTENTS

Introduction to Generative AI and Intelligent Tutoring Systems	5
<i>Anne M. Sinatra, Vasile Rus, Arthur C. Graesser, and Paige M. Lawton, Eds.</i>	
Section I – Conversation and Natural Language Interactions	9
<i>Vasile Rus, Paige M. Lawton, and Anne M. Sinatra, Eds.</i>	
Chapter 1 – Exploring the use of Generative AI in Conversation-Based Assessment	11
<i>Diego Zapata-Rivera, Carol Forsyth, Liang Zhang, and Arthur C. Graesser</i>	
Chapter 2 – On The Role of Generative AI in Conversational Intelligent Tutoring Systems	25
<i>Vasile Rus, Mihai Lintean, and Anne M. Sinatra</i>	
Chapter 3 – Large Language Models and Intelligent Tutoring Systems: Conflicting Paradigms and Possible Solutions	37
<i>Punya Mishra, Danielle S. McNamara, Gregory Goodwin, and Diego Zapata-Rivera</i>	
Chapter 4 – The Use of Large Language Models in an Intelligent Text Framework	47
<i>Scott Crossley, Joon Suh Choi, Langdon Holmes, Wesley Morris, Vasile Rus, and Mihai Lintean</i>	

Chapter 5 – Using Generative AI to Improve Student’s Writing 55
William Swartout, Paul Deane, and Benjamin Nye

Section II – Scenario-Based Generative AI and Concrete Examples 73
Anne M. Sinatra, Vasile Rus, and Arthur C. Graesser, Eds.

Chapter 6 – Applying Generative AI Methods to Simulation-Based Adaptive Military Training 75
Robert A. Sottolare

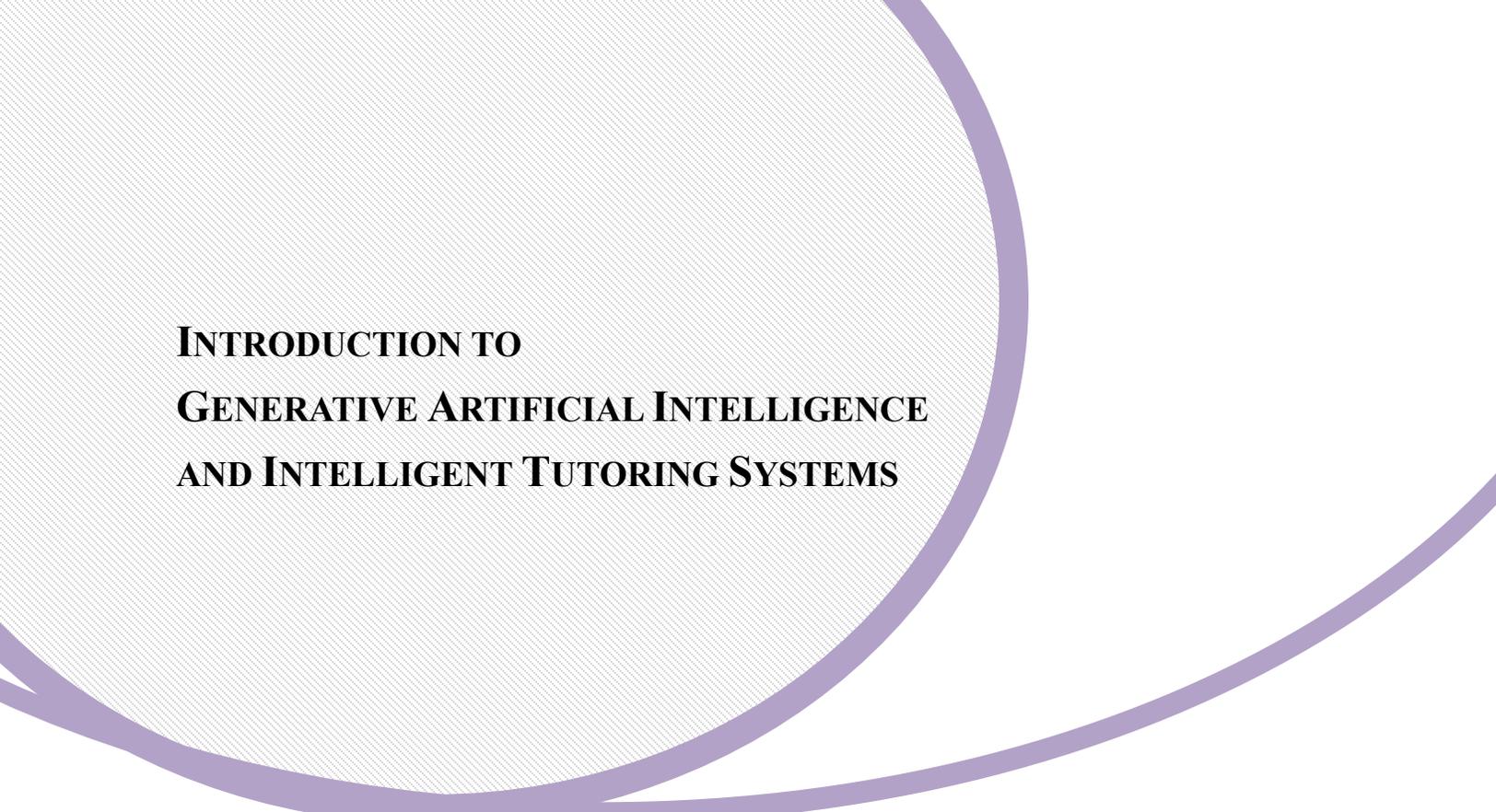
Chapter 7 - Adaptive Training Scenario Generation and Assessment with Generative AI 95
Andy Smith, Randall Spain, Wookhee Min, Bradford Mott, and James C. Lester

Chapter 8 – Generative AI for Assessing and Supporting Complex Performances in Teams 105
Peter W. Foltz, and Randall Spain

Chapter 9 – Generative AI for Concept Learning: Evolution of the Mobile Fact and Concept Training System 115
Andrew M. Olney

Chapter 10 – Applications for Generative AI in Intelligent Tutoring Systems in Army Education 127
Rebecca L. Robinson, Erica D. Kessler, Randy J. Brou, Robert A. Sottolare, and Diego Zapata-Rivera

Section III– Opportunities and Implications of Generative AI for Training Environments	137
<i>Arthur C. Graesser, Paige M. Lawton, and Anne M. Sinatra, Eds.</i>	
Chapter 11 – Generative AI for Instructional Systems Design Scalability, Efficiency, and Quality Control	139
<i>Ramona Shires, Robert McCormack, Shawn Weil, and Gregory Goodwin</i>	
Chapter 12 –Arizona State University CreateAI Platform	151
<i>Ishrat Ahmed, Paul Alvarado, Siddharth Jain, Tracy Arner, Elizabeth Reilley, and Danielle S. McNamara</i>	
Chapter 13 – Using Generative AI to Guide Experiential Learning	167
<i>Lisa N. Townsend, Paige M. Lawton, Robert A. Sottolare, Randall Spain, Richard Landers, Benjamin Goldberg, and Andy Smith</i>	
Chapter 14 - Opportunities for Leaveraging Generative AI for Creating Remediation Content in GIFT	177
<i>Anne M. Sinatra, Paige M. Lawton, and Randall Spain</i>	
Chapter 15 – Special Considerations and Next Steps for AI in Training Across the Services, Industry, and Academia	187
<i>Lisa N. Townsend, Danielle S. McNamara, Robert A. Sottolare, Benjamin Goldberg, and Wendi Van Buskirk</i>	
Biographies	217



**INTRODUCTION TO
GENERATIVE ARTIFICIAL INTELLIGENCE
AND INTELLIGENT TUTORING SYSTEMS**

***Anne M. Sinatra¹, Vasile Rus², Arthur C. Graesser²,
and Paige M. Lawton¹, Eds.***

*US Army Combat Capabilities Development Command (DEVCOM)
– Soldier Center¹*

The University of Memphis Institute for Intelligent Systems²

The *Design Recommendations for Intelligent Tutoring Systems* series has covered many different topics over the past thirteen years. Those topics have ranged from general components of intelligent tutoring systems (ITSs) (Learner Modeling, Instructional Management, Authoring Tools, Domain Modeling) to advanced elements (Assessment Methods, Team Tutoring, Self-Improving Systems, Data Visualization, Competency Based-Scenario Design) to overview discussions (SWOT [Strengths, Weaknesses, Opportunities and Threats] Analysis of ITSs) and applications (Professional Career Education). Each book and chapter additionally touches on the implications of the research for the Generalized Intelligent Framework for Tutoring (GIFT) software (Sottolare et al., 2012; Sottolare et al., 2017; Goldberg & Sinatra, 2023).

Each book in the *Design Recommendations for Intelligent Tutoring Systems* series has been associated with an Expert Workshop on the same topic. All of the workshops have been part of cooperative agreements between US Army Combat Capabilities Development Command (DEVCOM) Soldier Center and University of Memphis, with the current workshop being associated with cooperative agreement W912CG-24-2-0001. One of the goals of the expert workshops is to learn more about ITS capabilities that are being developed, and how these approaches/lessons learned could enhance the GIFT software (GIFT is freely available at <https://www.GIFTtutoring.org>). Invited experts in industry, academia, and government discuss the expert workshop topic, their applicable work, and suggestions for improving GIFT in a traditionally two day event. Both the University of Memphis and GIFT Teams participate in the workshop, help to guide discussion, and ask questions that will provide insight into current challenges in GIFT.

The expert workshop on the topic of Generative Artificial Intelligence (AI) in Intelligent Tutoring Systems, which is associated with this current book, was held in person in Orlando, Florida on September 24th and 25th, 2024, and included presentations about the applications of Generative Artificial Intelligence (AI) for ITSs. The goal of the workshop was to seek reports of early successes and challenges of using AI and Machine Learning (including Generative AI and large language models) to address various tasks related to our understanding of human learning and the development of ITSs. The University of Memphis team that participated in the workshop included Arthur C. Graesser, Vasile Rus, and Kyra Kurapati Williams. The US Army DEVCOM Soldier Center team who participated in the workshop included Benjamin Goldberg, Gregory Goodwin, Paige M. Lawton, Anne M. Sinatra, Randall Spain, and Lisa N. Townsend.

Sections of the Book

This book is organized into three sections covering aspects of Generative AI applications for ITSs:

- I. Conversation and Natural Language Interactions
- II. Scenario Based and Concrete Examples
- III. Opportunities and Implications for Generative AI for Training Environments

References

- Goldberg, B., & Sinatra, A.M. (2023). Generalized Intelligent Framework for Tutoring (GIFT) SWOT analysis. In Sinatra, A.M., Graesser, A.C., Hu, X., Goodwin, G., & Rus, V. (Eds.), *Design Recommendations for Intelligent Tutoring Systems, Volume 10: Strengths, Weaknesses, Opportunities and Threats (SWOT) Analysis of Intelligent Tutoring Systems*, pp. 9-26. US Army DEVCOM Soldier Center.
- Sottolare, R.A., Brawner, K.W., Goldberg, B.S. & Holden, H.K. (2012). *The Generalized Intelligent Framework for Tutoring (GIFT)*. Orlando, FL: U.S. Army Research Laboratory Human Research & Engineering Directorate (ARL-HRED).

Sottolare, R., Brawner, K., Sinatra, A. & Johnston, J. (2017). An Updated Concept for a Generalized Intelligent Framework for Tutoring (GIFT). Orlando, FL: *US Army Research Laboratory*. May 2017.

Acknowledgements

Research was sponsored by the US Army Combat Capabilities Development Command (DEVCOM) – Soldier Center (SC) and was accomplished under Cooperative Agreement Number W912CG-24-2-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the US Army DEVCOM SC or the US Government.



SECTION I – CONVERSATION AND NATURAL LANGUAGE INTERACTIONS

Vasile Rus¹, Paige M. Lawton², and Anne M. Sinatra², Eds.

*The University of Memphis Institute for Intelligent Systems¹
US Army Combat Capabilities Development Command (DEVCOM)
– Soldier Center²*

CHAPTER 1 – EXPLORING THE USE OF GENERATIVE AI IN CONVERSATION-BASED ASSESSMENT

Diego Zapata-Rivera¹, Carol Forsyth¹, Liang Zhang², and Arthur C. Graesser²
ETS Research Institute¹; The University of Memphis²

Introduction

Advances in Artificial Intelligence (AI) are facilitating the transformation and development of dialogue-based adaptive systems in educational settings. For example, researchers are exploring the use of Generative AI, specifically Large Language Models (LLMs), to generate tutoring feedback (e.g., hints, prompts) (Pardos & Bhandari, 2023) and to facilitate dialogue-based tutoring between students and an artificial agent (Hu, 2023). Foundational LLMs by themselves may not necessarily produce interactions that are appropriate in educational settings. Systems that build on top of LLMs (e.g., by improving their reasoning explainability and pedagogical capabilities using prompt engineering, fine-tuning techniques, semantic layers, or neuro symbolic approaches), have the potential to produce appropriate instructional and assessment materials and processes (e.g., instructional feedback).

We have explored the creation of Conversation-Based Assessments (CBAs), conversations between a human student and artificial agent(s) designed to assess skills in a natural context (Zapata-Rivera et al., 2023). These CBAs are well suited for assessing skills in a conversational context such as argumentation, scientific inquiry, language skills, and collaboration. Currently, we are exploring how LLMs and Evidence-Centered Design (ECD) (Mislevy et al., 2003) can be leveraged to create CBAs. Also, we have designed a multi-agent architecture that integrates several modules that leverage LLM agents in collaborative mechanisms for implementing subtasks, such as: administering tasks, tracing conversations, capturing learning knowledge, skills, and other attributes (KSAs) by analyzing learners' responses, and providing feedback with potential follow-up questions, prompts, and hints to further uncover student understanding capability. The multi-agent architecture includes both student-facing (i.e., agents that interact directly with the student such as a student agent or a scientist agent) and non-student facing agents (i.e., software agents that work on the background) with functionalities that complement foundational LLMs. Specifically, two *student-facing agents*, a *coordination agent*, a *formative assessor*, and a *summative assessor* provide additional support in areas such as pedagogy, bias and fairness, unit testing, evidence aggregation, and learner modeling.

In this chapter we will describe our prompt engineering approach, the multi-agent system, and how this functionality can support instruction and assessment. We present our current work that compares LLM CBAs to previously designed CBAs that use other types of natural language processing algorithms for semantic analyses (i.e., Regular Expressions, RegEx, Jurafsky & Martin, 2008; Latent Semantic Analysis LSA, Landaeur et al., 2007) and dialogue management tools (ASATA - AutoTutor Script Authoring Tool for Assessment, Nye et al., 2014). A comparison is made between two systems (with versus without LLMs) that potentially differ on the linguistic/discourse categories of the responses of artificial agents (i.e. tutor) to student input that is classified by the different algorithms (Forsyth et al., 2024). In addition, we discuss potential areas of future research. These areas include exploring the amount of feedback needed to create engaging conversations for assessment purposes, the quality of evidence collected from different types of dialogue moves, approaches for dealing with new versions of LLMs when they become available, a methodology to improve LLM understanding of English dialects, and the use of English accessibility features for multilingual learners. Finally, we speculate on how our work can inform design decisions

regarding new versions of GIFT (Sottolare et al., 2017; Goldberg & Sinatra, 2023) and STEEL-R (Goldberg et al., 2021) that leverage the capabilities of LLMs.

Conversation-based Assessments

Conversation-based assessments (CBAs) involve students interacting with artificial agents that provide students with opportunities to elaborate on their responses and explain their decisions. CBAs build on advances in dialogue systems research, artificial agents in education, and immersive learning environments (e.g., Clarke-Midura, Code, Dede, Mayrath & Zap, 2011; Graesser et al., 2001; Millis et al. 2011, Johnson & Lester, 2016; Lester et al., 2012).

In previous work, CBAs engage students in computer-based dialogues aimed at gathering evidence about their Knowledge, Skills, and other Attributes (KSAs). For example, CBAs assess KSAs in which conversational dialogue is a natural approach to collect evidence. These skills include argumentation, science inquiry, critical thinking, English language, and collaborative problem solving (So et al., 2015; Song et al., 2014; Zapata-Rivera, et al. 2014).

CBAs can provide students with multiple opportunities to elaborate on their responses (e.g., by rephrasing the question when questions are not clear and asking for additional information when students provide an incomplete answer). CBAs are usually used in conjunction with other types of tasks in technology-rich environments such as interactive scenarios and simulations (Zapata-Rivera et al., 2023).

Our prior work made use of ECD principles and the AutoTutor framework (Graesser, et al., 2001) to create conversations via ASATA (AutoTutor Script Authoring Tool for Assessment). ASATA uses Latent Semantic Analysis (LSA) and regular expressions (RegEx) to assess how well student responses semantically match expected answers, to categorize students' responses into speech acts, and determine how the system reacts to each type of response. We created conversation diagrams for implementation in the dialogue authoring tool (ASATA) to support collaborative development of CBAs by an interdisciplinary group of experts in assessment, dialogue, learning, human-computer interaction, and artificial intelligence in education (Zapata-Rivera et al., 2015). We are currently exploring the use of LLMs to implement CBAs with the hope that this landmark technology will increase efficiency in the generation of conversations and their quality. The next sections describe several aspects of our design and evaluation approach.

ECD Approach to Designing and Evaluating Conversations using LLMs

Our conversation design process starts with a clear definition of the construct(s) and the behaviors/interactions needed to support claims about students' mastery of them. ECD offers a principled methodology for assessment design. ECD supports the development of clear assessment arguments showing how tasks/conversations provide opportunities for students to demonstrate their KSAs and how the evidence collected is aggregated to support assessment claims. By following ECD principles, we support assessment validity and fairness. In the case of CBAs, we apply ECD principles to processes such as identifying the construct(s), the types of tasks/scenarios, structuring, testing and refining prompts internally, aggregating evidence and producing scores and reports. The process involves codesign activities with teachers and experts and iterative evaluation with users (see Figure 1).

LLMs have the potential for reducing the complexity of CBA design and implementation. Current versions of LLMs are already pre-trained on vast datasets, enabling developers to efficiently fine-tune them for specific tasks without extensive additional training. LLMs' offers sophisticated natural language processing (NLP) technologies that can simplify the creation of CBAs. However, foundational LLMs need to be augmented in various ways to produce appropriate conversations for learning and assessment

purposes. For example, their reasoning, explainability, and pedagogical capabilities could be improved by using prompt engineering, fine-tuning techniques, and semantic layers or neuro symbolic approaches (Moore et al., 2023; Pinto et al., 2024). Our approach leverages a combination of ECD, prompt engineering, and the implementation of a multi agent-architecture that includes a student model (Forsyth et al., 2024; Zapata-Rivera et al., 2024). In addition, we make use of Retrieval Augmented Generation (RAG) using relevant assessment design documentation such as Student Model, Evidence Model and Task Model Documentation (Mislevy et al., 2003) and Assessment Design Patterns (Mislevy & Riconscente, 2005). Assessment Rubrics, Conversation Diagrams (Zapata-Rivera et al., 2015), and Toulmin Diagrams (Toulmin, 1958) are used to ensure that the results of the LLM produces CBAs and evidence that are aligned with ECD. The next two sections introduce our prompt engineering approach and our multi-agent architecture.

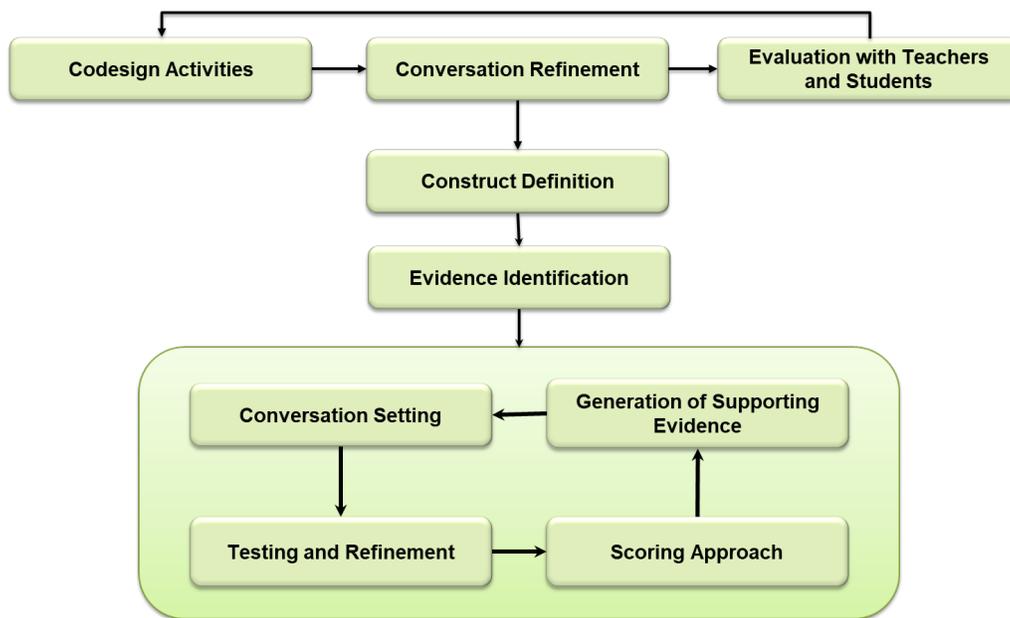


Figure 1. Design Architecture for LLM, ECD-based CBAs. Adapted from “Designing and Evaluating Evidence-Centered Design based Conversations for Assessment with LLMs”, by Zapata-Rivera, D. et al, 2024, Joint Proceedings of the Human-Centric eXplainable AI in Education and the Leveraging Large Language Models for Next Generation Educational Technologies Workshops (HEXED-L3MNGET 2024).CEUR-WS.org, online [https:// ceur-ws.org/Vol-3840/L3MNGET24_paper3.pdf](https://ceur-ws.org/Vol-3840/L3MNGET24_paper3.pdf). © 2024 by *Educational Testing Service. All rights reserved. Reprinted with permission.*

Prompt Engineering Approach

Prompt engineering involves developing principled instructions for an LLM in basic English (the prompt) with the goal of increasing the probability of the desired output. Prompt engineering is similar to computer programming in the sense that the designer is providing instructions to an algorithm (in this case, the LLM) to produce a result. It differs from computer programming in that it is less constrained and hence the desired output is not guaranteed. Still, it may be possible to provide guidelines for the development of effective prompts.

Developing effective prompts is surprisingly harder than expected. It requires an iterative approach to refining a prompt with a variety of strategies that take into account the order of conversational actions and necessary repetitions of instructions. Some of these strategies are general and have been documented in the research literature while others are specific to the goal of generating CBA tasks with ECD elements.

Specifically, we created 5 components to ensure the LLM remains on the domain topic and elicits evidence of the student’s KSAs in alignment with ECD. These 5 components include (1) Introduction and Domain, (2) Conversational Schema, (3) Modeling Behavior of Persona, (4) More Description and Restating and (5) What to Do with the Evidence.

Introduction and Domain

The goal of this component is to explain to the LLM its role as the teacher or fellow student peer and the role of the human student (or user) while considering the human student’s existing knowledge. We also set the stage to follow ECD within the conversations by explaining the domain and specific topic at the beginning of the prompt. Specifically in the introduction, we use a pattern referred to as the persona pattern (White et al., 2023) to explain to the LLM its role in relation to the student, such as “you are a knowledgeable friend of mine (the student)” and “your name is X.” It is necessary, given this role or persona, that we define the aforementioned domain knowledge to set the stage for the conversation. An example domain and topic could be science inquiry skills within the topic of predicting volcanic eruptions. As seen in Table 1, this element of the prompt produces discourse moves from the LLM such as “Hi, I’m John, your friendly and funny study partner. I’m here to help you improve your science inquiry skills in the context of a volcano eruption. Don’t worry, I won’t make you feel like you’re walking on lava 😊.” After introducing the domain and topic, we then define the role of the student that the LLM is chatting with by explaining the KSAs that the student already possesses. This is referred to as the “audience pattern” by White et. al. (2023). An example of this pattern within the given context is, “you can assume I know the basic information about volcanoes.” In table 1, you can see how this affects the conversation when the LLM states “You know some basic concepts about volcanoes, right?”. Constraining the conversation to this topic is extremely important because the LLM has access to such a broad range of information in its training set, much of which is irrelevant to the specific aspects of the domain and topic. Therefore, by providing more information about the student’s knowledge, we are able to begin constraining the LLM to gathering evidence from the student on a specific domain. However, this task still requires iterative refinement and a clear list of ancillary topics the LLM should not discuss by mentioning information such as “do not ask me (the student) information I’m already aware of.”

Table 1. Sample Conversation with prompt engineering without Retrieval Augmented Generation (RAG)

© 2024 by Educational Testing Service. All rights reserved. Reprinted with permission.

TUTOR	Hi, I’m John, your friendly and funny study partner. I’m here to help you improve your science inquiry skills in the context of a volcano eruption. Don’t worry, I won’t make you feel like you’re walking on lava. 😊 Let’s start with a warm-up question: What are some signs that a volcano is about to erupt?
STUDENT	I don't know [metacognitive response]
TUTOR	Come on, you can tell me something. You know some basic concepts about volcanoes, right? Think about what happens inside a volcano before it erupts. What are some physical or chemical changes that might indicate an increase in pressure or temperature?
STUDENT	seismic activity changes [partial answer]
TUTOR	That’s right! Seismic activity changes are one of the signs that a volcano is about to erupt. Can you explain why seismic activity changes indicate an impending eruption?
STUDENT	its at an increased rate [partial answer]
TUTOR	Yes, that’s part of the reason. Seismic activity changes are caused by the movement of magma and gas inside the volcano. When the pressure or temperature increases, the magma and gas try to escape through cracks and fissures in the earth’s crust. This creates vibrations that can be detected by seismometers. An increased rate of seismic activity means that the magma and gas are moving faster and more violently, which can indicate an impending eruption. Can you name another sign that a volcano is about to erupt?
STUDENT	Please repeat the question [metacommunicative]
TUTOR	Sure, I’ll repeat the question. Can you name another sign that a volcano is about to erupt?

STUDENT	Dog [Irrelevant]
TUTOR	I'm sorry, but that's not a sign that a volcano is about to erupt. A dog is an animal, not a physical or chemical change. Please try to stay focused on the topic. We are practicing science inquiry skills in the context of a volcano eruption. Can you name another sign that a volcano is about to erupt?

Conversational Schema

The conversational schema is the second aspect of the prompt which provides the rules of conversation based on pedagogical aspects as well as assessment principles. Basically, the schema tells the LLM how to proceed within a conversation which is necessary to have a pedagogically-sound CBA that gathers evidence about the student. In the context of a formative assessment, we leverage the Socratic method for the teaching discourse moves with a similar prompt that has been formulated by Hu et al., (2023). For example, we tell the LLM to not ask questions that require yes or no answers as this would make it more difficult to get information from the student about their knowledge of the topic. This is illustrated in Table 1 when the LLM asks the question “What are some signs that a volcano is about to erupt?” which clearly requires a detailed answer from the student. We also tell it to require the student to provide explanations for numerical answers as LLMs (and all NLP) are notoriously bad at math. Furthermore, we align with the prior research of (Graesser et al., 2001; Graesser et al., 1995) in telling the LLM to remain alert for misconceptions and address targeted questions with definitive answers. After explaining this to the LLM, we then must once again add in more negations and rules gained from iterative refinement to constrain the discourse moves. For example, we tell the LLM to keep its turns short (100 words or less) to avoid lengthy explanations.

Modeling Behavior of Persona

In this component of the prompt, it is necessary to provide the LLM with information on how to model the behavior of the persona of a knowledgeable friend helping the student to ensure that the artificial agent remains on the domain and continues to teach while eliciting evidence of the student’s KSAs. Negotiations are integral to accomplishing this goal as the persona has been described in the Introduction and Domain. However, as previously mentioned, the LLM has access to a large corpus and therefore much of the attempts to model behavior largely rely on constraining the LLM by telling it what not to do. For example, we must remind the LLM to not tell the students about their teaching style. Furthermore, the accurate classification of the student’s answers is imperative to ensure appropriate responses from the LLM’s persona. Therefore, we rely on categorizing speech acts of students that are identified as common in previous work of Graesser and Person (1994). For example, if a student provides a common answer of “I don’t know”, the LLM is instructed to respond with “Come on, tell me something”. If the student’s answer is incomplete, then the LLM is instructed to provide clues but not provide the answer. The LLM is also instructed to correct misunderstandings either directly or in an indirect way with a question to nudge the student to have a sudden realization and self-discover their error (which actually is the essence of the Socratic method). Moreover, the LLM also accounts for the common speech act of students saying “please repeat” by repeating the previous answer. In this fashion, the LLM has responses prepared for common speech acts of students. To ensure that these speech acts are modeled correctly, the LLM is given example responses tagged as correct, incorrect, metacognitive (“I don’t know”), metacommunicative (“please repeat”), incomplete answers, misconceptions and other uncategorized answers allowing for the LLM to have information about the response and a production rule to follow for the correct feedback based on this response. In Table 1, one can see how these various speech acts are handled within a sample conversation.

More Description and Restating

LLMs have an issue of forgetting aspects of a prompt and must be reminded. The main goal of step 4 is to provide additional information about the expected interaction with the student. This information includes guidance based on teaching and learning principles. The current version of the prompt implements the

following process:(a) You pose a question and then wait for my answer. Humor is always welcomed; (b) You assess my understanding by examining the correctness and amount of evidence provided towards the topic selected; (c) Based on my answer, you adjust your subsequent questions; and (d) When enough evidence is provided, you select a different topic within the same domain area and repeat the process. This sequence of conversational moves can be seen in Table 1 as well. This approach is an augmented version of Hu et al. (2023)'s Socratic method of teaching. We must also remind the LLM of its persona to ensure that the conversation remains appropriate and on domain. For example, we prompt the LLM to "Remember, your name is now John, you are my most reliable study partner, although you sometimes make mistakes, you are much better than me, so your patient tutoring is very helpful to me. I really like your sense of humor!". Note that in the restating of aspects of the prompt, there are often additional instructions such as "I only understand English". Once again, these types of constraints are discovered through iterative refinement and placed strategically to avoid repetition when unnecessary as there is a character limit to the prompts that the LLMs can process.

What to Do with Evidence

Finally, we must instruct the LLM on what to do with the evidence produced from the conversations. This aspect is extremely important as we want to follow the guidelines of ECD. As such, we instruct the LLM to produce a Toulmin Diagram of the discourse move, the standards it aligns with, and the reasoning for the alignment with these standards (Zapata-Rivera et al., 2024). In the multi-agent system, we further aid the LLM to accomplish this goal by providing a specific rubric designed by experts. Furthermore, we do ask for a final score but at this stage have little confidence in the LLM provided score and rather defer to analysis of the Toulmin diagram itself.

Multi-Agent LLM Architecture for CBA

Multi-agent systems have been studied and applied in multiple fields (Dorri et al., 2018), including intelligent tutoring systems (Šarić-Grgić et al., 2018) and learner modeling (Zapata-Rivera & Greer, 2004). Multi-agent systems can be used to solve complex problems in distributed and parallel computational contexts. Multi-agent issues, such as level of autonomy, communication strategies, and coordination, are some of the areas that have been investigated that have relevance to our project. Recently, multi-agent approaches have been proposed to the creation of Generative AI systems to enhance tutoring feedback, learning experiences, and increase engagement. For instance, Zhang, Zhang-Li et al. (2024) proposed SimClass, a system that leverages LLM-based agents to simulate real classroom roles, including teacher, teaching assistant, session controller, and classmates, to enhance learning engagement and effectively manage classroom dynamics. Yu et al. (2024) integrated multiple LLM-driven agents to perform tasks such as preparing tutoring resources (e.g., course materials, lecture notes, instructional videos), dynamically adjusting teaching processes based on student interactions, and creating personalized learning scenarios with roles like AI teaching assistants and customizable AI classmates, thereby enhancing both teaching and learning in online education. Zhang, Lin et al. (2024) developed a dialogue-based ITS, *Socratic Playground for Learning*, which has a multi-agent workflow to create learning scenarios, generate questions, and engage in dialogues with learners. Additional research in this area is ongoing, further exploring the potential of multi-agent systems in AI educational context. These advancements pave the way for leveraging LLM-based multi-agent systems to handle complex tasks in CBAs, such as generating feedback, delivering dialogue prompts, and assessing learners' KSAs to inform and enhance test-taking strategies.

The proposed multi-agent LLM architecture is designed for CBAs, specializing in the creation of a collaborative mechanism that leverages multiple LLM agents, each equipped with task-specific capabilities to facilitate automatic processes in specialized domains. As shown in Figure 2, the multi-agent workflow

for CBA integrates various components to facilitate the adaptive assessment of learners' KSAs. This architecture integrates multiple agents, including formative and summative assessors, a coordination agent, an expert agent (referred to as Dr. Garcia), a peer student agent (referred to as Art), and a learner model powered by an LLM to facilitate the tracking and evaluation of learner progress. It also features a shared memory component, along with additional functions and tools (e.g., RAG) that is customized or enhanced for specific task implementations (e.g., assessment for expert agent).

Formative Assessor

This agent is a critical component of the system because it evaluates the student's turn-level responses to ensure accurate categorizations of student contributions (e.g. correct, meta-cognitive, incomplete). The agent leverages an LLM and a structured rubric that is aligned with relevant standards (e.g., NGSS). The agent assesses the accuracy, relevance, and completeness of student inputs while considering the context of the conversation. Based on the assessment, it dynamically determines the most appropriate feedback and subsequent action, such as prompting Dr. Garcia to seek further elaboration or guiding Peer Art to redirect off-topic responses humorously. Assessment based feedback is passed to Dr. Garcia or the peer student agent to refine or adapt it based on other goals (e.g., pedagogical goals). This agent's mechanism ensures tailored, adaptive feedback that enhances learning while providing researchers with valuable insights into the assessment of student progress within dialogue-based scenarios.

Summative Assessor

This agent is responsible for evaluating the overall mastery of KSAs by analyzing the learner's responses in a structured dialogue and other available evidence. Using prompts tailored to the session's context, such as predicting thunderstorms or selecting appropriate weather stations, the summative assessor constructs Toulmin diagrams. These diagrams map claims, evidence, and warrants in the learner's discourse to assess how well their arguments align with relevant standards (e.g., NGSS). Additional information in AI generated Toulmin diagrams have potential for informing scoring, assessment documentation and instruction. The evaluation process involves leveraging predefined rubrics, which includes student goals, activities, leading questions, keywords, NGSS-aligned dimensions, and examples of "good" versus "partial" answers. This enables the summative assessor to provide a comprehensive summary of the learner's KSAs and generate scores based on rubrics developed by researchers and educators. This information is used to update the learner model. Currently, we are evaluating LLM generated scores with those produced by human experts.

Expert Agent (Dr. Garcia)

Dr. Garcia serves as the expert agent within the multi-agent system, embodying a knowledgeable and engaging mentor to guide students through the learning process. Using a Socratic questioning style, Dr. Garcia facilitates critical thinking and conceptual understanding by posing thought-provoking questions, assessing responses, and providing constructive feedback. With templates tailored to specific scenarios, such as predicting thunderstorms or selecting weather stations, Dr. Garcia dynamically adapts feedback based on the student's input and understanding. The agent leverages detailed domain-specific instructions, NGSS-aligned dimensions, and a structured four-step process to ensure that the student demonstrates mastery of the topic before progressing. Dr. Garcia also humorously engages with students, encourages detailed explanations, and ensures alignment with predefined educational goals. This agent's role is pivotal in fostering science inquiry skills, tracking conceptual progress, and enhancing the educational experience in a personalized and interactive manner. This agent can implement different pedagogical strategies.

Peer Student Agent (Art)

The peer student, Art, is designed to serve as a collaborative and supportive study partner for the learner, fostering engagement through humor and relatable interactions. Unlike the expert agent, Dr. Garcia, Art intentionally introduces minor mistakes to stimulate the human student's critical thinking and deeper reflection, while still offering guidance to keep the learner focused. Art employs Socratic questioning to encourage the learner to refine their reasoning and enhance their understanding of complex topics by using tailored prompts and scenarios, such as predicting thunderstorms or selecting weather stations. The role of the peer student agent can change depending on the pedagogical strategy implemented.

Coordination Agent

This agent that is powered by an LLM is responsible for orchestrating interactions between various agents in the learning environment. Serving as the central coordinator, the coordination agent enforces conversational rules, such as determining the number of turns before ending the conversation, deciding which agent (e.g., Dr. Garcia, Art, or the formative assessor) should respond at each step, and ensuring smooth transitions between roles. The LLM capabilities of the coordination agent enhance its ability to interpret context and guide decision-making dynamically. Additionally, the coordination agent incorporates ethical safeguards by detecting inappropriate language through predefined rules and databases, responding by redirecting conversations or involving human oversight when necessary. This integration of LLM-powered functionality ensures the system operates cohesively, balances interaction dynamics, and maintains a safe, and efficient learning environment.

Learner Model

The Learner Model is designed to monitor, evaluate, and model a learner's KSAs through dialog-based interactions and other available evidence, partially leveraging LLM-powered capabilities for learning tracking and knowledge tracing in dialogue scenarios (Neshaei et al., 2024; Scarlatos, Baker & Lan, 2025). It captures relevant learner data, including process logs, response data, and conversation history, to provide continuous tracking of the learner's progress and activities. Based on this data, it conducts assessments and predictions to evaluate the learner's current understanding and anticipate future learning needs. The learner model supports personalization by enabling precise feedback and facilitating meaningful interactions that address both cognitive and non-cognitive KSAs, ensuring a comprehensive and data-driven approach to learning.

Shared Memory

This component plays a pivotal role in managing both short-term and long-term contextual information. It bridges the limited context windows inherent in LLMs by retrieving and organizing relevant learning records, ensuring that the system maintains a cohesive understanding of the learner's journey. Shared memory storage is a critical feature of this architecture, enabling seamless information sharing across agents. It supports context retrieval for generating feedback, prompts, or tailored assessments, ensuring alignment with the learner's progress and historical interactions (evidence aggregation and integration of evidence mechanisms in the learning model). Additionally, the memory component aids in inferring patterns from learning records, enabling the generation of personalized questions and adaptive interventions. This integration ensures that the system provides a dynamic, data-driven, and collaborative environment for effective learning and assessment, fostering deeper understanding and engagement.

RAG Reinforced Assessment

RAG is utilized to augment the assessor's efficacy in evaluating the learner's progress while facilitating the generation of more personalized and adaptive feedback aimed at assessing the learner's KSAs. By integrating external knowledge sources covering domain-specific knowledge and leveraging components from ECD, such as assessment design documents, rubrics, design patterns, and conversation diagrams. RAG ensures that assessments are not only precise but also deeply aligned with educational standards and learning objectives. This process is fundamentally driven by the principles of effective prompt engineering. This structured approach ensures that evaluations are anchored in well-defined pedagogical and assessment frameworks.

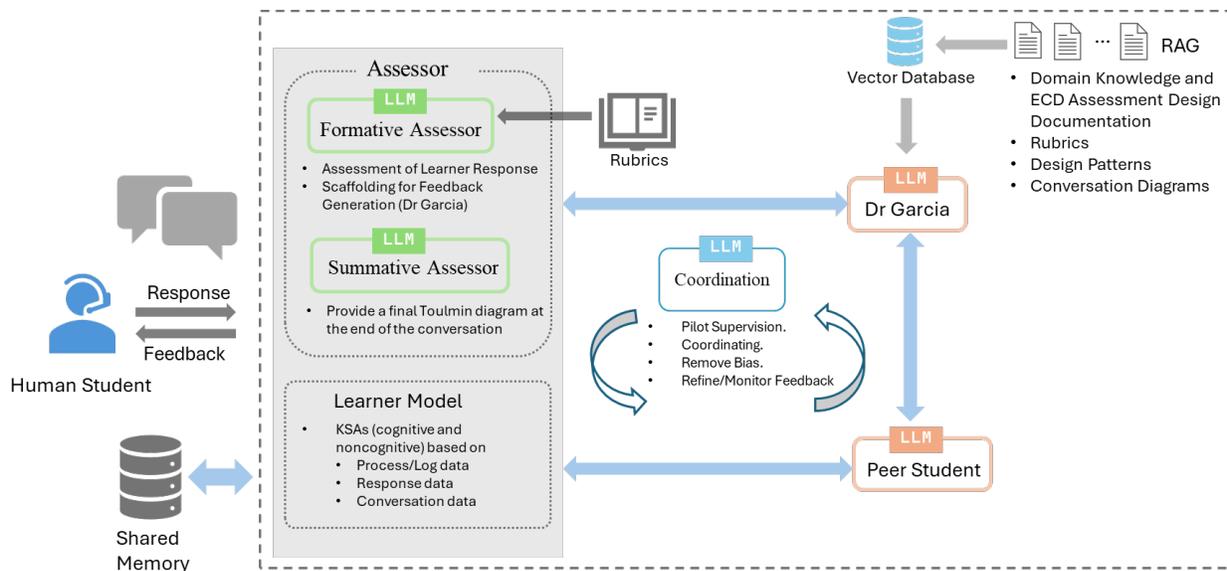


Figure 2. Multi-agent workflow for conversation-based assessment (CBA) framework. © 2024 by *Educational Testing Service*. All rights reserved. Reprinted with permission.

In Table 2, we present example LLM conversations in the context of volcano eruptions to illustrate how RAG enhances the tutor's capabilities, demonstrating its potential to enrich the functionality and effectiveness of our real-world multi-agent CBA system. Table 2 shows two sample conversations (without RAG and with RAG) that can take place in the volcano scenario. One features a RAG-augmented tutor agent leveraging documents that encompass scientific knowledge on volcanic eruptions, seismic activity, and seismometer deployment, principles of ECD for evidence-based assessment, triologue-based methodologies for evaluating inquiry skills, and game-like mechanics to enhance engagement and assessment. The second one excludes RAG. Each scenario consists of three dialogue turns between the tutor and the student, utilizing the GPT-4o model as the foundational model for agent interactions. Specifically, the RAG-enhanced tutor agent utilizes an external knowledge base and assessment information to deliver detailed explanations and address student uncertainties. Its reasoning process involves three steps: retrieving relevant knowledge and assessment information using RAG, generating an initial response based on the student's input and the retrieved context to promote deeper understanding, and refining the response to ensure alignment with assessment information (e.g. rubrics and other assessment documentation) for comprehensive and accurate feedback. The second scenario without RAG relies solely on the LLM inherent capabilities.

Expert evaluations of the sample conversations reveal that when RAG is applied, the responses generated are contextually constrained, relevant, and provide deeper insights aligned with the goals of the scenario.

In contrast, without RAG, the conversation lacks access to an external knowledge base and structured assessment, often resulting in responses that deviate from the scenario's objectives and lack the depth required for effective tutoring. For example, in the RAG-augmented scenario, when the student states that “more tremors of different types can be used to infer eruption alert levels for a volcano”, the tutor acknowledges this and deepens the conversation by asking about the different types of seismic activity and how they indicate various stages of volcanic activity, providing immediate and detailed explanations. In contrast, in the scenario without RAG, while the tutor also acknowledges the student's point by saying “That’s a good start, Peter”, it offers limited expansion and redirects the student to external resources for further information instead of providing in-depth explanations within the conversation.

Table 2. Sample Conversation with and without RAG

© 2024 by Educational Testing Service. All rights reserved. Reprinted with permission.

Role	Conversation Without RAG	Conversation With RAG
TUTOR	Hello Peter, it’s your trusty study partner John here, ready to shake things up with seismic activity! Let’s dive right in. Can you explain how seismic activity is related to volcanic eruptions?	Hello Peter, it’s your trusty study partner John here, ready to shake things up with seismic activity! Let’s dive right in. Can you explain how seismic activity is related to volcanic eruptions?
STUDENT	more tremors of different types can be used to infer eruption alert levels for a volcano	more tremors of different types can be used to infer eruption alert levels for a volcano
TUTOR	That's a good start, Peter. Tremors can indeed indicate the likelihood of an eruption. But can you explain how the different types of tremors, such as harmonic tremor and volcanic tectonic earthquakes, specifically contribute to predicting an eruption?	That's a good start, Peter! Seismic activity, such as tremors, can definitely indicate potential volcanic eruptions. However, let's dig a bit deeper. Can you explain the different types of seismic activity that are typically observed around volcanoes and how they might indicate different stages or levels of volcanic activity?
STUDENT	Low frequency shows magma moving, and when the chamber is about to open, there should be high-frequency tremors indicating rocks cracking.	Low frequency shows magma moving, and when the chamber is about to open, there should be high-frequency tremors indicating rocks cracking.
TUTOR	Okay, Peter, you're on the right track! Low-frequency tremors can indeed signal magma movement, and the high-frequency ones indicate rock fracturing. But let's dig deeper: how do scientists distinguish between these tremors and determine their sources? What tools or methods do they use to analyze these seismic signals?	Great job, Peter! You've identified two key types of seismic activity: low-frequency tremors linked to magma movement and high-frequency tremors indicating rock cracking as the magma forces its way up. Let's build on this. How might the patterns or frequency of these seismic activities change as an eruption becomes more imminent? What other seismic indicators might you expect to see or monitor as part of a comprehensive volcanic surveillance system?
STUDENT	I don't know	I don't know
TUTOR	No worries, Peter! To get a better understanding, you can start by looking into seismographs and seismometers, which are tools scientists use to record and analyze seismic waves. Additionally, you might find it helpful to explore how spectrograms can visually represent different types of tremors. Two great resources to get you started are the USGS Volcano Hazards Program website and the Incorporated Research Institutions for Seismology (IRIS) website. Once you've checked those out, let's discuss how these tools help distinguish between different seismic activities.	Peter, no worries! Let’s crack this open like a volcanic rock. Think about how seismicity evolves as magma moves toward the surface. What happens to the rate and intensity of tremors? Also, consider if there are other seismic indicators, like harmonic tremor or volcanic-tectonic earthquakes. How might these help in predicting an eruption?

Notably, this multi-agent approach allows for the definition and execution of *unit testing* that can be deployed when a new or a different version of an LLM becomes available. This facilitates keeping up with advances in Generative AI. For example, unit testing can be integrated as a component of the system to determine the changes and modifications needed due to the integration of new LLM versions. The multi-agent framework has been preliminarily applied in real-world scenarios, with testing currently underway. Further insights and advancements are expected to be explored in future research endeavors.

Initial Results

We have compared the types of CBA responses produced using RegExp/LSA with responses generated by LLM-based CBAs. Initial results included comparing the responses to common dialogue moves (Zapata-Rivera et al., 2024) and analysis of linguistic features of CBA responses (Forsyth et al. 2024). The LLM used was Microsoft Copilot GPT-4 (OpenAI, 2024). A set of responses were randomly selected from a dataset of 105 middle-school students who completed a version of the volcano scenario that makes use of RegExp/LSA and ASATA. The responses generated by the system were compared to simulated responses within the LLM-based CBA's. Results showed that interactions with LLM CBAs are more natural and require less NLP development effort (from weeks down to days). For example, when there are responses to irrelevant responses, the LLM CBA subtly brings the topic back to the domain by connecting the irrelevant response to the topic; in contrast, the RegExp/LSA version merely asked students to explain the relevance of their response to the topic, which was how human designers decided to handle it. Also, the LLM CBA responses included analogies and some humor. Open natural responses can potentially improve engagement, but may be challenging to score (Zapata-Rivera et al., 2024).

In a further investigation, we compared linguistic features of the pedagogical agent's responses between the two versions to discern key differences in these responses. To conduct this investigation, the responses were analyzed via TextEvaluator (Sheehan et al., 2014) which produces principal components that measure different language/discourse levels of text complexity in the common core standards, along with qualitative analysis. Results indicated that the LLM generated CBA created artificial agent discourse moves that were more complex by all measures. This was further exemplified by the LLM producing analogies, truly mixed initiative dialogue (allowing the students to ask for help) and motivational language. Implications suggest that the LLM goes into greater depth of the topics and therefore may be more useful in aiding students to gain a deeper understanding of the material. However, this hypothesis requires additional testing (Forsyth et al., 2024).

Challenges of using LLMs for CBA include (a) evaluating various approaches for implementing components of the multi-agent platform (e.g., evaluating the efficiency and accuracy of various approaches to implementing the coordination agent); (b) implementing various pedagogical strategies that align well with formative assessment purposes; (c) evaluating the use of LLMs for improving automated scoring of CBAs, evaluating the potential of using Toulmin diagrams for informing scoring, assessment documentation and teacher instruction; (d) integrating evidence from various types of tasks into the learner model and evaluating its accuracy; and (e) implementing and evaluating LLM-based CBAs for other constructs (e.g., collaboration, communication, and critical thinking).

Application to GIFT and STEEL-R

Our multi-agent architecture, including our approach to manage evidence identification and aggregation using LLMs, can inform the integration of LLMs in GIFT (the Generalized Intelligent Framework for Tutoring) and STEEL-R (Synthetic Training Environment Experiential Learning for Readiness). The use of LLMs in the production of conversations for assessment purposes has resulted in a reduction in time and effort. However, additional effort is required to ensure that conversations are appropriate. Agents in the multi-agent architecture can implement pedagogical strategies based on process data and learner model information (e.g., proficiency estimates, learner progression levels, engagement, persistence, interest levels). Similar results can be expected when integrating LLMs in the GIFT framework.

LLMs in CBAs can generate more natural conversations with longer interactions compared to previous approaches using RegExp and LSA. However, these conversations do not necessarily address all the important aspects of relevant constructs and may not always be pedagogically appropriate. For example,

the use of humor may result in inappropriate jokes which need to be monitored. However, an iterative prompt refinement approach and the multi-agent AI platform that we have described here can facilitate the process of producing conversations that are suitable for assessment and learning contexts. The use of platform agent components such as the formative and summative assessors, the expert and student agents, the coordination agent and the learner model can support the generation of these conversations. Also, techniques such as RAG have the potential for making the conversations more relevant based on curated content. Assigning scores based on conversational and other data is a challenging task that we are addressing by identifying evidence from CBAs, and aggregating it with evidence from other sources (e.g., responses to other tasks). We recommend implementing approaches in GIFT and STEEL-R to deal with fairness. Clear evidence paths that are implemented in STEEL-R can be extended to deal with some of these issues.

Transparency, and fairness are important aspects to consider when using LLMs in learning and assessment contexts. We are addressing these issues by customizing our prompts based on ECD principles and applying RAG, monitoring the responses generated by the LLM (e.g., the coordination agent), and employing human-in-the loop approaches. However, additional work needs to be conducted to detect and make modifications that handle these issues. Related work in neuro-symbolic approaches has the potential for improving transparency and explainability of Generative AI systems (Colelough & Regli, 2024; Zhang & Sheng, 2024; Zapata-Rivera & Arslan, 2024). However, it is beyond the scope of this chapter to address neuro-symbolic approaches that identify relationships between explainable symbolic models and neural models that defy explanations that are easy to understand.

The use of Generative AI brings opportunities and challenges for the development and evaluation of adaptive learning and assessment systems. We have described various approaches for enhancing foundational LLMs to support appropriate interactions for instruction and assessment purposes in conversation-based assessment and learning. Enhancing LLMs with ECD principles and relevant assessment design documentation (e.g., via RAG) can produce relevant and appropriate conversations for assessment purposes. The use of learner model information can play an important role in supporting LLM Agents' interactions. The use of LLMs can support learner modeling processes such as evidence identification and evidence aggregation processes. Both GIFT and STEEL-R can benefit by integrating Generative AI to enhance scenario design and assessment components.

Acknowledgements

This work was funded in part by the National Science Foundation and the Institute of Education Sciences under Grant #2229612. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the U.S. Department of Education.

References

- Clarke-Midura, J., Code, J., Dede, C., Mayrath, M., & Zap, N. (2011). Thinking outside the bubble: virtual performance assessments for measuring complex learning. In: Mayrath, M.C., Clarke-Midura, J., Robinson, D. (Eds.), *Technology-Based Assessments for 21st Century Skills: Theoretical and Practical Implications From Modern Research*. Information Age, Charlotte, NC, pp. 125–147
- Colelough, B. C., & Regli, W. (2024). Neuro-Symbolic AI in 2024: A Systematic Review. *Proceedings of the First International Workshop on Logical Foundations of Neuro-Symbolic AI (LNSAI 2024)*. CEUR-WS.org, online <https://ceur-ws.org/Vol-3819/paper3.pdf>
- Dorri, A., Kanhere, S. S., & Jurdak, R. (2018). Multi-agent systems: A survey. *Ieee Access*, 6, 28573-28593.

- Forsyth, C.M., Zapata-Rivera, D., Graf, A. & Jiang Y., (2024). Complex Conversations: LLMs vs. Knowledge Engineered Conversation-based Assessment. In *Proceedings of the 17th International Conference on Educational Data Mining*, pp 868-871. DOI: 10.5281/zenodo.12729976
- Goldberg, B. S., DeFalco, J. A., Hoffman, M., & Burmester, E. (2021). User feedback on a hybrid team tutoring strategy. In *Proceedings of the Challenges and Advances in Team Tutoring Workshop during the Artificial Intelligence in Education Conference 2021* (pp. 1–9). CEUR-ws.org, online, <https://ceur-ws.org/Vol-3096/paper1.pdf>
- Goldberg, B., & Sinatra, A. M. (2023). Generalized Intelligent Framework for Tutoring (GIFT) SWOT analysis. In *Design Recommendations for Intelligent Tutoring Systems: Volume 10-Strengths, Weaknesses, Opportunities and Threats (SWOT) Analysis of Intelligent Tutoring Systems* (pp. 9–26). US Army DEVCOM Soldier Center
- Graesser, A.C., & Person, N.K., (1994). Question asking during tutoring. *Am. Educ. Res. J.* 31, 104–137.
- Graesser, A. C., Person, N., Harter, D., & the Tutoring Research Group. (2001). Teaching tactics and dialog in AutoTutor. *International Journal of Artificial Intelligence in Education*, 12, 257–279.
- Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, 9(6), 495–522. <https://doi.org/10.1002/acp.2350090604>
- Hu. X. (2023). Empowering education with LLMs - the next-gen interface and content generation. [demo] Presented at *The Inter. Conference on Artificial Intelligence in Education*. (Tokyo, Japan, July 03-07, 2023). AIED 2023.
- Johnson, W.L., & Lester, J.C. (2016). Face-to-face interaction with pedagogical agents, twenty years later. *Int. J. Artif. Intell. Educ.* 26 (1), 25–36.
- Jurafsky, D., & Martin, J. (2008). *Speech and Language Processing*. Prentice Hall, Englewood, NJ.
- Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). Handbook of latent semantic analysis. Mahwah, NJ: Erlbaum.
- Mislevy, R.J., Steinberg, L.S., and Almond, R.G. 2003. On the structure of educational assessments. *Meas. Interdiscip. Res. Perspect.* 1, 3–62.
- Lester, J. C., Rowe, J. P., & Mott, B. W. (2012). Narrative-centered learning environments: A story-centric approach to educational games. In *Emerging technologies for the classroom: A learning sciences perspective* (pp. 223-237). New York, NY: Springer New York.
- Millis, K., Forsyth, C., Butler, H., Wallace, P., Graesser, A.C., & Halpern, D., (2011). Operation ARIES! A serious game for teaching scientific inquiry. In: Ma, M., Oikonomou, A., Lakhmi, J. (Eds.), *Serious Games and Edutainment Applications*. Springer-Verlag, London, pp. 169–196.
- Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessments. *Meas. Interdiscip. Res. Perspect.* 1, 3–62.
- Mislevy, R., & Riconscente, M. (2005). Evidence-centered Assessment Design: Layers, Structures, and Terminology (*PADI Technical Report 9*). Menlo Park, CA: SRI International.
- Moore, S., Tong, R., Liu, Z., Hu, X., Lu, Y., Joleen, L., Khosravi, H., Denny, P., Singh, A., Brooks, C., Stamper, J., & Cao, C. (2023). *AIED2023 Empowering Education with LLMs - the Next-Gen Interface and Content Generation*. AIED 2023. Tokyo, Japan.
- Neshaei, S. P., Davis, R. L., Hazimeh, A., Lazarevski, B., Dillenbourg, P., & Käser, T. (2024). Towards modeling learner performance with large language models. arXiv preprint arXiv:2403.14661, 2024.
- Nye, B.D., Graesser, A.C. & Hu, X. (2014). AutoTutor and Family: A Review of 17 Years of Natural Language Tutoring. *Int J Artif Intell Educ* 24, 427–469. <https://doi.org/10.1007/s40593-014-0029-5>
- Pardos, Z. A., & Bhandari, S. (2023). Learning gain differences between ChatGPT and human tutor generated algebra hints. Retrieved from *arXiv:2302.06871v1*
- Pinto. J.D., Worden, E., Botelho, A., Cohausz, L., Cohn, C., et al., (2024) Joint Proceedings of the Human-Centric eXplainable AI in Education and the Leveraging Large Language Models for Next Generation Educational Technologies Workshops (HEXED-L3MNGET 2024) co-located with 17th International Conference on Educational Data Mining (EDM 2024). Atlanta, Georgia, USA, July 14, 2024. <https://ceur-ws.org/Vol-3840/>
- Šarić-Grgić, I., Grubišić, A., Stankov, S., & Štula, M. (2019). An agent-based intelligent tutoring systems review. *International Journal of Learning Technology*, 14(2), 125–140.
- Scarlatos, A., Baker, R. S., & Lan, A. (2025). Exploring knowledge tracing in tutor-student dialogues using llms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*. 249–259.
- Sheehan, K. M., Kostin, I., Napolitano, D., & Flor, M. (2014). The Text Evaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *The Eleme. School J.*, 115, 184–209.

- So, Y., Zapata-Rivera, D., Cho, Y., Luce, C., & Battistini, L., (2015). Using dialogues to measure English language skills. In: García Laborda, J., Sampson, D.G., Hambleton, R.K., Guzman, E. (Eds.), *Journal of Educational Technology and Society. Special Issue: Technology Supported Assessment in Formal and Informal Learning*. pp. 21–32.
- Song, Y., Sparks, J., Brantley, W., Oliveri, M., & Zapata-Rivera, D. (2014). Designing Game Activities to Assess Students' Argumentation Skills. *Paper Presented at the Annual Meeting of the American Educational Research Association (AERA)*, Philadelphia.
- Sottolare, R. A., Brawner, K. W., Sinatra, A. M., & Johnston, J. H. (2017). An updated concept for a Generalized Intelligent Framework for Tutoring, R. A., Goldberg, B. S., Brawner, K. W., & Holden, H. K. (2012). A modular framework to support the authoring and assessment of adaptive computer-based tutoring systems (CBTS). In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference* (pp. 1-13).
- Toulmin, S. E. (1958). *The uses of argument*. Cambridge, United Kingdom: University Press.
- Yu, J., Zhang, Z., Zhang-li, D., Tu, S., Hao, Z., Li, R. M., Li, H., et al. (2024). From MOOC to MAIC: Reshaping online teaching and learning through LLM-driven agents. arXiv preprint, arXiv:2409.03512.
- Zapata-Rivera, D., & Arslan, B. (2024). Learner Modeling Interpretability and Explainability in Intelligent Adaptive Systems. In: Santoianni, F., Giannini, G., Ciasullo, A. (eds) *Mind, Body, and Digital Brains. Integrated Science*, vol 20. Springer, Champp. 95–109. https://doi.org/10.1007/978-3-031-58363-6_7
- Zapata-Rivera, J. D., & Greer, J. (2004). Inspectable Bayesian student modelling servers in multi-agent tutoring systems. *International Journal of Human Computer Studies*. 61(4), 535-563. [Online] Available: <http://dx.doi.org/10.1016/j.ijhcs.2003.12.017>
- Zapata-Rivera, D., Forsyth, C. M., Graf, A., & Jiang, Y. (2024). Designing and Evaluating Evidence-Centered Design based Conversations for Assessment with LLMs. *Joint Proceedings of the Human-Centric eXplainable AI in Education and the Leveraging Large Language Models for Next Generation Educational Technologies Workshops (HEXED-L3MNGET 2024)*. CEUR-WS.org, online https://ceur-ws.org/Vol-3840/L3MNGET24_paper3.pdf
- Zapata-Rivera, D., Jackson, T., & Katz, I.R. (2015) Authoring Conversation-based Assessment Scenarios. In R. A. Sottolare, A. C. Graesser, X. Hu, and K. Brawner (Eds.), *Design Recommendations for Intelligent Tutoring Systems Volume 3: Authoring Tools and Expert Modeling Techniques*. (pp. 169–178). U.S. Army Research Laboratory.
- Zapata-Rivera, D., Jackson, T., Liu, L., Bertling, M., Vezzu, M., & Katz, I.R., (2014). Science inquiry skills using dialogues. In *the 12th International Conference on Intelligent Tutoring Systems*. pp. 625–626.
- Zapata-Rivera, D., Sparks, J. R., Forsyth, C. M., & Lehman, B. (2023). Conversation-based assessment: Current findings and future work. In *International Encyclopedia of Education (Fourth Edition)*. R.J. Tierney, F. Rizvi, & K. Ercikan, Eds. Elsevier. 504–518. doi:10.1016/B978-0-12-818630-5.10063-6
- Zhang, L., Lin, J., Kuang, Z., Xu, S., & Hu, X. (2024). SPL: A Socratic Playground for Learning Powered by Large Language Models. *CEUR Workshop Proceedings. Educational Data Mining 2024 Workshop: Leveraging Large Language Models for Next Generation Educational Technologies*.
- Zhang, X., & Sheng, V. S. (2024). Neuro-Symbolic AI: Explainability, Challenges, and Future Trends. arXiv preprint arXiv:2411.04383.
- Zhang, Z., Zhang-Li, D., Yu, J., Gong, L., Zhou, J., Liu, Z., ... & Li, J. (2024). Simulating classroom education with llm-empowered agents. arXiv preprint arXiv:2406.19226.

CHAPTER 2 – ON THE ROLE OF GENERATIVE AI IN CONVERSATIONAL INTELLIGENT TUTORING SYSTEMS

Vasile Rus¹, Mihai Lintean², and Anne M. Sinatra³

The University of Memphis¹; EdClub, Inc.²; US Army Combat Capabilities Development Command (DEVCOM) - Soldier Center³

Introduction

This chapter discusses the role of Generative Artificial Intelligence (Generative AI) and more specifically the role of Large Language Models (LLMs) in conversational intelligent tutoring systems (ITSs). Generative AI is a set of recently developed AI architectures based on deep neural networks that can generate text and other forms of outputs, e.g., images, in response to user inputs, called prompts, specifying what the user wants. The prompts can be as simple as “Tell me the weather today in Memphis, TN” to very elaborated prompts that ask the LLM to act like an expert, i.e., role assignment such as “You are a tutor.”, and solve a particular task such as explaining a Python code example or describing the structure of an argumentative essay.

In many respects, LLMs are an impressive technology providing state-of-the-art performance in many tasks. Despite their performance, LLMs suffer from a number of weaknesses such as hallucinogenic behavior, inconsistency, non-deterministic behavior, and data contamination (Zhao et al., 2023; Morris, 2024; Mirzadeh, 2024).

Particularly relevant for our discussion here is LLMs’ ability to “chat” with the user, i.e., have a multi-turn conversation, similar to conversational ITSs which is the major education technology on which we focus in this chapter. The chat-based or conversational LLMs started with GPT 3.5 which can interact with the user via conversation. This dialogue-based interaction allows ChatGPT to “answer follow up questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests” (OpenAI, 2024). It should be noted that GPT 3.5 was released soon after InstructGPT which was trained to follow instructions given in a prompt in order to provide a response (Ouyang et al., 2022). In ChatGPT, multi-turn dialogue is implemented using a mechanism in which prior prompts and responses become context for the next conversation turn, i.e., the next user prompt and prior context are used to generate the next ChatGPT response. The dialogue history or context serves more than just that, e.g., it serves as some kind of implicit user model, as no explicit user model is being maintained. In adaptive ITSs, a fine-grain, accurate user/learner model is needed in order to adapt instruction to each learner, which means a separate, explicit learner model must be maintained when using LLMs for such ITSs. In fact, given the absence of such fine-grain, accurate, and explicit open learner models, LLMs should be used for augmenting ITSs using a task-specific approach as explained later.

When using LLMs for conversational ITSs, there are some unique challenges and opportunities. ITSs with conversational dialogue form a special category of educational technologies (Rus et al., 2013). These conversational ITSs are based on explanation-based constructivist theories of learning and the collaborative constructive activities that occur during human tutoring. Conversational ITSs have several advantages over other types of ITSs. They encourage deep learning as students are required to explain their reasoning and reflect on their basic approach to solving a problem. Conceptual reasoning is more challenging and beneficial than mechanical application of mathematical formulas. Furthermore, conversational ITSs have the potential of giving students the opportunity to learn the language of scientists, an important goal in science literacy. Developing conversational ITSs requires authoring the major components of any ITS, conversational or not, such as the domain model, learner model, pedagogical model, and

interaction/interface model while keeping in mind the major peculiarity of conversational ITSs, which is the fact that the main form of interaction is natural dialogue and student inputs are mainly natural language statements. This peculiarity brings up unique challenges such as dialogue management, question/hint and feedback generation, and assessment of student natural language statements. This chapter provides a summary of our views and experience with respect to using LLMs for developing conversational ITSs. Furthermore, we discuss more broadly whether LLMs are good for learning in their current state, i.e., if they benefit learners if the learners freely use LLMs. In this regard, we make two major claims: (1) LLMs should be used for developing conversational ITSs using a task-specific approach as opposed to a more holistic approach and (2) there is a need to develop an educational wrapper around current LLMs in order to make them effective for learning. Last but not least, we will discuss potential applications of LLM-based conversational tutors for US Army applications and make recommendations for the Generalized Intelligent Framework for Tutoring (GIFT) (Sottolare et al., 2017) and STEEL-R (Goldberg, 2020). GIFT is an empirically-based, service-oriented framework of tools, methods and standards to make it easier to author adaptive instructional systems (AIS), manage instruction and assess the effect of AIS, components and methodologies (Sottolare et al., 2017). STEEL-R is a combination of existing and new software that captures and evaluates the experiential competence of individuals and teams (Goldberg, 2020). STEEL-R incorporates competency-based experiential learning and modified versions of the US Army Combat Capabilities Development Command (DEVCOM) Soldier Center's GIFT, and other existing components.

The Role of Generative AI for ITSs

As pointed out earlier, LLMs/Generative AI tools can generate text, images, videos, and other content, e.g., code examples, and can provide state-of-the-art solutions to many natural language processing tasks such as question answering, question generation, machine translation, and semantic similarity. Some of those tasks such as question answering, question generation, and semantic similarity are of key importance to conversational tutors in particular. For instance, semantic similarity is of key importance to understanding what students say in response to a tutor question and therefore assessing their mastery level (part of the learner model) in conversational ITSs. More generally, we make the case that LLMs can be used to develop all major components of ITSs such as content creation (e.g., instructional items that can serve a particular learning goal such as code examples meant to offer learners the opportunity to master a particular programming concept), domain models, student models, pedagogical models, and interaction models. We make our case in the context of dialogue-based ITSs as most of our work is in the context of such ITSs. As already noted, chat capable LLMs fit well with such ITSs, as LLMs' main focus is text-to-text generation, i.e., map a textual prompt into a textual response. Multi-modal input and output, i.e. language mixed with other types of content such as images, is possible but language is the primary output of existing LLMs. In this chapter, we focus on text-to-text LLMs only.

There are many ways in which LLMs can be used to develop ITSs as noted in our recent work (Rus, 2024). We outlined three major approaches: (1) holistic ITS specification in a one-page prompt as suggested by OpenAI, the creator of ChatGPT, (2) an ITS emulation approach, and (3) the task-specific approach that relies on existing ITS frameworks that have been developed and refined over recent decades. We argued for the latter, task-specific use of LLMs in ITSs in general and conversational ITSs in particular (Rus, 2024). The main advantage of this approach is that task specific prompts and output curation strategies can be designed which should lead to higher quality, more robust, less risky (i.e., minimize hallucinations), and more open and adaptive training systems. The main goal is to use LLMs to improve capabilities of current ITSs in a systematic, transparent, and reproducible manner while also speeding up the process and reducing development costs. This is in contrast with approaches which aim at entirely replacing existing ITS architectures and relying exclusively on LLMs as in the one-page prompt ITS approaches (Rus, 2024). This one-page prompt full-specification of an AI-Tutor was suggested by Open-AI on their website. This one-page prompt, which basically instructs the AI Tutor how to drive the interaction with the student, seems

bold yet too risky and infeasible. Our recommended task specific approach has other advantages such as enabling us to better understand the pros and cons of using LLMs for ITSs, which is much needed at this early, hype stage of Gen-AI. Furthermore, we argue for ITS developers to build or augment their existing ITS platforms with LLM-based components only after clear evidence that LLMs can lead to reliable and consistent improvements, and that well understood and documented processes with respect to how to use LLMs in ITS development are in place. In this approach, existing ITS platforms will drive the interaction with the learners based on decades of development, testing, and refinement while making use of LLMs where clear evidence justifies such use.

Implementing Conversational ITSs Using Generative AI

We propose two methods for implementing conversational ITSs using the task-specific approach. In the first method, LLMs can be used to decompose the complex task of tutoring into components and then the same or other LLMs can be prompted to solve each component. As an example of this approach, a separate chapter in this book presents the use of LLMs for assisting students with argumentative essay writing (Swartout et al., 2026). The first step in developing the tool was to prompt LLMs to decompose the task of essay writing into smaller steps, basically, asking an LLM for the structure of an argumentative essay, followed by using the LLMs to assist with each of the smaller steps, e.g., to help with the introduction paragraph, which is supposed to state the main thesis of the essay. This method has the same major weakness as the more holistic, one-page prompt ITSs mentioned earlier in that they are too risky - less controllable and harder to manage – although they are more transparent than the more holistic approach. As a way to mitigate the risk associated with overly relying on LLMs for both the task decomposition, component level prompting and integration of step level responses, a two-phase strategy can be utilized which is implemented in a separate application. This external application/wrapper will first prompt LLMs to decompose the targeted task and then prompt the LLMs to solve each subtask of the more complex tasks and integrate the results in a meaningful, coherent manner that would maximize students’ learning and learning experience while providing researchers and developers some level of control and transparency. This is the approach adopted by Deane (Swartout et al., 2026). The external application that orchestrates the two-phase strategy is similar to the tutoring platform in the second method described shortly, although it may not be an existing, full ITS platform. Another advantage of having an external application that orchestrates the overall task decomposition, subtasks, and integration of step-level results is that the external application can also store the prior user session so that the user can continue from where they left off. Current LLMs do not provide such prior session memory such that users can return and continue from where they left off in their prior session.

In a second method, existing tutoring platforms are used to orchestrate the overall solution of the complex task of tutoring while LLMs are used for task-specific solutions. An example of this approach has been illustrated previously by our team (Banjade et al., 2024; Oli et al., 2023; Oli et al., 2024). Basically, we use our DeepTutor/iCODE platform to orchestrate the conversational tutor-student interaction while using LLMs for task specific support, e.g., auto-assessment of student natural language responses (Oli et al., 2024) or generating code explanations for our iCODE tutor which provides adaptive scaffolding of students’ code comprehension processes (Oli et al., 2023).

We will focus next on how well LLMs can address specific tasks in conversational ITSs. Based on our experience so far, LLMs can provide competitive (but not perfect) solutions to specific tasks essential to any ITS and can help develop and author content as well as other items such as assessment instruments or personalized content (e.g., instructional tasks that fit learner’s background and career goals in order to make such tasks more meaningful and engaging) needed for ITS development. We have explored the use of LLMs for each major component of an ITS except the interaction/interface model, that is, we have investigated LLMs for domain modeling tasks, student modeling tasks, and pedagogical modeling tasks. Specifically,

we have explored these tasks in the context of a conversational ITS for code comprehension based on the instructional strategy of scaffolded self-explanation (Oli et al., 2023). For instance, we as well as others have shown that LLMs can automatically generate high quality code examples and accompanying explanations of those code examples which can act as benchmark explanations to assess students' explanations of the code (Oli et al., 2023). The LLMs can generate code examples and benchmark explanations with very high accuracy although not perfectly, which means the LLM-generated explanations must be checked by human experts. Code examples and benchmark explanations have typically been developed manually by human experts which was tedious, expensive, and a major scalability bottleneck. The advantage of using LLMs is that experts only need to spend a fraction of their effort for generating explanations as now they only have to check and correct, if needed, LLM-generated explanations. For our particular use, the generated code explanations did not follow particular theories that guide our iCODE tutor, e.g., code comprehension and learning theories, which means human experts had to spend more time to change the explanations to fit our purposes (Oli et al., 2023). Furthermore, it should be noted that iCODE focuses on learners in Computer Science 1/Computer Science 2 courses, i.e., intro-to-programming courses which use small code examples for which there is plenty of training data for LLMs in e-books and websites. Such electronic textbooks and websites provide ample repositories of simple code examples and accompanying explanations. For tasks where training data is not so widely available or not publicly available as in military applications, the performance of currently available LLMs may differ substantially.

The benchmark explanations are needed to assess students' mastery level (part of the learner model) as well as inform the next tutorial move, e.g., what feedback and next scaffolding move may be needed, which is part of the pedagogical model. LLMs can help automatically assess the correctness of students' explanations, not directly, but indirectly by prompting the LLMs to semantically compare the student explanation to the corresponding benchmark explanation (Oli et al., 2024). That is, if prompted directly to assess students' explanations, LLMs perform significantly worse than when prompted to assess the semantic similarity between a student explanation and a benchmark explanation.

There are a number of lessons learned from our experience of using LLMs for various tasks in conversational ITSs. First, the exact modelling of a target task when using LLM-based methods is very important. For instance, prompting LLMs to predict semantic similarity on a scale of [0-1], i.e., normalized score between 0 and 1, yields significantly superior performance compared to prompting to predict similarity on other scales (1-5 Likert scale). Second, there is a wide range in performance across various LLMs with GPT-4 leading to the best results in most cases. Third, LLMs perform comparably well, in particular GPT models, to fine-tuned encoder-based models. This is in general true about ChatGPT, i.e., it provides state-of-the-art solutions on many tasks on par with more fine-tuned, dedicated solutions to these tasks. That is, it is a general model that works as well as more specialized solutions on many tasks – a Swiss knife kind of tool. While it does not completely solve many of these tasks, it provides state-of-the-art solutions.

LLMs have been explored by others for implementing conversational ITSs or components of such ITSs. While we focused so far on specific ITS-related tasks and components, others have focused on more holistic ITS development using LLMs. We review next several such holistic approaches.

Schmucker and colleagues (2023) used LLMs to emulate an existing tutoring framework, the Expectation-Misconception Tailored (EMT) dialogue framework used in AutoTutor (Graesser et al., 2008). While this approach is trying to emulate the EMT framework using LLMs, the result is a conversational companion for a textbook or other sources of 'lesson content' that form the input to the approach. The proposed process is simple and the resulting ITS is basic (similar to an enhanced question answering system for an e-book). For instance, there is no student model that can reflect at each moment a students' level of mastery. The instructional tasks are verification questions about the content of an e-book as opposed to more complex tasks such as problems to be solved by the learner. It is not clear what the nature of these verification

questions is either. Are they shallow, factoid questions (what?) or deeper, more conceptual questions (why?). Another shortcoming of this approach is that there are no misconceptions identified or documented, instead the authors indicate they rely on “GPT4’s ability to detect factually incorrect information in the learner’s responses during the active teaching process” (p.4). This limits the number of learning opportunities such as immediately correcting well known misconceptions the learner may articulate. However, it has recently been shown that LLMs are not very good at identifying missing parts or incorrect parts in student generated responses when prompted directly to do so (Banjade et al., 2024). They do a much better job when the task is reframed as a semantic similarity task for which the correct and incorrect (misconceptions) are readily available.

Sonkar and colleagues (2023) describe a design framework called Conversational Learning with Analytical Step-by-Step Strategies (CLASS) for building advanced ITSs using LLMs. The CLASS framework was proposed to empower ITSs with two key features: 1) a carefully curated scaffolding dataset to train and enhance the LLM with problem-solving strategies allowing the ITS to provide improved step-by-step guidance to learners; and 2) through the help of a second conversational dataset, CLASS improves on the teacher-student natural language interaction with more accurate and more engaged conversations. While state-of-the-art ChatGPT was used to facilitate the creation of the two datasets, the authors used a separate, open-source LLM (Vicuna-13b, Chiang et al, 2023) for the teacher-student conversations. This strategy allowed them to train their own custom LLM model for ITS conversations on a specific topic, in this case biology, from the two datasets. A second reason for using an in-house LLM is a privacy and ethics concern that most educational institutions have: by not feeding the conversations back into an open GPT system, the learner data is protected. The framework was then used to develop SPOCK, a conversational-based tutoring system for biology. The focus of CLASS and SPOCK is on problem solving or helping students answer “hard problems”/questions in biology corresponding to levels 4-6 in Bloom’s taxonomy. It is not clear if the “hard problems” are indeed from levels 4-6 in Bloom’s taxonomy as the shown prompts only ask for “hard, challenging problem which can be broken down into subproblems.” For instance, the problem shown in appendix B.3.1. focuses on “What is mitochondria?” which is a definitional question. Definitional questions belong to level 1 in Bloom’s taxonomy. Also, the problems do not seem to have degrees of difficulty such that they could be appropriately assigned for students with varied mastery levels according to Zone of Proximal Development principles. Furthermore, the scaffolding and conversational datasets are synthetically generated by LLMs, and it is not clear how realistic they are, i.e., properly modeling actual student responses. No detailed evaluation is reported of various components of the system such as the number of turns of the LLM-generated conversations. No learner model is being used in either generating the “hard problems” or in the scaffolding and conversation data and corresponding trained components. It is an interesting effort which will require substantial improvement and evaluation. The authors also acknowledge that their LLM was prone to occasional inaccuracies and hallucinations, indicating that additional guardrails are needed to further improve on the accuracy of the LLM responses during the teacher-student interaction. Overall, we think their approach falls under the category of one-page prompt approaches to develop ITSs using LLMs (Rus, 2024) as indicated by the inference prompt in appendix A.4. (Sonkar et al., 2023). Both efforts by Sonkar and colleagues (2023) and Schmucker and colleagues (2023) use LLMs to generate questions/problems from e-textbooks and then develop tutorial dialogues around those questions/problems. Notably, both these studies indicate impressive LLM performance, e.g., Sonkar and colleagues (2023) indicate perfect (100%) or nearly perfect performance (upper 90s), depending on the performance criteria, when reporting the ability of LLMs to generate tutorial dialogues. They report results only for the scaffolding dataset and use criteria such as “Is the solution to the main problem factually correct?”. There is no manual creation of dialogues involved for training and evaluation purposes in both the efforts by Sonkar and colleagues (2023) and Schmucker and colleagues (2023) as is the case for the work presented next.

Al-Hossami and colleagues (2024) explore the role of LLMs to create a Socratic code debugger for novice programmers. Socratic debugging is defined by the authors as a conversation between a knowledgeable

programmer and a beginner student who wants help fixing a buggy solution for a simple computational problem. To develop the Socratic debugger, they manually created a dataset of multi-turn Socratic interactions meant to help beginners fix “buggy” solutions to simple computational problems. The dataset was then used for testing the Socratic debugging abilities of a number of LLMs. This is a key feature of this work, in contrast to the effort of Sonkar and colleagues (2023) which is presented above, which generated the data automatically. Not only did Al-Hossami and colleagues manually create Socratic interactions, but they did so in a manner that accounts for different “student abilities”. Instructors can often think of multiple ways of guiding the student at any particular turn in the conversation depending on their abilities, therefore, the dataset contributors were asked to provide all alternative utterances that they think could help the student at every turn in the conversation. The resulting Socratic debugger was evaluated both manually and automatically. The manual evaluation involved three human evaluators. At each instructor dialogue turn, the evaluators manually examined each LLM-generated utterance to determine if it was an appropriate Socratic utterance at that turn. The results indicated precision and recall ranging from single digit (recall of 5.5%) to high 50s (recall of 57.5%) depending on the type of prompt, simple versus chain of thought. The typical performance was precision and recall in the low 40s, i.e., 40% of LLM-generated utterances were correct (precision) and 40% of expert-generated utterances were also generated by LLMs (recall). A key feature of this work is the systematic and rigorous evaluation and accounting for learners’ abilities, i.e., including some kind of learner model.

Discussion and Recommendations for Future Research

The role of LLMs for conversation ITSs could be transformative although it is too early to make a final conclusion. As already noted, chat-based LLMs seem to be a very good fit for chat-based, conversational ITSs. However, this quick assessment of the role of LLMs for conversational ITSs based on the key interface model for both chat-based LLMs and conversational ITSs is deceiving as we argued in this chapter. LLMs cannot straightforwardly implement conversational ITSs, e.g., there is not explicit learner model or session memory implemented in currently available LLMs, not to mention their weaknesses such as inconsistency and unstable behavior such as being surprisingly sensitive to variations in prompt wording or across multiple models, multiple generations of the same model, or even the same model when repeated several times (Oli et al, 2023; Morris, 2024). To fully understand the role and potential of LLMs for conversational ITSs, there is a need for solid, systematic studies that can document the strengths and weaknesses of LLMs for such ITSs. As a way to monitor progress, there is a need for benchmark data sets for conversational ITSs and other ITSs similar to benchmark datasets for main Natural Language Processing (NLP) tasks such as question answering, machine translation, or summarization. Such benchmark datasets can be used to measure and report the performance of LLMs on key ITS tasks as new LLMs are being released, similar to HELM – Holistic Evaluation of Foundation Models (Stanford, 2024). Potential ITS tasks for which benchmark datasets should be created in order to monitor LLM performance are: (personalized) content generation, question answering, question/hint generation, automated assessment/learner modelling, and knowledge component extraction/domain modelling.

More broadly, the impact of LLMs on education is both promising and unsettling. For instance, there will likely be a shift in what skills are needed to do well in a world where Generative AI can assist users in solving various tasks. For instance, as we claimed in Rus (2023), comprehension becomes even more important now in the Generative AI era as users must comprehend (in order to assess) what Generative AI tools produce. That is, there will be a major shift from generation processes to comprehension processes when using such tools and therefore, training of the future workforce should emphasize such skills. It is important to note that there is still a generation task the user must do, which is generating/creating the prompt. Nevertheless, LLM-assisted prompt generation (or prompt completion, i.e., complete a prompt the user starts creating) can reduce or eliminate this generation task too with the user only needing to specify a number of key requirements as bullet points, thus leaving all the details of actually creating the prompt to

such tools that can assist with prompt generation/completion. In fact, our prediction is being voiced by others as well (Denny et al., 2023). Denny and colleagues (2023) note that, in the context of Computer Science (CS) education, “there is now a shift in importance towards code reading, comprehension and evaluation of LLM-generated code.” (p. 296). We believe this shift will be broad, across all domains, not only in CS education. On a similar note, Denny and colleagues (2023) argue that CS students will have to focus now on “prompt problems” as opposed to coding problems, i.e., students will have to master writing prompts that will effectively ask LLMs to write a specific code for them and comprehend and assess/evaluate the generated code. This reflects a bigger shift in problem solving or performing almost any task assisted by LLMs. Indeed, instruction following, chat-capable LLMs seem to be leading to a paradigm shift in problem solving with computers in the sense that the users specify in natural language the requirements of the task to be solved and LLMs, if rightly prompted, will provide the solution or assist with finding the solution. This is in contrast to the current approach where first, requirements are specified by domain experts/business professionals in natural language and/or some other formats, e.g., diagrams, and then business analysts together with programmers translate those requirements into code-based solutions that computers can interpret and execute. There are many implications of this shift such as the need to master the natural language-based specifications, i.e., the prompting, to be aware of risks and weaknesses of LLMs and corresponding mitigation strategies, as well as the need for LLM output comprehension and evaluation skills as already noted.

LLMs do seem to have the ability to do task decomposition extending the capabilities of search engines to solve complex, multi-step tasks (White, 2024). LLMs show promise with helping users in navigating and resolving/completing complex tasks (see the argumentative essay writing task mentioned earlier). In contrast, traditional search engines are not able to understand the complex user tasks and therefore not able to properly assist users with those tasks. It is up to the user to break down a complex task into subtasks and smaller steps, plan a solution search strategy which may eventually involve asking a search engine or someone on a web forum to assist with a specific step, one at a time, and then assess all those small step findings, refine the search as needed, and put together and compose a coherent solution from the final responses to each step. As already noted, LLMs exhibit some early promises of both task decomposition and solution integration thus alleviating users of significant effort and making them more productive. The downside of this task decomposition and solution integration is that for novices/learners it may be counterproductive. Both White (2024) and Shein (2024) suggest that LLMs in their raw form may not be appropriate for learning. We agree: LLMs in their current form are doing too much for the learners who do not have a chance to fully engage all learning processes they are supposed to engage for effective learning. In fact, one main thesis of this chapter is that LLMs need to be augmented or wrapped in some educational wrapper in order to be conducive to learning, i.e., in their raw form, they do too much for the learner and therefore run the risk of being used as solution providers as opposed to learner companions. As evidence for this thesis, Shein (2024) described an experiment with LLMs versus traditional search engines in which they showed that students learn more when using a search engine whereas they are more productive on new tasks when using LLMs. That is, LLMs are good for helping complete tasks, even new ones, but not for learning and mastering the underlying concepts and processes. Shein (2024) notes: “Working hard and struggling is actually an important way of learning. When you’re given an answer, you’re not struggling and you’re not learning. And when you get more of a complex problem, it’s tedious to go back to the beginning of a large language model and troubleshoot it and integrate it” (p. 13).

Mastering a domain is still needed. One can argue the opposite – that mastering a domain is not needed anymore given that LLMs can do complex task decomposition and also handle step-level subtasks. There is some validity to this argument for tasks where there is a way to assess the correctness and completeness of the output of LLMs. However, given that LLMs are not perfect and providing correct solutions that can be assessed is not guaranteed, the typical use case is the user asking for some output which they need to comprehend and assess by themselves which means they need to master the domain. Furthermore,

prompting an LLM to solve a task for the user requires some expertise. Even though prompt completion tools can assist the core user, the need must be specified by the user.

It is beyond the scope of this chapter to discuss more broadly the opportunities and challenges of LLMs in education. Xu and colleagues (2024) or Wang and colleagues (2024) provide an overview of LLMs for education. We will just emphasize two important aspects indicated also by Mannekote and colleagues (2024) which must be addressed for LLMs to be more impactful for ITSs more broadly. One is the need to develop more computing efficient language models for use in educational settings with limited resources (Hoffmann et al., 2022). Secondly, the ability to interpret the behavior and output of these AI powered agents is crucial for trustworthiness. Though these neural network based AIs have been seen for the most part as black-box systems, there is recent work focused more on the reasoning of the LLM in producing the output (as we have seen with the latest ChatGPT-o1 model). The new concept of Open Learner Models (OLMs) could enable learners to view and interact with the model's representation of their knowledge (Bull, 2020). These can revolutionize educational technologies by making learning processes more adaptive and personalized, though their implementation should be guided by "strong ethical and pedagogical standards" (Mannekote et al., 2024).

To sum up, LLMs do show a lot of promising potential for improving ITSs and there is early promise on various tasks such as content generation, question/hint generation, and auto-assessment. More rigorous and systematic research is needed to fully understand their true potential and shortcomings.

Recommendations for GIFT and STEEL-R Overall

The advent of LLMs could have many implications for the future of simulation and training in terms of skills, processes, and the development of technologies for training and simulation such as GIFT and STEEL-R.

As already noted earlier, first there must be a systematic and rigorous effort spent on understanding what LLMs can and cannot do for the education ecosystem. Part of this first phase, benchmarking key ITS-related tasks for monitoring LLM performance, is critical. Then, the focus should be on identifying new skills, e.g., prompting strategies, prioritizing "old" skills such as comprehension, developing novel processes of learning, and developing advanced learning technologies such as conversational ITSs.

One potential application of LLM-based conversational ITSs is for after action reviews (AARs). The AAR framework is complex as it involves many phases, players, guidelines, rules, content areas, processes and strategies (Townsend et al., 2018). For instance, the AAR involves many phases during an exercise, before the AAR, during the AAR, and before the next exercise. Examples of such phases are Observe Performance, Review Performance Objectives and Goals, Establish Tactical Timeline, Conduct Force of Four Discussions, and Set Goals. There are other phases supporting the main phases such as AAR Facilitator/Instructors AAR Preparation. Furthermore, the AAR framework includes many players with different roles such as observers/controllers (OC) and instructors, the AAR facilitator (Platoon Leader), and squad members. In addition, there are guidelines (e.g., Guidelines 1-5) and rules (20/80 rule: 20% of the time the AAR Facilitator/Instructors speak, 80% of the time Squad Members speak), content areas (e.g., TC3, ACA), and activities and processes to be implemented to achieve certain goals such as sensemaking (Townsend et al., 2018).

Given the complexity summarized above, specific tasks in the process could be targeted for conversational tutor support. The tasks and training can target both instructors and squad members. Examples of AAR-related tasks that could be implemented and augmented with conversational tutors in GIFT are illustrated below:

- A question generation (QG) tool can help the instructors/AAR facilitator and researchers with automatically generating and selecting questioning to be used during the AAR. The QG tool would follow the questioning techniques and guidelines mentioned in the various sources we consulted. The QG tool can be developed into a questioning training tool for instructors, i.e., a training tool that facilitates “Questioning Techniques Practical Application Role Play” for instructors.
- Automatic assessment of AARs can be facilitated through the development of a tool that analyzes how effective AARs are and provides feedback. AARs are effective if they follow the guidelines and rules and implement processes and strategies as specified in the AAR framework. For instance, AARs are ineffective if only the leaders speak or when they fail to set tactical and team goals for improvement.
- Implement the Force-of-Four phase with one learner/squad member. Implementing the Force-of-Four phase means implementing its four major steps: identify where trigger events caused a challenge, review how the challenge was addressed, discuss alternate solutions, and discuss alternate solutions’ impact on the mission. While the focus will be on implementing the Force-of-Four phase, it will require input from the other phases, e.g., review the most critical events which are identified during the Observe Performance phase when instructors and the AAR facilitator (Platoon Leader) choose the most important critical events and squad behaviors for review (Guideline 2). Furthermore, each instructor defines major training objectives and then identifies related errors and examples of good performance (Guideline 3). This latter part is important for implementing the Force-of-Four phaser automatically as conversational tutors need benchmark responses, i.e., examples of good performance to automatically assess student/learner/squad member responses. LLMs can be explored as a potential replacement for existing solutions, e.g., to what extent they can generate examples of good performance and/or assess student responses without the availability of a benchmark response.

For all the exemplified tasks above, there are many details that need to be ironed out such as the input for each task, e.g., multimodal input such as a video of the simulated exercise or a more abstract description of a timeline and major events with video segments of each of those major events for the learner to review. Who are the human and virtual participants in the simulated AAR process? Some possible scenarios: a) one human player being part of a team with multiple Non-Player Characters and a virtual AAR instructor; b) all team players are human and the facilitator is a human assisted by a multi-modal AI agent to guide the AAR process; c) similar to b) except some team players are NPC; and d) similar to b) or c) but the AAR process is driven by an AI facilitator.

The AAR framework or parts of it could be implemented using a multi-agent approach where there is a facilitator conversational agent which drives the interaction/conversation whereas other agents can play the role of instructors and/or other squad members. As a first step, the focus could be on one squad member who interacts with a conversational agent to review their personal performance during a prior mission before participating in the actual AAR with instructors and other squad members. Even more, such an individualized self-reflection guided by a conversational agent may be also scheduled after the actual AAR. In such a post-AAR session, the squad member may be asked by the agent tutor to reflect again on their performance or another squad members’ performance (peer review). In those cases, the conversational agent would prompt the squad member to reflect on their performance during a prior mission and provide feedback, e.g., the conversational agent could show the squad members’ actual performance.

Conclusions

We discussed in this chapter the role of LLMs on conversational ITSs and more broadly the role of LLMs on education. Early work on using LLMs to develop ITSs leads to mixed results depending on the goal and evaluation methodology therefore the impact of LLMs on conversational ITSs is still yet to be fully

understood. We argued that LLMs in their current state are not good for learning, i.e., broadly speaking, as they do too much for the learner. Some kind of educational wrapper must be built around LLMs to be inductive of learning. Last but not least, we made recommendations for GIFT and STEEL-R indicating how conversational ITSs based on LLMs could be used for tasks such as AARs and how LLMs are reshaping the set of skills future generations need.

Acknowledgements

This research was sponsored by DEVCOM SC-TSD-STTC and was accomplished under Cooperative Agreement Number W912CG-24-2-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of DEVCOM SC-TSD-STTC or the US Government. The US Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation herein. This work has also been partially supported by the following grants awarded to Dr. Vasile Rus: the Learner Data Institute (NSF award 1934745); CEdPad (NSF award 1822816); and iCODE (IES award R305A220385). The opinions, findings, and results are solely those of the authors and do not reflect those of NSF and IES, or DoD. Neither NSF nor IES have approved or endorsed its content.

References

- Al-Hossami, E., Bunescu, R., Smith, J., and Teehan, R. (2024). Can Language Models Employ the Socratic Method? Experiments with Code Debugging. In Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1 (SIGCSE 2024). Association for Computing Machinery, New York, NY, USA, 53–59. <https://doi.org/10.1145/3626252.3630799>.
- Banjade, R., Oli, P., & Rus, V. (2024) Identifying Gaps In Students’ Explanations of Code Using LLMs, Proceedings of 25th International Conference on Artificial Intelligence in Education, July 8-12, 2024, Recife, Brazil
- Bull, S. (2020). There are open learner models about! IEEE Trans. Learn. Technol. 13, 425–448. doi: 10.1109/TLT.2020.2978473
- Chiang, W., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., and Xing, E.P. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.
- Denny, P., Leinonen, J., Prather, J., Luxton-Reilly, A., Amarouche, T., Becker, B.A., & Reeves, B.N. (2023). Prompt Problems: A New Programming Exercise for the Generative AI Era. Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1.
- Goldberg, B (2020). STE Experiential Learning-Readiness (STEEL-R) Strategy. STTC White Paper.
- Graesser, A. C., Rus, V., D’Mello, S., and Jackson, G. T. (2008). AutoTutor: Learning Through Natural Language Dialogue That Adapts to the Cognitive and Affective States of the Learner. In Current Perspectives on Cognition, Learning, and Instruction: Recent Innovations in Educational Technology That Facilitate Student Learning, ed. D. H. Robinson and G. Schraw, 95–125. Charlotte, NC: Information Age Publishing.
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., et al. (2022). “An empirical analysis of compute-optimal large language model training,” in Advances in Neural Information Processing Systems, Vol. 35, eds. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh (Red Hook, NY: Curran Associates, Inc.), 30016–30030.
- Mannekote A, Davies A, Pinto JD, Zhang S, Olds D, Schroeder NL, Lehman B, Zapata-Rivera D and Zhai C (2024) Large language models for whole-learner support: opportunities and challenges. Front. Artif. Intell. 7:1460364. doi: 10.3389/frai.2024.1460364
- Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2024). GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. <https://arxiv.org/abs/2410.05229>
- Morris, M.R. (2024). Prompting Considered Harmful. Commun. ACM 67, 12 (December 2024), 28–30. <https://doi.org/10.1145/3673861>

- Oli, P., Banjade, R., Tamang, L.J., Rus, V. (2023). The Behavior of Large Language Models When Prompted to Generate Code Explanations, NeurIPS'23 Workshop on Generative AI for Education (GAIED), New Orleans, 15 December, 2023.
- Oli, P., Banjade, R., Chapagain, J., and Rus, V. (2024) Automated Assessment of Students' Code Comprehension using LLMs, Proceeding of AAAI 2024 Workshop on AI for Education - Bridging Innovation and Responsibility, February 26-27, 2024, Vancouver, Canada.
- OpenAI. (2024). <https://openai.com/index/chatgpt/> (accessed on November 17, 2024)
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. 2024. Training language models to follow instructions with human feedback. In Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22). Curran Associates Inc., Red Hook, NY, USA, Article 2011, 27730–27744.
- Rus, V., D'Mello, S., Hu, X., & Graesser, A.C. (2013). Recent Advances in Conversational Intelligent Tutoring Systems, *AI Magazine*, 34,(3):42-54.
- Rus, V. (2023). Generative AI And Its Impact: An Educator and AIED Researcher's View. Accessed Online on March 25, 2024.
<https://drive.google.com/file/d/1SCwkqVfI1tHHNvnpQQIWhzFQl9xZyjq/view?usp=sharing>
- Rus, V. (2024). Large Language Models and Their Implications for Conversational Tutors and GIFT. In Proceedings of the 12th Annual GIFT Users Symposium. Orlando, FL: US Army Combat Capabilities Development Command - Soldier Center. ISBN 978-0-9977258-6-5.
- Schmucker, R., Xia, M., Azaria, A., and Mitchell, T. (2023). Ruffle and Riley: Towards the Automated Induction of Conversational Tutoring Systems, NeurIPS 2023 Workshop on Generative AI for Education (GAIED), December, 2023
- Shein, E. (2024). The Impact of AI on Computer Science Education. *Communications of the ACM*, 67(9), September 2024
- Sonkar, S., Liu, L., Mallick, D.B., & Baraniuk, R. (2023). CLASS: A Design Framework for Building Intelligent Tutoring Systems Based on Learning Science principles. Conference on Empirical Methods in Natural Language Processing.
- Sottolare, R. A., Brawner, K. W., Sinatra, A. M., & Johnston, J. H. (2017). An updated concept for a Generalized Intelligent Framework for Tutoring, R. A., Goldberg, B. S., Brawner, K. W., & Holden, H. K. (2012). A modular framework to support the authoring and assessment of adaptive computer-based tutoring systems (CBTS). In Proceedings of the Interservice/Industry Training, Simulation, and Education Conference (pp. 1-13).
- Stanford. (2024). <https://crfm.stanford.edu/helm/lite/latest/#/> (accessed on November 17, 2024)
- Swartout, W., Deane, P., & Nye, B. (2025). Using Generative AI to Improve Student's Writing, in Sinatra, A.M., Graesser, A.C., Lawton, P.M., and Rus, V. (Eds.), Design Recommendations for Intelligent Tutoring Systems: Volume 12 - Generative AI in Intelligent Tutoring Systems. Orlando, FL: US Army Combat Capabilities Development Command - Soldier Center.
- Townsend, L. & Johnston, J. & Ross, W. & Milham, L. & Riddle, D. & Phillips, H. (2018). An Integrated After Action Review (IAAR) Approach: Conducting AARs for Scenario-Based Training Across Multiple and Distinct Skill Areas. 10.1007/978-3-319-91122-9_20.
- Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P.S., & Wen, Q. (2024). Large Language Models for Education: A Survey and Outlook. ArXiv, abs/2403.18105.
- White, R. W. (2024). Advancing the Search Frontier with AI Agents. *Communications of the ACM*, 67(9), September 2024
- Xu, H. & Gan, W. & Qi, Z. & Wu, J. & Yu, P. (2024). Large Language Models for Education: A Survey. 10.48550/arXiv.2405.13001.
- Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J., & Wen, J. (2023). A Survey of Large Language Models. <https://doi.org/10.48550/arXiv.2303.18223>.

CHAPTER 3 – LARGE LANGUAGE MODELS AND INTELLIGENT TUTORING SYSTEMS: CONFLICTING PARADIGMS AND POSSIBLE SOLUTIONS

Punya Mishra¹, Danielle S. McNamara¹, Gregory Goodwin², and Diego Zapata-Rivera³

Learning Engineering Institute, Arizona State University¹; US Army Combat Capabilities Development Command (DEVCOM) – Soldier Center²; ETS Research Institute³

Introduction

The advent of Large Language Models (LLMs) has fundamentally disrupted our thinking about educational technology. Their ability to engage in natural dialogue, provide contextually relevant responses, and adapt to learner needs has led many to envision them as powerful tools for personalized learning. This emergence raises important questions about their relationship with Intelligent Tutoring Systems (ITSs), which have long been the gold standard for computer-based personalized instruction through their structured, discipline-focused approach. While the potential for integrating these technologies is compelling, significant theoretical and practical challenges remain. This paper examines these challenges and proposes new ways of conceptualizing the relationship between LLMs and ITSs to enhance both personalization and learning outcomes.

Background: Intelligent Tutoring Systems

ITSs, a subset of intelligent systems in education, represent a significant advancement in computer-aided learning. These sophisticated software platforms incorporate artificial intelligence (AI) components to create dynamic, adaptive learning environments. Unlike traditional computer-assisted instruction, ITS actively track and analyze students' work, providing tailored feedback and hints in real-time. By collecting and processing data on individual student performance, including both cognitive and non-cognitive variables, these systems can make nuanced inferences about learners' strengths and weaknesses, subsequently recommending personalized learning paths (Kulik & Fletcher, 2015; Shute & Zapata-Rivera, 2012). The development of such systems requires not just technical capabilities but also careful consideration of stakeholder needs through participatory design approaches involving both educators and learners (Jurenka et al., 2024). Recent studies have also demonstrated that these systems can achieve accuracy rates of over 90% in assessing student responses and providing appropriate feedback (Oli et al., 2023). However, despite the precision and adaptability of ITS, their implementation in classrooms has faced significant limitations, especially in contrast to the ease of use and adaptability of LLM-based solutions.

Key Components and Challenges of ITSs

The design of ITSs, it is argued, has at its foundation three key components. These systems must ideally possess: (1) knowledge of the learner (student model), (2) knowledge of the domain (expert model), and (3) knowledge of teaching strategies (pedagogical model). In recent years, LLMs have emerged as potential tools for enhancing these components of ITSs, offering flexible and highly responsive language-based capabilities that traditional ITSs may lack. In this context, LLMs can potentially assist in developing all three components, though careful curation and expert oversight remain essential (Rus, 2024). The integration of these knowledge bases marks a shift from earlier, knowledge-free computer-assisted instructional programs. A key distinguishing feature of ITSs is their capacity to diagnose students' errors with precision and adapt instruction based on these diagnoses. Furthermore, the development of these

components requires comprehensive evaluation frameworks that span quantitative, qualitative, automatic, and human evaluations to ensure their effectiveness and safety (Jurenka et al., 2024; Shahzad et al., 2024).

Despite the considerable potential of ITSs to revolutionize education, their widespread adoption in classroom settings remains limited. This gap between potential and implementation can be attributed to two primary factors: the substantial cost associated with developing and deploying these systems, and the inherent limitations in current educational measurement approaches. Traditional educational assessment relies heavily on broad indicators like grades and attendance, while ITSs can measure student performance at a much more granular level - this measurement mismatch creates significant challenges for maximizing these systems' educational value. Beyond these traditional challenges, there is a need for robust evaluation frameworks and responsible development practices when incorporating new AI technologies into educational systems. This includes careful consideration of safety, bias, and ethical implications through systematic evaluation processes involving both educators and learners (Al-kfairy et al., 2024; Jurenka et al., 2024; Warr et al., 2024a). Further, while ITS development and deployment can be prohibitively expensive, LLMs present an opportunity to mitigate some of these costs by utilizing pre-trained models and readily adaptable language technologies. While utilizing existing LLMs can potentially reduce certain development costs compared to traditional programming approaches, developing LLMs from scratch remains extremely expensive, and even using existing models can incur significant operational costs at scale (Benram, 2024; Buzby, 2024). Additionally, LLMs introduce new challenges requiring careful prompt engineering and output curation (Rus, 2024). For instance, Warr et al. (2024a) have demonstrated systemic and implicit bias in these systems' responses when asked to evaluate student essays (see also Goldshtein et al., 2024a, 2024b). The latter issue is particularly significant, as it underscores the need for more sophisticated assessment methodologies that can fully leverage the adaptive capabilities of ITSs.

The Promise and Limitations of LLMs in Educational Contexts

The advent of LLMs has been proposed as a promising avenue for enhancing ITS capabilities (Cao, 2023; Chen et al., 2023; Schmucker et al., 2024). In terms of the student model, LLMs can potentially identify, capture, and analyze vast amounts of learner data, including written responses, question-answering patterns, and conversational nuances, to build a more comprehensive and dynamic representations of the learner's knowledge state (Schmucker et al., 2024). Moreover, they can leverage this understanding to generate appropriate materials, tasks, and feedback as part of the adaptive learning cycle (Shute & Zapata-Rivera, 2012). For example, an LLM could analyze free-text responses from students in real-time, identifying misconceptions in an introductory physics course and providing targeted hints or resources. However, empirical studies show that LLMs' assessment capabilities vary significantly based on how the task is framed, with semantic similarity-based approaches outperforming direct assessment methods (Oli et al., 2023).

While these capabilities are promising, recent large-scale studies emphasize the importance of comprehensive evaluation frameworks to validate LLM performance. For instance, real-world deployments in university settings have shown that LLMs require careful scaffolding and oversight to effectively support student learning (Jurenka et al., 2024). For the expert model, LLMs trained on extensive domain-specific corpora can serve as robust repositories of knowledge, capable of generating accurate and contextually relevant information across a wide range of subjects (Chen et al., 2023; Mugaanyi et al., 2024). Recent research demonstrates that LLMs can extract domain models from textbooks with precision rates in the 60-70% range, though relation extraction remains more challenging (Rus, 2024). Sixty to seventy percent, though impressive, can be quite problematic in educational contexts, particularly when combined with the fact that these modes often (confidently) confabulate information. Perhaps most significantly, in the realm of the pedagogical model, LLMs and chatbots can leverage their natural language processing capabilities to implement sophisticated teaching strategies, adapting their communication style, explanation depth, and

question complexity based on the learner's responses and inferred comprehension level (Cao, 2023). This adaptive approach could potentially mimic the nuanced interactions of skilled human tutors, offering explanations, asking probing questions, and providing scaffolded support in a more natural and engaging manner (Hu, 2023; Pardos & Bhandari, 2023; Zapata-Rivera et al., 2024).

Fundamental Challenges with LLMs

However, a critical examination of LLMs reveals certain inherent characteristics that challenge their suitability for ITS roles (Nye et al., 2023). In brief, we propose that there are two interconnected arguments regarding the nature of LLMs that complicate their ability to serve as a foundation for ITSs: (1) the inherent necessity for these models to engage in what can be termed "hallucination," and (2) the consequent variability in their outputs. This variability has been systematically documented in recent studies examining LLM-generated educational content, where outputs vary significantly based on prompt wording and temperature settings (i.e., the degree of randomness in responses) - even when using identical inputs (Oli et al., 2023). These challenges necessitate a structured approach to responsible development that goes beyond traditional testing. Recent proposals have suggested using a combination of automatic evaluations, human expert ratings, and real-world user studies to assess both the technical performance and pedagogical value of LLM-enhanced systems. Such evaluation frameworks have identified specific areas where LLMs excel (such as promoting engagement) and areas where they need improvement (such as maintaining consistent pedagogical quality) (Jurenka et al., 2024).

Fundamentally, LLMs extrapolate beyond their training data due to their generalization capabilities. LLMs have been described as being "stochastic parrots" (Bender et al., 2021) due to the fact that their outputs are probabilistically generated word by word (or feature by feature) based on their training data. The implications of these characteristics for educational applications require careful consideration. Research has shown that successful deployment of LLMs in educational settings demands both automated safeguards and human oversight, with particular attention to safety, and the quality of pedagogical interactions. Studies involving real students have demonstrated that while LLMs can effectively support learning, their deployment requires careful consideration of student needs and appropriate guardrails. There is indeed some evidence that this is too simplistic a view since these models appear to have developed some higher order conceptual structures (Anthropic, 2024) as evidenced by their ability to perform a range of complex tasks on which they have not been trained. For instance, work by Oli et al. (2023) suggests that while LLMs can generate high-quality educational content with correctness rates above 90%, they still require expert oversight and curation, particularly for theory-driven educational content.

LLMs are merely generating words based on their training data, with little connection to the world. Essentially, LLMs "make stuff up" which leads to what have been colloquially termed as "hallucinations." It is important to note, however, that these misrepresentations are not necessarily a flaw but rather an intrinsic feature of the technology (Xu et al., 2024).

A corollary of this generative nature of LLMs is that they exhibit significant output variability even when provided the same input. This variability is not just random – it is highly context-dependent, with the same question receiving markedly different responses based on how it's framed. Recent research (Warr et al., 2024b) demonstrates substantial unexplained variance in responses from LLMs, persisting across different versions and resistant to mitigation attempts (e.g., using retrieval augmented generation - RAG). This variability is particularly pronounced in educational contexts, where studies have found that factors such as temperature settings and prompt wording can significantly impact the pedagogical quality of outputs (Oli et al., 2023). These small shifts can move interactions in unpredictable directions and conversations can stray increasingly far from intended learning paths without educators or learners being aware of the deviation. This is particularly concerning for learners who, by just where they are in their learning journey,

do not have the judgment to question the responses from AI. This can create a kind of conversational drift that can magnify small errors as the conversation progresses.

Additionally, research has shown that while LLMs can effectively assess student responses when framed as semantic similarity tasks, they perform poorly when directly asked to identify missing or incorrect parts in student work (Rus, 2024). These facts combined with the research that shows a significant prevalence of stereotyping has significant implications for their ethical use in student evaluation systems. This output variability can significantly undermine ITSs' intended purpose to deliver consistent, high-quality instruction. For example, a study by Warr et al. (2024b) found that identical prompts resulted in different feedback for student essays, raising concerns about fairness and consistency. Moreover, recent research has shown that a user's prior domain knowledge significantly impacts their ability to effectively utilize generative chatbots, with laypersons particularly limited in their capacity to engage with unfamiliar domains due to lack of expertise (Imundo et al., 2024).

These fundamental attributes of LLMs—hallucination and variability—conflict with the precise, consistent functioning required in traditional ITS roles. The tendency to hallucinate compromises the accuracy of student modeling and the reliability of domain knowledge presentation. Studies have shown that even state-of-the-art LLMs struggle with maintaining consistency in educational outputs, with performance varying significantly across different subject domains and task types (Rus, 2024). The high degree of output variability undermines the consistency necessary to implement effective pedagogical strategies.

While ITSs excel in delivering structured, accurate feedback for well-defined problems, LLMs offer the ability to handle more open-ended, creative inquiries, which could lead to a more holistic educational experience if properly integrated. Consequently, LLMs may be ill-suited for the core functions of ITSs, potentially leading to inconsistent instruction and misleading information.

Reconceptualizing the Role of LLMs in Education

In light of these limitations, we propose a shift in how we conceptualize the role of LLMs in educational contexts. Rather than serving as authoritative tutors within an ITS framework, LLMs may be more effectively utilized as thought partners in the learning process. This reconceptualization aligns with emerging evidence from real-world educational deployments. For instance, studies show that students primarily use LLM tutors for specific learning support tasks like understanding concepts and debugging code, rather than as comprehensive replacements for traditional instruction (Sheese, 2024). This aligns with findings that while generative chatbots show promise in supporting expert cognition and novice training, their effectiveness is significantly constrained when working with laypersons who lack sufficient domain knowledge to critically engage with the AI's outputs (Imundo et al., 2024). Furthermore, successful implementation of LLM-based educational tools requires not only careful attention to student feedback but also continuous refinement based on real-world usage patterns, suggesting an iterative design approach rather than a one-time deployment. This aligns with recent empirical findings suggesting that LLMs are most effective when used to augment rather than replace existing ITS components, with task-specific implementations showing more promise than general-purpose tutoring approaches. In this capacity, their ability to generate varying perspectives and engage in open-ended dialogue becomes an asset, fostering critical thinking, idea exploration, and information literacy. This approach leverages the strengths of LLMs while maintaining the crucial role of human expertise in education, aligning with contemporary pedagogical philosophies that emphasize active learning and critical evaluation of information. It has been suggested that LLMs could generate high-quality educational content and explanations, but their effectiveness is maximized when integrated within existing educational frameworks and subject to expert oversight. We believe that the synergy between these technologies could potentially address a significant gap in current educational technology approaches, creating a more holistic and engaging learning experience.

This reconceptualization particularly aligns with LLMs' context-dependent nature. While their varying responses to differently framed versions of the same question can be problematic in traditional tutoring roles that demand consistency, this sensitivity to context becomes an asset when LLMs serve as thought partners. In this role, their ability to approach problems from multiple angles based on how they are framed can help students understand the importance of perspective and problem formulation - key aspects of creative and critical thinking. This characteristic aligns with contemporary pedagogical approaches that emphasize the value of viewing concepts from multiple perspectives and understanding how different framings can lead to different insights.

ITSs excel in the realm of structured, disciplinary knowledge acquisition. Their strength lies in their ability to provide precise, personalized instruction and assessment, adapting to individual learning patterns and ensuring mastery of core concepts. Studies demonstrate that ITS platforms achieve optimal results when handling well-defined, rigorous domains where consistency and precision are paramount. This systematic approach is crucial for building a solid foundation of knowledge and skills. However, this rigorous focus on accuracy and mastery can sometimes lead to a learning experience that, while effective, may lack inspiration or fail to ignite a deeper passion for the subject matter.

LLMs could be leveraged to engage students in personalized, creative activities that make use of student model information. As such, LLMs would offer a complementary set of capabilities that align well with the more creative and exploratory aspects of learning. Their ability to generate novel connections, provide varying perspectives, and engage in open-ended, creative discourse has great potential to serve as an "intelligent creative buddy." This is consistent with findings that suggest that LLMs can effectively generate varying educational content, with particularly strong performance in creating explanations and examples that promote active learning, though the quality varies based on prompt design and model parameters. This aspect of LLMs could be instrumental in fostering intellectual curiosity, encouraging lateral thinking, and helping students develop a more nuanced and multifaceted understanding of the subject matter. By offering unconventional viewpoints or drawing unexpected parallels, LLMs could challenge students to think beyond the confines of traditional disciplinary boundaries, potentially leading to more innovative and interdisciplinary approaches to problem-solving.

A Synergistic Framework: Combining ITSs and LLMs

The thoughtful integration of these technologies in educational settings has strong potential to create a powerful learning ecosystem (Maity & Deroy, 2024). Successful integration requires a comprehensive evaluation-driven approach encompassing multiple dimensions: automatic evaluations for rapid development iteration, human expert evaluations for pedagogical quality, and real-world user studies for understanding actual impact. This multifaceted evaluation framework helps ensure that integrated systems maintain both technical accuracy and pedagogical effectiveness. This integration is most successful when following a task-specific approach, where LLMs are used to enhance specific ITS components rather than attempting to replace entire systems (Rus, 2024). ITSs could provide the structured, foundational knowledge and skills, ensuring accuracy and mastery, while LLMs could augment this learning by making them more conversational, offering creative interpretations, generating thought-provoking questions, and inspiring students to explore the broader implications and applications of their knowledge (Sifaleras, 2024).

Indeed, there is some preliminary evidence that such hybrid approaches can achieve both high accuracy in content generation while maintaining pedagogical rigor necessary for effective instruction (Oli et al., 2023). This dual approach could cater to both the need for rigorous academic standards and the equally important goal of nurturing creativity, critical thinking, and a genuine love for learning.

The design of the Generalized Intelligent Framework for Tutoring (GIFT) supports an additional hybrid approach to using LLMs (Sottolare et. al., 2017). GIFT is a modular framework designed to support and facilitate reuse, customization, or substitution of its individual components. These components include a learner model that includes a representation of the current knowledge state and skill level of the learner, a pedagogical module that represents learning approaches for different types of content and learner ability, and a domain module that represents the concepts, skills, and competencies to be imparted to the learners. There are other elements of GIFT that provide a user interface and the ability to connect to different devices, sensors, applications, and other services.

LLMs are currently being evaluated in a couple of different ways for implementation within GIFT. One involves using LLMs to classify team members' speech acts to support assessment of team performance (Spain et al., 2024). In this use-case, a team communication analysis framework has been incorporated into a demonstration involving crew gunnery. The team communication framework leverages an LLM through iterative prompting to classify team communications into one of nine categories of dialogue acts. These include categories such as: acknowledgements, providing information, and action requests. One advantage of this level of abstraction is that dialogue acts are generalizable across tasks and therefore easier to assess against theoretical team process models. Tests of this approach found that GPT-4 had an accuracy rate of about 77% as compared to human raters. While clearly not perfect, this is a significant step towards automation of a challenging task. As LLM's evolve, it will be interesting to see how quickly they get better at this kind of classification task. Once team speech acts are classified they can be assessed for errors and used to provide feedback to learners through GIFT's domain knowledge file, which uses context to determine which speech acts are needed by which individuals at various points in the scenario.

A second way in which LLM's are being considered for incorporation into GIFT is in support of a more conversational tutor, at least in terms of remediation feedback (Hu, 2023; Louwerse et al., 2009; Zapata-Rivera et. al., 2024). In this implementation, carefully structured course content would be delivered to learners through the existing GIFT framework. When a learner does not appear to understand a concept, the learner needs to be presented with some type of remediation. One of GIFT's primary adaptive feedback strategies is embodied in the Interactive, Constructive, Active, Passive framework (ICAP - Chi & Wylie, 2014). More cognitive engagement required in the active and constructive types of remediation is expected to produce better understanding and learning at the expense of taking more time than passive feedback (e.g., simply presenting information to the learner). Depending on the learner's characteristics, the complexity of the content, and the criticality of the training, different levels of engagement may be appropriate. GIFT seeks to find the optimal level of engagement necessary to train the content. It is not hard to see that in a course with many concepts, creating the necessary remedial material is a significant burden for course authors. The idea here is to use carefully constructed prompts to have an LLM assist in generating this content on demand.

In this particular use case, if a learner fails to understand a concept, GIFT would generate a set of prompts designed to solicit passive to increasingly more interactive types of feedback to help the learner master the concept sufficiently to move forward with additional content. By having GIFT—rather than the learner—generate these structured prompts, we can better control and constrain the LLM's responses. While this approach is still conceptual, this systematic prompt generation could help mitigate some of the known challenges with variability and hallucinations in current LLMs. The potential educational value of properly constrained, LLM-generated content makes this a promising area for investigation.

Future Research Directions

Future research, we argue, should investigate the factors that impact the effective implementation of this synergy with a careful consideration of the strengths and limitations of each, and through that develop

frameworks for the future design of pedagogical tools. This investigation should include systematic evaluation of LLM outputs across different educational contexts, as research has shown significant variations in performance across different domains and task types. This potential synergy also opens up a rich array of research questions, a few are listed below.

Integration mechanisms: How can we effectively integrate ITSs and LLMs to create a seamless learning experience that balances structured knowledge acquisition with creative exploration? What architectural frameworks would best support this integration? How can we systematically evaluate and mitigate the variability in LLM outputs when integrated into ITS frameworks? How can we implement comprehensive evaluation frameworks that assess both technical performance and pedagogical value? What role should automatic versus human evaluation play in different stages of development?

Impact on Learning & Cognition: How does exposure to both precise ITS instruction and more exploratory LLM interactions impact students' cognitive development, particularly in areas such as critical thinking, creativity, and metacognition? Further, what are the measurable impacts on both short-term knowledge acquisition and long-term conceptual understanding when utilizing this combined approach? How can we best leverage LLMs' semantic analysis capabilities while accounting for their limitations in direct assessment tasks? How can we systematically collect and incorporate student feedback and usage patterns to improve system effectiveness? What metrics best capture the actual impact on student learning in real-world settings?

Personalization strategies: How can we leverage the strengths of both technological approaches to create more nuanced and effective personalization strategies that address not just knowledge gaps, but also motivational and creative needs? What role should prompt engineering and output curation play in ensuring consistent, high-quality personalized learning experiences? How can we ensure responsible development practices while maintaining personalization capabilities? What safety frameworks and evaluation protocols are needed to protect student interests? How can we responsibly leverage LLMs capabilities in the generation of materials/activities/feedback to support personalization?

Domain-specific applications: How does the effectiveness of this combined approach vary across different academic disciplines? Are there certain subjects or skill areas where this synergy is particularly beneficial or potentially problematic? How can we optimize LLM performance for specific educational domains while maintaining pedagogical consistency?

Addressing these research questions would not only advance our understanding of AI in education but also provide crucial insights for the development of more effective, engaging, and holistic learning technologies that take advantage of the differential powers of ITSs and AI—combining the precision of ITSs with the creative potential of LLMs, ultimately enhancing both cognitive and motivational learning outcomes.

Acknowledgements

The research reported here was supported by the Arizona State University (ASU) Learning Engineering Institute and the Institute of Education Sciences, U.S. Department of Education, through Grant R305T240035 to Arizona State University. The opinions expressed are those of the authors and do not represent views of ASU, the Institute of Education Sciences, or the U.S. Department of Education.

References

- Al-kfairy, M., Mustafa, D., Kshetri, N., Insiew, M., & Alfandi, O. (2024). Ethical challenges and solutions of generative AI: An interdisciplinary perspective. *Informatics, 11* (3), 58. MDPI.
- Anthropic. (2024, May 21). *Mapping mind to language model*. <https://www.anthropic.com/news/mapping-mind-language-model>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).
- Benram, G. (2024, October 4). *Understanding the cost of Large Language Models (LLMs)*. TensorOps. <https://www.tensorops.ai/post/understanding-the-cost-of-large-language-models-llms>
- Buzby, H. (2024, June 25). *Breaking down the cost of large language models*. Qwak. <https://www.qwak.com/post/llm-cost>
- Cao, C. (2023). Leveraging large language model and story-based gamification in intelligent tutoring system to scaffold introductory programming courses: A design-based research study. *arXiv preprint arXiv:2302.12834*.
- Chen, Y., Ding, N., Zheng, H. T., Liu, Z., Sun, M., & Zhou, B. (2023). Empowering private tutoring by chaining large language models. *arXiv preprint arXiv:2309.08112*.
- Chi M.T. and Wylie R. (2014) The ICAP framework: linking cognitive engagement to active learning outcomes. *Educational Psychology, 49*(4), 219–243.
- Goldshtein, M., Ocumpaugh, J., Potter, A., Roscoe, R.D. (2024a). The social consequences of language technologies and their underlying language ideologies. *Lecture Notes in Computer Science, 271*-290. https://doi.org/10.1007/978-3-031-60875-9_18
- Goldshtein, M., Alhashim, A. G., & Roscoe, R. D. (2024b). Automating bias in writing evaluation: Sources, barriers, and recommendations. *The Routledge International Handbook of Automated Essay Evaluation, 421*-444. <https://doi.org/10.4324/9781003397618-25>
- Hu, X. (2023, July 3-7). Empowering education with LLMs - *The next-gen interface and content generation* [Conference demo]. International Conference on Artificial Intelligence in Education, Tokyo, Japan.
- Imundo, M. N., Watanabe, M., Potter, A. H., Gong, J., Arner, T., & McNamara, D. S. (2024). Expert thinking with generative chatbots. *Journal of Applied Research in Memory and Cognition*. Advance online publication. <https://dx.doi.org/10.1037/mac0000199>
- Jurenka, I., Kunesch, M., McKee, K. R., Gillick, D., Zhu, S., Wiltberger, S., Phal, S. M., Hermann, K., Kasenberg, D., Bhoopchand, A., Anand, A., Pisljar, M., Chan, S., Wang, L., She, J., Mahmoudieh, P., Rysbek, A., Ko, W.-J., Huber, A., ... Ibrahim, L. (2024). Towards responsible development of generative AI for education: An evaluation-driven approach.
- Kulik, J. & Fletcher, J. D. (2015). Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. *Review of Educational Research, 86*(1), 42-78. <https://doi.org/10.3102/0034654315581420>
- Louwerse, M.M., Graesser, A.C., McNamara, D.S., & Lu, S. (2009). Embodied conversational agents as conversational partners. *Applied Cognitive Psychology, 23*, 1244-1255.
- Maity, S., & Deroy, A. (2024). Generative AI and its impact on personalized intelligent tutoring systems. *arXiv preprint arXiv:2410.10650*.
- Mugaanyi, J., Cai, L., Cheng, S., Lu, C., & Huang, J. (2024). Evaluation of large language model performance and reliability for citations and references in scholarly writing: Cross-disciplinary study. *Journal of Medical Internet Research, 26*. <https://doi.org/10.2196/52935>
- Nye, B. D., Mee, D., & Core, M. G. (2023). Generative large language models for dialog-based tutoring: An early consideration of opportunities and concerns. *LLM@ AIED, 78*-88.
- Oli, P., Banjade, R., Tamang, L.J., Rus, V. (2023). The behavior of large language models when prompted to generate code explanations. *NeurIPS'23 Workshop on Generative AI for Education (GAIED)*.
- Pardos, Z. A., and Bhandari, S. (2023). Learning gain differences between ChatGPT and human tutor generated algebra hints. *arXiv preprint arXiv:2302.06871*. <https://doi.org/10.48550/arXiv.2302.06871>
- Rus, V. (2024). *Large language models and their implications for conversational tutors and GIFT*. Proceedings of the GIFT Symposium, 1-15.
- Schmucker, R., Xia, M., Azaria, A., & Mitchell, T. (2024, July). Ruffle & Riley: Insights from designing and evaluating a large language model-based conversational tutoring system. *International Conference on Artificial Intelligence in Education, 75*-90. Cham: Springer Nature Switzerland.

- Shahzad, M. F., Xu, S., & Zahid, H. (2024). Exploring the impact of generative AI-based technologies on learning performance through self-efficacy, fairness & ethics, creativity, and trust in higher education. *Education and Information Technologies*, 1-26.
- Sheese, B., Liffiton, M., Savelka, J., & Denny, P. (2024, January). *Patterns of student help-seeking when using a large language model-powered programming assistant*. Paper presented at the Australian Computing Education Conference (ACE), Sydney, NSW, Australia. <https://doi.org/10.1145/3636243.3636249>
- Shute, V. J., & Zapata-Rivera, D. (2012). Adaptive educational systems. In P. Durlach, & A. Lesgold (Eds.), *Adaptive technologies for training and education*. New York, NY: Cambridge University Press. 7-27.
- Sifaleras, A. (2024). *Generative intelligence and intelligent tutoring systems*. 20th International Conference, ITS 2024, Thessaloniki, Greece, June 10–13, 2024, proceedings, part I. Springer Nature.
- Sottolare, R. A., Brawner, K. W., Sinatra, A. M., & Johnston, J. H. (2017). An updated concept for a Generalized Intelligent Framework for Tutoring (GIFT). *GIFTtutoring.org*, 1-19.
- Spain, R., Min, W., Roberts, N., Kumaran, V., Pande, J., & Lester, J. (2024, August). Leveraging TCAT for Advanced Team Communication Analysis and Performance Assessment in GIFT. In *Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym12)* (p. 69).
- Warr, M., Oster, N., & Isaac, R. (2024a). Implicit bias in large language models: Experimental proof and implications for education. *Journal of Research on Technology in Education*, 0(0), 1–24. <https://doi.org/10.1080/15391523.2024.2395295>
- Warr, M., Pivovarova, M., Mishra, P., & Oster, N. J. (2024b). Is ChatGPT racially biased? The case of evaluating student writing. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4851112>
- Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*. <https://doi.org/10.48550/arXiv.2401.11817>
- Zapata-Rivera, D., Forsyth, C. M., Graf, A., & Jiang, Y. (2024). *Designing and evaluating evidence-centered design based conversations for assessment with LLMs*. Proceedings of EDM 2024 Workshop: Leveraging Large Language Models for Next Generation Educational Technologies.

CHAPTER 4 – THE USE OF LARGE LANGUAGE MODELS IN AN INTELLIGENT TEXT FRAMEWORK

Scott Crossley¹, Joon Suh Choi¹, Langdon Holmes¹, Wesley Morris¹, Vasile Rus², and Mihai Lintean³
Vanderbilt University¹; The University of Memphis²; EdClub, Inc.³

Introduction

Textbooks have historically served as valuable resources that enrich the learning experiences of students beyond what a teacher can orally present in a classroom. The digitization of textbooks has provided an efficient and widely accessible alternative to paper texts. However, reading digital texts engages readers in patterns that are different from reading printed texts. Studies indicate that some readers tend to read slower when reading from a screen (Dillon, 1992) and engage in a more shallow, fragmented reading behavior powered by skimming and keyword spotting (Dyson & Haselgrove, 2000; Horton et al., 1995; Levy, 1997).

Advances in artificial intelligence (AI) through the use of large language models (LLMs) have led to the next revolution in textbooks: intelligent texts.¹ Intelligent texts include interactive, natural-language-processing-based features that can make the learning process more dynamic. An early version of an intelligent textbook was *Inquire Biology* (Chaudhri et al., 2013). Besides the basic features of an electronic textbook, such as key term definitions, note taking or image interaction, *Inquire* used an early version of an AI reasoning system and a manually curated knowledge base to answer logic questions about the content of the book, as well as suggest verbatim questions (i.e., questions for which the answer is available verbatim in the source text) to the reader for review. More recently, researchers have started to develop conversational tutoring systems that can answer questions about texts using LLMs. *EduQuiz* (Dijkstra et al., 2022) is such a system and was designed to generate context-dependent quizzes from reading comprehension texts. The quizzes comprised multi-choice questions with correct and distractor answers and were generated by a GPT-3 model that was fine-tuned on text-quiz pairs. Another recent example of intelligent texts that focus on question generation is *Explanatory AI for Education* (YAI4Edu; Sovrano et al., 2024), which enhances textbooks through automatic question generation and explanatory answers. Following a similar pathway to *EduQuiz*, YAI4Edu uses Text-to-Text Transfer Transformers (T5; Raffel et al., 2020) to extract specialized knowledge graphs from texts to identify the most relevant questions from the content. Explanatory answers are automatically extracted from the books with the help of the knowledge graph and the pre-trained deep language model. YAI4Edu generates verbatim questions and primary archetypal questions as well as questions of secondary archetypes that make use of primary archetypes (i.e. the system will ask for examples, manners, substitutions, and origins). *ExplanatoryGPT* (Sovrano et al., 2023) also focuses on question-answering using LLMs but provides additional guardrails. The system is seen as an enhancement over simply using ChatGPT or other LLMs because it provides interactive and user-centered explanations to textbook related questions while mitigating hallucinations and memory shortcomings common in early ChatGPT versions.

¹ A historical review of using AI to enhance textbooks in order to make them more interactive, engaging, adaptive, and intelligent was reported in Brusilovsky, Sosnovsky, and Thaker, (2022).

Intelligent Texts for Enhanced Lifelong Learning (iTELL) Framework

More advanced AI technologies have also been explored to augment digital textbooks that enable levels of interactivity, engagement, and adaptation beyond just question generation. Examples of these intelligent text frameworks include the Intelligent Texts for Enhanced Lifelong Learning (iTELL) framework (described in depth in this chapter) and the Ruffle&Riley system. The Ruffle&Riley system builds interactive tutorial dialogue around AI-generated “review questions” from digital lesson texts (Schmucker et al., 2024). Students are prompted to answer these review questions, and scaffolding and feedback is provided as needed in an interactive, student-adapted manner. Inspired from traditional expectation misconception tailoring (EMT) tutoring approaches (Graesser et al., 2004), a list of expectations is generated for each question within the system. Potential misconceptions are handled on-the-fly and rely on LLMs to detect and remediate factually incorrect information.

Given recent developments in intelligent textbooks, this chapter introduces the iTELL framework. iTELL is an advanced framework that can convert any type of machine-readable text into interactive, intelligent text powered through AI within a web-app. We discuss how LLMs in iTELL provide reading interventions and feedback to learners and current steps to ensure the system is adaptive to learners’ backgrounds and skills. Lastly, we discuss lessons learned and how intelligent textbooks can be adapted into both the GIFT (Sottolare et al., 2017) and STEEL-R (Goldberg, 2020) frameworks to increase learning opportunities for Warfighters.

iTELL converts traditional learning materials to intelligent texts using a content authoring system. As part of this process, learning materials are split into pages and chunks, where a chunk is a subsection within a page usually under a single sub-header, most often consisting of 1-3 paragraphs of text or a single instructional video. From this structured format, iTELL generates learning activities based on theories of reading comprehension (Self Explanation Reading Training; McNamara, 2004, 2017) and cognitive engagement (the Interactive, Constructive, Active, and Passive framework; Chi & Wylie, 2014) and provides opportunities for users to generate knowledge about what they read and watch through constructed responses items (CRIs) and summary writing. The CRIs and summaries are scored automatically by LLMs specifically trained to generate scores which inform qualitative feedback to students (Morris et al., 2024a).

For instance, a constructed response item is generated for each chunk using GPT-3.5-turbo with human-in-the-loop. Each chunk in an iTELL volume has a 1/3 chance of spawning an accompanying constructed response item for a reader, with a minimum of one constructed response item per page. Users are required to submit at least one response to a spawned item before proceeding to the next chunk. Readers’ constructed responses are scored for correctness using two separate fine-tuned LLMs, Bilingual Evaluation Under-study with Representations from Transformers (BLEURT, Sellam et al. 2020) and Masked and Permuted Language Modeling (MPNet, Song et al., 2020), both of which report an accuracy of ~ .80 (Morris et al., 2024b) on question/answer pairs in the Multi-Sentence Reading Comprehension (MultiRC) dataset (Khashabi et al., 2018). If both models agree that the construct response is correct, it is scored correct. If both models agree that the constructed response is incorrect, it is scored incorrect. If the models disagree, the constructed response is scored partially correct. These models are also used to provide feedback to readers when re-answering CRIs that they answered incorrectly.

For summaries, iTELL readers receive feedback on Language Borrowing by calculating the proportion of overlapping trigrams between the summary and the source (i.e., containment; see Broder, 1998). Students also receive feedback on Relevance using cosine similarity between the text embedding of the summary and the text embedding of the source. Once the summary is judged to not be plagiarized and relevant, it is scored by a single LLM introduced in Morris et al. (Morris et al., 2024c; Morris et al., 2023) on Content (i.e., whether the summary reproduces the core contents of the source). iTELL also includes an AI-powered

chatbot that is based on Llama 3 with retrieval augmented generation (RAG) to ensure answers are on topic and appropriate prompting to enact guardrails against cheating and other types of misuse. The chatbot is always available to students to activate if they have questions about text-related material or the iTELL system. The chatbot also affords users the opportunity to make inferences and elaborate on the text using self-explanations during re-reading exercises. The feedback from the AI tools integrated into iTELL is used in several different ways, including to guide learning, correct misconceptions, review missed topics, prepare for upcoming materials, make links between the texts and the real world, and help elaborate on what users have learned.

iTELL is specifically designed to support adult learners. Common themes that characterize the adult learner are that adults benefit from greater autonomy and agency in their learning experiences (Knowles, 1975; Rogers, 1994; Smith, 1983). iTELL supports adult learners to engage in self-directed learning (SDL) through self-managing, self-monitoring, and self-modifying (Costa & Kallick, 2004). iTELL is autonomous and requires no human oversight, allowing users to access the system on their schedule, mirroring the self-paced nature of traditional reading materials (self-management). iTELL further supports SDL by creating opportunities for learners to reflect on their reading process via think-aloud activities and to explain their understanding via self-explanation questions (self-monitoring). Autonomous learning is supported by providing learners with instant feedback throughout the reading process, with features that allow learners to optionally seek detailed feedback (self-modification). Additionally, iTELL's provision of short, directed, and concrete online tasks can be helpful in assisting learners who are not fully capable of self-directed learning become learners with higher self-direction (Fidishun 2000; Grow, 1991). Distinct learning behaviors and needs are accommodated through interventions that are activated adaptively based on learner profiles within iTELL. Concretely, iTELL supports a learner-centered approach by meeting most of the recommendations on accommodating different learning preferences in an online learning environment. This includes ensuring students can review materials at will, ensuring materials are provided in multiple modes, allowing enough time for students to master the content and using strategies such as guided questioning (Cercone, 2008).

iTELL has been tested in a variety of classrooms and vocational educational settings for adult learners with a focus on community college students and students in more traditional university settings that focused on computer science, data science, or STEM related classes. Analyses of iTELL data show iTELL leads to learning gains for students. For instance, reading time analyses indicate that readers who spent more time engaged with an iTELL text tended to score better on constructed response items, to write summaries that were more relevant, and to engage in re-reading (Choi et al., 2024). User experience data indicates that students' experiences with the AI tools within iTELL are positive. Overall, students reported that the CRIs and summary tasks were easy to work with and helped them improve their learning. Students also reported that the AI feedback they received was generally accurate and that overall satisfaction with iTELL was high (Crossley et al., 2024; Morris et al., 2024a). An A/B test of the iTELL system reported a small but meaningful relationship between pre- and post-test score differences for students that used an iTELL text version versus those that used a digital version. The mean score differences indicated that students that used iTELL showed gains of ~5% versus students that did not. Follow up analyses indicated that iTELL may lead to greater learning gains for lower proficiency students (Crossley et al., 2024).

As mentioned earlier, AI-powered chatbots are integrated into iTELL and data analyses are underway to assess how engaging readers in self-explanations using a dialogic chatbot called the Strategic Thinking And Interactive Reading Support (STAIRS 1.0) system can increase learning gains. Self-explanation is the process of explaining aloud the meaning of a text through paraphrasing, elaboration, logic, prediction, or bridging and can heighten literacy processes and improve reading comprehension (Chi et al., 1994, Duke & Pearson, 2009). In iTELL, when a reader shows low comprehension of text based on summary scores, that reader is asked to re-read a text chunk based on interaction data within the system. After re-reading, a chatbot guides the reader through a self-explanation of the text to engage that reader in strategies that

enhance comprehension. STAIRS 1.0 makes the iTELL reading experience more interactive and introduces readers to strategies that can improve text comprehension in their future reading. In a recent randomized controlled trial, we found no significant improvement in content scores for first and second summary attempts for participants who did not receive STAIRS support, while participants who received STAIRS support reported a significant and meaningful improvement in content scores on their second summary attempt.

Decoder-only LLMs (Llama 3.1 8B, Dubey et al., 2024) are also being integrated into CRIs. If a user answers a CRI incorrectly such that both LLM scoring models report that the answer is wrong, STAIRS 2.0 (Stimulated Think Aloud for Interactive Reading Support) will trigger. STAIRS 2.0 is inspired by previous work utilizing think-alouds in classroom settings (Oster, 2001). In a think-aloud, learners vocalize their thoughts as they read, which produces artifacts of their cognitive processes. These artifacts can be utilized to analyze and reflect upon learners' cognitive processes. STAIRS 2.0 is a think-aloud exercise in which a dialogic agent models the think-aloud process for the user and engages the user in a dialogue based on theories of reading comprehension about their own cognitive processes as they are re-reading. This exercise is designed to promote metacognitive awareness of the reading process and self-regulated learning. The think-aloud process is modeled by re-generating the chunk as a text stream (token stream) in which the dialogic agent interjects "think-alouds," which are thoughts about the reading that model the cognitive processes used by effective readers. After the think-aloud process is modeled, iTELL AI asks the learner to write what they were thinking as they read the chunk and engages the user in a structured dialog intended to encourage further reflection about effective reading strategies.

The next steps in the development of iTELL using LLMs will include the integration of two new interventions using reduced redundancy tasks in which users are required to complete a summary of the material with some of the words deleted or gapped (i.e., partially deleted). These tasks are designed to construct a learner model of the user in order to calibrate personalization systems. The first task is a C-test (Klein-Bradley, 1985) which users complete at the beginning of each text. This task has two purposes. First, it is designed to assess student reading proficiency at the beginning of an iTELL deployment so that the iTELL learner's record store can be updated early and interventions personalized to reading ability. The second purpose is to provide readers with an early overview of the reading material and to ask them to make predictions about what they are going to read, which is an important strategy for successful reading comprehension (Smith, 1975). LLMs will be used for automatic rational gap selection where LLM predictions of Item Response Theory parameters will optimize the amount of assessment information with a minimal number of gaps (Yancey et al., 2024). In addition to the use of C-tests at the start of each text, we also plan on implementing an alternative to summary writing that is based on a fill-in-the-blanks approach (i.e., a cloze test). For the cloze tests, we will use LLMs to automatically identify key words and phrases to gap in order to assess the users' comprehension of information from the text (Yang et al., 2021). Additionally, we will develop automatic semantic scoring metrics so that users will receive full or partial credit for responses that are semantically correct even when they are not exact matches to the original text (Kleijn et al., 2019).

Recommendations for GIFT and STEEL-R Overall

iTELL has much in common with the Generalized Intelligent Framework for Tutoring (GIFT; Sottolare et al., 2017). Like GIFT, iTELL goes beyond the functions of a traditional shell tutor and can provide adaptive content and instructions for any domain. The content management system employed by iTELL allows content creators to quickly and easily generate modules on any content domain. Also, like GIFT, iTELL's tutor-user interface uses short- and long-term learner modeling and raw user input to inform module-decisions. For instance, iTELL's evaluation module decides whether STAIRS 1.0 is activated based on a

learner's current learning state derived from the learner's current raw input as well as their previous learning activity logs. Based on the learner model and the decision made by the evaluation module, the feedback module then produces and streams domain-relevant questions designed to scaffold the learner's think-aloud process, adaptively targeting the most relevant portion of the text for the learner to re-read. Unlike STEEL-R, iTELL does not use a learning management system and all telemetry data are stored in a Supabase instance that constantly interacts with the iTELL frontend to log data using a REST API (Application Programming Interface) formatted in JSON (JavaScript Object Notation). For all LLM interactions including summary scoring, constructed response scoring, and STAIRS, iTELL's frontend makes API calls to iTELL along with API calls to an external server with GPU (Graphics Processing Unit) capacities powered by FastAPI and vLLM. The resulting data, which includes scoring model results, generated feedback, key words, and chatbot dialogs are stored in the Supabase instance to constitute the learner model.

iTELL (or a similar intelligent text application) could be integrated within GIFT and STEEL-R to increase the acquisition of skills within the Warfighter community through AI interventions. In particular, iTELL could assist with training processes where digital texts are used (textbooks, training manual, and other training documents). We have found that reading interventions (e.g., summaries and constructed responses) and their accompanying AI feedback can lead to increased learning gains. Proficiency in these tasks is also related to reading time itself, providing predictive validity for the approach. Users also indicate that the reading interventions help them better understand the text.

Thus, our primary recommendations for the GIFT and STEEL-R initiatives are to integrate LLMs that can be used to provide feedback to readers about their comprehension processes. Additionally, the ability to incorporate AI powered chatbots that can engage readers in conversations about the text they are reading, including the ability to help them self-explain their reading, would be a powerful addition to the GIFT and STEEL-R initiatives.

Conclusions

This chapter has introduced how LLMs can be used within intelligent textbooks to provide interactive feedback to learners that should make text more engaging and motivating. The chapter provides a deep dive into the iTELL framework and demonstrates how LLMs within the framework can be used to provide feedback to students about short answer responses and about the quality of text summarizations. Additional LLM interventions within iTELL that leverage dialogic AI chats are also discussed. Future directions for the iTELL framework and for intelligent texts in general are provided with a focus specifically on recommendations for GIFT and STEEL-R. Overall, strong links are provided between GIFT and STEEL-R that should allow for the adoption of intelligent texts within a general tutoring framework and learning management system.

References

- Broder, A. Z. (1998). On the resemblance and containment of documents. *Proceedings Compression and Complexity of SEQUENCES 1997* (Cat no 97TB100171), 21–29.
<https://doi.org/10.1109/SEQUEN.1997.666900>.
- Brusilovsky, P., Sosnovsky, S., & Thaker, K. (2022). The return of intelligent textbooks. *AI Magazine*, 43(3), 337-340.
- Cercone, K. (2008). Characteristics of adult learners with implications for online learning design. *AACE review*, 16(2), 137-159.

- Chaudhri, V. K., Cheng, B., Overholtzer, A., Roschelle, J., Spaulding, A., Clark, P., Greaves, M., & Gunning, D. (2013). Inquire Biology: A Textbook that Answers Questions. *AI Magazine*, 34(3), 55-72. <https://doi.org/10.1609/aimag.v34i3.2486>
- Chi, M. T., De Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive science*, 18(3), 439-477.
- Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist* 49 (4), 219-243.
- Choi, J., Morris, W., Holmes, L., Crossley, S. (2024). Focus Time and Writing Performance in an Intelligent Textbook. *Proceedings of the 17th International Conference on Educational Data Mining (EDM)*. Atlanta, GA.
- Costa, A. L., & Kallick, B. (2004). Launching self-directed learners. *Educational leadership*, 62, 51-57.
- Crossley, S. A., Choi, J. S., Morris, W., Holmes, L., Joyner, D. & Gupta, V. (2024). Using Intelligent Texts in A Computer Science Classroom: Findings from an iTELL Deployment. Proceedings of 8th Educational Data Mining in Computer Science Education Workshop (CSEDM 2024) at the 17th International Conference on Educational Data Mining (EDM). Atlanta, GA.
- Duke, N. K., & Pearson, P. D. (2009). Effective practices for developing reading comprehension. *Journal of education*, 189(1-2), 107-122.
- Dijkstra, R., Genç, Z., Kayal, S., & Kamps, J. (2022). Reading Comprehension Quiz Generation using Generative Pre-trained Transformers. *Proceedings of the Fifth International Workshop on Intelligent Textbooks*.
- Dillon, A. (1992). Reading from paper versus screens: A critical review of the empirical literature. *Ergonomics*, 35(10), 1297-1326.
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Srivankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., ... Zhao, Z. (2024). The Llama 3 Herd of Models (arXiv:2407.21783). arXiv. <https://doi.org/10.48550/arXiv.2407.21783>
- Dyson, M., & Haselgrove, M. (2000). The effects of reading speed and reading patterns on the understanding of text read from screen. *Journal of Research in Reading*, 23(2), 210-223.
- Fidishun, D. (2000). Teaching Adult Students to Use Computerized Resources: Utilizing Lawler's Keys to Adult Learning to Make Instruction More Effective. *Information Technology and Libraries*, 19(3), 157.
- Goldberg, B (2020). STE Experiential Learning-Readiness (STEEL-R) Strategy. STTC White Paper.
- Graesser, A. C., Lu, L., Jackson G. T., Mitchell, H., Ventura, M., Olney, A., & Louwerse, M. (2004). Autotutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, 36, 180–192.
- Grow, G. O. (1991). Teaching Learners To Be Self-Directed. *Adult Education Quarterly*, 41(3), 125–149. <https://doi.org/10.1177/0001848191041003001>
- Horton, W., Taylor, L., Ignacio, A., & Hoft, N. L. (1995). *The Web page design cookbook: All the ingredients you need to create 5-star web pages*. John Wiley & Sons, Inc.
- Khashabi, D., Chaturvedi, S., Roth, M., Upadhyay, S., & Roth, D. (2018). Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.
- Kleijn, S., Pander Maat, H., & Sanders, T. (2019). Cloze testing for comprehension assessment: The HyTeC-cloze. *Language Testing*, 36(4), 553–572. <https://doi.org/10.1177/0265532219840382>
- Klein-Braley, C. (1985). A cloze-up on the C-Test: A study in the construct validation of authentic tests. *Language Testing*, 2(1), 76–104. <https://doi.org/10.1177/026553228500200108>
- Knowles, M. S. (1975). *Self-directed learning: A guide for learners and teachers*. The Adult Education Company.
- Levy, D. M. (1997, July). I read the news today, oh boy: Reading and attention in digital libraries. In *Proceedings of the Second ACM international Conference on Digital libraries* (pp. 202-211).
- McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes*. 38 (1), 1-30.
- McNamara, D. S. (2017). Self-explanation and reading strategy training (SERT) improves low-knowledge students' science course performance. *Discourse Processes*. 54 (7), 479-492.
- Morris, W., Crossley, S., Holmes, L., Ou, C., McNamara, D., & Dascalu, M. (2023). Using Large Language Models to Provide Formative Feedback in Intelligent Textbooks. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky* (Vol. 1831, pp. 484–489). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36336-8_75

- Morris, W., Choi, J., Holmes, L., Gupta, V., & Crossley, S. A. (2024a). Automatic Question Generation and Constructed Response Scoring in Intelligent Texts. *Workshop on Leveraging Large Language Models for Next Generation Educational Technologies at the 17th International Conference on Educational Data Mining (EDM)*. Atlanta, GA..
- Morris, W., Crossley, S., Holmes, L., Ou, C., Dascalu, M., & McNamara, D. (2024b). Formative Feedback on Student-Authored Summaries in Intelligent Textbooks Using Large Language Models. *International Journal of Artificial Intelligence in Education*. <https://doi.org/10.1007/s40593-024-00395-0>
- Morris, W., Holmes, L., Choi, J. S., & Crossley, S. (2024c). Automated Scoring of Constructed Response Items in Math Assessment Using Large Language Models. *International Journal of Artificial Intelligence in Education*. <https://doi.org/10.1007/s40593-024-00418-w>
- Oster, L. (2001). Using the Think-Aloud for Reading Instruction. *The Reading Teacher*, 55(1), 64–69.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140), 1-67.
- Rogers, M. (1994). Learning about global futures: An exploration of learning processes and changes in adults. D. Ed. Thesis, University of Toronto, Toronto, ON, Canada, 1994.
- Schmucker, R., Xia, M., Azaria, A., Mitchell, T. (2024). Ruffle & Riley: Insights from Designing and Evaluating a Large Language Model-Based Conversational Tutoring System. In: Olney, A.M., Chounta, IA., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds) *Artificial Intelligence in Education. AIED 2024. Lecture Notes in Computer Science*, vol 14829. Springer, Cham. https://doi.org/10.1007/978-3-031-64302-6_6
- Sellam, T., Das, D., & Parikh, A. P. (2020). BLEURT: Learning Robust Metrics for Text Generation (Version 5). arXiv. <https://doi.org/10.48550/ARXIV.2004.04696>
- Smith, F. (1975). The role of prediction in reading. *Elementary English*, 52(3), 305-311.
- Smith, R. M. (Ed.). (1983). *Helping adults learn how to learn*. Jossey-Bass.
- Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). MPNet: Masked and Permuted Pre-training for Language Understanding (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2004.09297>
- Sottolare, R. A., Brawner, K. W., Sinatra, A. M., & Johnston, J. H. (2017). An updated concept for a Generalized Intelligent Framework for Tutoring, R. A., Goldberg, B. S., Brawner, K. W., & Holden, H. K. (2012). A modular framework to support the authoring and assessment of adaptive computer-based tutoring systems (CBTS). In *Proceedings of the Interservice/Industry Training, Simulation, and Education Conference* (pp. 1-13).
- Sovrano, F., Ashley, K., & Bacchelli, A. (2023). Eliminating Hallucinations: GPT-based Explanatory AI for Intelligent Textbooks and Documentation. *Proceedings of the Fifth International Workshop on Intelligent Textbooks*.
- Sovrano, F., Ashley, K., Brusilovsky, P.L., & Vitali, F. (2024). How to Improve the Explanatory Power of an Intelligent Textbook: A Case Study in Legal Writing. *International Journal of Artificial Intelligence in Education*. <https://doi.org/10.1007/s40593-024-00399-w>
- Yancey, K. P., Runge, A., LaFlair, G., & Mulcaire, P. (2024). BERT-IRT: Accelerating Item Piloting with BERT Embeddings and Explainable IRT Models. In E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, & Z. Yuan (Eds.), *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 428–438). Association for Computational Linguistics. <https://aclanthology.org/2024.bea-1.35>
- Yang, A. C. M., Chen, I. Y. L., Flanagan, B., & Ogata, H. (2021). Automatic Generation of Cloze Items for Repeated Testing to Improve Reading Comprehension. *Educational Technology & Society*, 24(3), 147–158.

CHAPTER 5 – USING GENERATIVE AI TO IMPROVE STUDENT’S WRITING

William Swartout^{1,2}, Paul Deane³, and Benjamin Nye^{1,2}

USC Institute for Creative Technologies¹; USC Center for Generative AI and Society²;
Educational Testing Service³

Introduction

The release of ChatGPT on November 30, 2022 represented a breakthrough in artificial intelligence (AI) and at the same time a major disruption. For the first time, it provided an easy-to-use system that had broad and robust capabilities for natural language generation and understanding. It could competently create essays on almost any topic, and although it had some significant limitations, such as a tendency to occasionally make things up, it was widely seen as something that would significantly impact how people work. The education community was taken aback since students now had at their disposal a “homework machine” (Williams & Abrashkin, 1958) that could quickly produce plausible responses to writing assignments (Jimenez, 2023). Educators were concerned that students would use Generative AI to cheat on assignments and some school districts banned access to ChatGPT from their networks and began to use tools designed to detect if assignments were written by Generative AI.

In examining Generative AI and education, we have become convinced that the “deny and detect” approach to dealing with Generative AI is short-sighted and counterproductive. There are two reasons for this. First, the tools to detect whether text has been written by Generative AI are not generally reliable. Although some studies have reported that a few detectors are 100% accurate (Walters, 2023), other studies show low accuracy for the same detectors on different datasets (Caulfield, 2023). Furthermore, the detectors have both false positive and false negative error rates. False positives are particularly problematic because they risk accusing a student of cheating when they did not. As Generative AI becomes more sophisticated, it is likely to be even more difficult to detect accurately. Second, and perhaps more importantly, if we deny students the use of Generative AI while they are in school and then graduate them into a world that will use it extensively, we will not have prepared students for their future.

In contrast to this “deny and detect” group of educators, another group emerged which might be called the “embrace and enhance” group. This group sought ways in which education could be re-structured so that Generative AI could be used in novel ways to improve student’s critical thinking rather than suppressing critical thinking as can occur if students rely too much on Generative AI to do the work for them.

In this chapter, we describe two writing tools, 1) an Outlining/Planning tool developed at the ETS Research Institute and 2) the Army Writing Enhancement (AWE) tool developed at the USC Institute for Creative Technologies. Both tools are aimed at using Generative AI to support students as they write papers, but not in the manner one might expect. Rather than using Generative AI to create first drafts of assigned writing exercises these tools use Generative AI to help students think about their paper and what content should be included and then to analyze and critique the student’s paper as it takes shape. We begin describing the two tools, their approach to writing and examples of their use. We follow that with a discussion of the Generative AI prompts that were required to make the tools work. As we performed this research, one of the things we discovered was that simple, naïve prompts did not produce useful results. As we will describe, the prompts that work are quite lengthy and need to specify the desired output precisely. At the same time, these prompts are quite general in the sense that they largely work across different large language models (LLMs) and over a broad range of domains that the essays may be about.

Connections To Related Work

Research on the ETS Outlining/Planning and AWE for writing enhancement represents a small part of a broader ecosystem of research on how LLM tools can improve writing. These tools fall into a few distinct categories, which include: writing assistance (e.g., grammar correction), broader automated writing evaluation, interactive coaching, and improving peer feedback.

Research on automated writing evaluation with earlier language models (pre-GPT 3.5) generally shows improvements on a specific essay. However, less evidence was available to determine where these build generalizable skills and, concerningly, in some cases the duration of the intervention (e.g., dosage) did not impact benefits (Zhai & Ma, 2023). That said, research such as Butterfuss et al. (2022) found that interactive coaching targeted to student needs did improve only the specific strategies trained for their revised paper, which is a better indicator of learning gains. A meta-analysis which distinguished between post-task performance and a follow-up task showed significant benefits for the post-task, $r = .57$, but only non-significant positive benefits on follow-up tasks, $r = .27$ (Fleckenstein et al., 2023). That study also found greater effect sizes with higher duration. Across both meta-analyses, the authors suggest that automated writing evaluation support has a high degree of variation and that further work is needed to determine more effective categories of tools. Emerging tools which use LLM dialogs to support writing are still pending evaluation, with even highly visible tools such as Khanmigo having only limited evidence of benefits (Shetye, 2024).

This indicates that a deeper analysis and categorization is likely required which aligns to work such as VanLehn's meta-analysis on intelligent tutoring systems (VanLehn, 2011), which indicated that more granular tutoring produced greater learning gains, such as step-based systems (e.g., hints, breaking down problems) outperforming answer-based systems (e.g., corrective feedback). Similarly, stronger gains should be expected from systems which engage learners in deeper thinking and structured help on their writing, as opposed to only corrective feedback on lower-level grammar or structure issues. Emerging research on LLMs has shown benefits for both AI feedback directly (Meyer et al., 2024) and for using AI to help students provide better peer feedback (Schultze et al., 2024). However, further research is needed to categorize and compare these types of feedback.

When considering this broader ecosystem, the tools presented below can be categorized as including:

- **Interactive Coaching:** A step-by-step conversation leads the learner through specific activities related to their writing. The conversation is not characterized by a single LLM prompt, but by a combination of multiple prompts which are combined by a structured workflow or agents. The system responds based on the state of a) the writing product and b) the user's inputs.
- **Multi-Criteria Formative Feedback:** The learner does not receive a single, immediately-actionable suggested change, but instead the student must make a personal evaluation of their priorities about which AI feedback and criteria align to their writing goals. The formative feedback and AI coaching are intended to provide modeling and scaffolding (e.g. cognitive steps or evaluation criteria to consider, or steps of a strategy to practice).
- **User Contribution & Reflection:** The learner is prompted explicitly to plan or discuss their writing product, separately from modifying the writing directly. The learner is expected to make meaningful changes about the content and purpose of the writing, rather than lower-level changes that only rephrase existing ideas.

By contrast, in the work below, we are not focused on other common types of writing support, such as extensive grammar assistance, automated summative grading (e.g., high-quality numerical estimates, potentially against a multi-criteria rubric), or explicit support for multilingual learners. These other types of tools are likely to benefit learners as well, but in this work we are primarily interested in the interconnection between writing and thinking (two-way process), rather than the translation from thinking to writing (one-way process).

An Interactive Planning Tool: Using Generative AI to Scaffold Writing Processes

While Generative AI can be used in open-ended chats, to provide whatever response the user requests, open-ended applications depend, to a lesser or greater extent, on the user knowing what questions to ask. This can be problematic in a tutorial context, especially for complex performance tasks like writing, where the development of expertise requires users to internalize a complex panoply of strategies that help them organize different stages of the writing process. One way to solve this problem is to create an AI-powered tool that scaffolds strategy use. This approach can be illustrated using an AI-powered interactive outlining tool developed at the ETS Research Institute. This tool is designed to walk students through the process of creating an outline for an argument essay, using Generative AI to produce both examples and feedback. Figure 1 shows basic features of this tool’s interface.

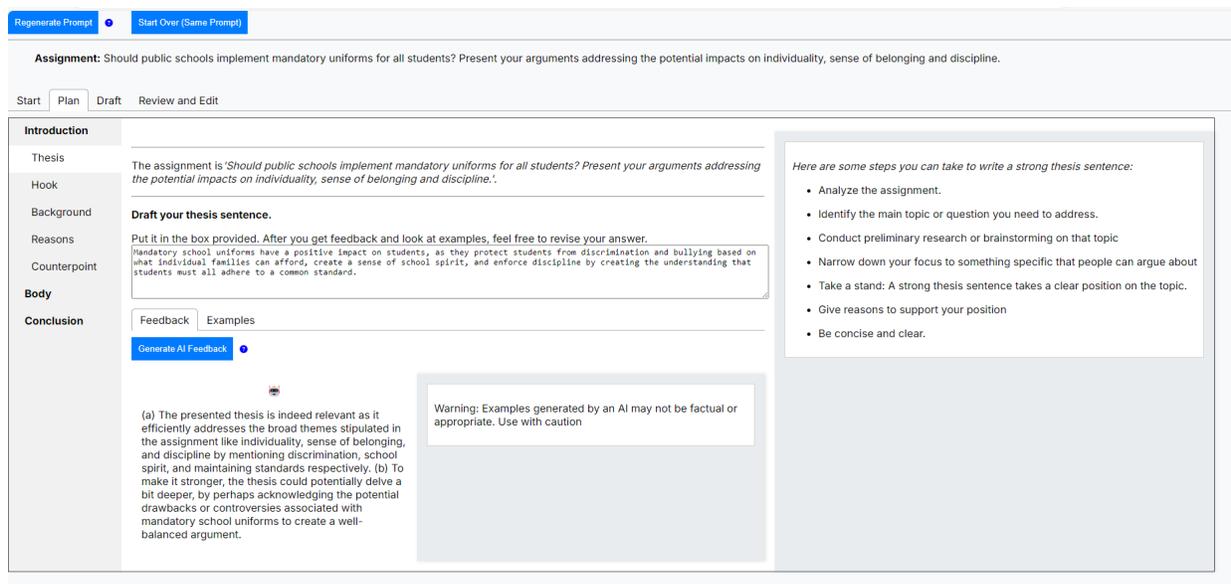


Figure 1. A screen from the interactive outlining tool, supporting the writing of a thesis sentence.

As Figure 1 illustrates, users have freedom to navigate to different parts of the outline (introduction/body/conclusion, or within the introduction, to key parts like the thesis and the hook, or lead-in). Once at a specific location in the outline, such as the thesis, they can draft their own text and obtain AI feedback. In deliberate practice mode, the tool generates random assignments and allows students to use AI to generate examples of specific writing strategies for different parts of the outline. This approach relies on prompt engineering behind the scenes and captures detailed logs of student actions for review (including records of cut-and-paste actions and use of AI to generate examples in deliberate practice mode). The tool is intended to be used under teacher direction, with teachers controlling when students access the tool and in which mode.

Prompt Design for The ETS Writing Tool. We have found that in critiquing, the design of the prompt that produces the critique is critical. A simple naïve prompt, such as “Critique my essay suggesting counter-arguments to my main thesis” may produce some results, but they are likely to be lower quality (e.g., vague) or inconsistent in their focus. The prompt needs to be designed to essentially lead the LLM through the process of creating the critique, specifying at some detail what should be included and what should be left out. In many ways, a well designed prompt or set of prompts represents a cognitive process: specific cues to consider, a chain of thought for breaking down the problem, and the details of what should be included in the output.

For example, the feedback shown in Figure (1) was generated by the following prompt:

```
{
  "role": "system",
  "content": "I am your writing coach. How can I help you?"
},
{
  "role": "user",
  "Content":
    "The assignment is 'Should public schools implement mandatory uniforms for all students? Present your arguments addressing the potential impacts on individuality, sense of belonging and discipline.'

    The thesis sentence is: 'Mandatory school uniforms have a positive impact on students, as they protect students from discrimination and bullying based on what individual families can afford, create a sense of school spirit, and enforce discipline by creating the understanding that students must all adhere to a common standard.'

    If the thesis sentence is ", say so and stop writing. Otherwise, evaluate whether the thesis statement is relevant to the topic of the assignment and if it is not, say so and stop writing. Otherwise, (a) In a single sentence, briefly evaluate how appropriate the thesis sentence is to the assignment, and (b) In a second sentence, suggest what kinds of revisions might make the thesis sentence stronger. Follow these instructions, but do not talk about them: Do not mention the need to provide reasons. Do not talk about mentioning the reason to provide reasons. Do not provide example revisions. Do not mention examples or evidence. Do not discuss how your response conforms to these directions."
}
```

This prompt was constructed by combining information collected from different parts of the user interface. The assignment was generated by clicking the “Regenerate Prompt” button, which ran a different prompt through ChatGPT. This prompt asked ChatGPT to generate a list of 300 possible writing prompts, and then select one randomly. This strategy was necessary to overcome the model’s tendency to favor a small list of stereotypical prompts, if asked to generate a single prompt directly. The thesis sentence was generated by the user, who typed it into the input box shown in Figure (1). The rest of the prompt is a template into which the assignment and the thesis sentence are inserted, and reflects prompt engineering designed to produce appropriate feedback, while suppressing ChatGPT’s tendency to produce responses that are too long or include references to aspects of its instructions that need to be kept in the background. A critical part of this design is that the context changes depending on where the user is in the outline. For example, if the user has written a thesis and provided supporting reasons, when they navigate to the body section of the outline, this information will be used to provide context when the AI needs to generate content or feedback for specific body paragraphs (see Figure 2):

Regenerate Prompt Start Over (Same Prompt)

Assignment: Should public schools implement mandatory uniforms for all students? Present your arguments addressing the potential impacts on individuality, sense of belonging and discipline.

Start Plan Draft Review and Edit

Introduction		
Body		An essay typically contains 4-7 body paragraphs, corresponding to important supporting reasons and counterpoint/rebuttal.
Paragraph 1	The assignment is 'Should public schools implement mandatory uniforms for all students? Present your arguments addressing the potential impacts on individuality, sense of belonging and discipline.'	<ul style="list-style-type: none"> You need to decide what evidence you must provide to convince people to accept your argument in each paragraph. One way to do that is to come up with a list of questions that you need to answer.
Support	The thesis sentence is: 'Mandatory school uniforms have a positive impact on students, as they protect students from discrimination and bullying based on what individual families can afford, create a sense of school spirit, and enforce discipline by creating the understanding that students must all adhere to a common standard.'	
Paragraph 2		
Paragraph 3		
Paragraph 4	The topic sentence is: 'If current policies on school uniforms are not changed, peer pressure to wear the latest or most expensive fashion trends could increase, leading to unnecessary competition and distraction from learning.'	
Paragraph 5		
Detailed Rebuttal	What should my evidence be? Questions I need to answer.	
Conclusion	Put your questions in the boxes provided. After you get feedback and look at examples, feel free to revise them.	
	<input type="text"/>	
	<input type="text"/>	

Figure 2. Screen designed to support body paragraph development, illustrating context inheritance from the thesis and reason screens.

In this example, the assignment, thesis sentence, and topic sentence are embedded in the AI prompts for this screen, helping to focus the feedback appropriately. Another important feature of the tool is that it is designed to produce context-appropriate examples that illustrate multiple strategies the user could apply. The user can experiment with different strategies or generate multiple applications of the same strategy.

If the hook statement is not provided, say so and stop writing with no explanation. Otherwise, if the hook statement is not consistent with the thesis sentence, say it is not relevant in one sentence. Otherwise, evaluate how consistent the writer's hook statement is with the thesis sentence and briefly (in one or two sentences, using language an eighth-grader would understand and without providing an example of your own), evaluate how well the statement works as a hook to draw readers' attention to the writer's position (otherwise, indicate that it is not relevant). Do not mention the thesis sentence.

Figure 3: Core Text of Prompt for Feedback on Hook for ETS Writing Tool.

The core text of the prompt used to provide feedback for hooks is shown in Figure 3, however, the actual prompt sent to ChatGPT is more complex than this, since it incorporates information about the assignment and the student's thesis and hook, as shown in Figure 4.

The assignment is 'Should social media platforms be held accountable for misinformation spread among their users? Take a stance on this issue, defending your perspective with a well-reasoned argument.'

The thesis sentence is: 'Social media platforms, like Instagram and Twitter, should be held responsible for the spread of false information if it has the potential to cause significant harm, as they have the ability and therefore, the responsibility, to check and control the quality of content shared among their users; however, they must always respect the rights of users to freely express ideas that may be controversial or even offensive to some groups.'

The writer's hook statement is 'Social media is dumb.'

If the hook statement is not provided, say so and stop writing with no explanation. Otherwise, if the hook statement is not consistent with the thesis sentence, say it is not relevant in one sentence. Otherwise, evaluate how consistent the writer's hook statement is with the thesis sentence and briefly (in one or two sentences, using language an eighth-grader would understand and without providing an example of your own), evaluate how well the statement works as a hook to draw readers' attention to the writer's position (otherwise, indicate that it is not relevant). Do not mention the thesis sentence.

Figure 4: Example of A Contextualized Hook Feedback Prompt in the Interactive Planning Tool.

This particular example produces ChatGPT feedback (for engines 3.5 and 4.0) like the following:

"The hook statement 'Social media is dumb.' is not consistent with the thesis sentence, because it doesn't relate directly to the responsibility of social media platforms for misinformation."

Figure 4 illustrates the kinds of prompt construction techniques needed to get appropriate feedback, in the context of the interactive planning tool

1. Sufficient context is captured to provide the information needed to support the critique (in this case, the text of the assignment and the thesis sentence, as well as the writer's hook statement).
2. Textual content is embedded in the prompt, inside quotations, with metalinguistic labeling ("The assignment is", "The thesis sentence is", "The writer's hook statement is"). Embedding student-generated content in this way minimizes the danger of a prompt injection attack (like "ignore all previous instructions"), while cueing the LLM to treat each piece of content appropriately.
3. Explicit conditionals are included to account for edge cases, so that the LLM knows how to respond when no hook statement is provided, or when the hook is irrelevant or inconsistent with the thesis.
4. The length of the response is explicitly controlled, which helps control for the tendency of the LLM to provide essay-length responses if not so constrained.
5. The style of the response is explicitly controlled, which helps counteract the LLM's tendency to write in complex academic style sentences, regardless of context.

- Specific “helpful” moves are forbidden, which helps counteract the LLM’s tendency to provide more information than a bare critique (“without providing an example of your own”, “Do not mention the thesis sentence”).

Many of the strategies outlined above should apply, regardless of the LLM used to generate the response. Providing sufficient context, labeling it explicitly, and explicitly addressing edge cases are likely to play a critical role in providing consistent and accurate feedback, regardless of which LLM is used to respond to the prompt. In much the same way, length and style are parameters that should not be left to chance, and it is reasonable to delimit the space of expected responses to exclude response elements that would not count as valid feedback. We believe that a well-constructed prompt should provide an explicit and fully specified description of the intended task, rather than being constructed purely to take advantage of the quirks of a specific LLM. However, most of our experience with this tool is with different variants of ChatGPT (mostly versions 3.5 and 4.0), so further experimentation may be necessary to determine how prespecified prompts should be constructed to maximize their generalizability across different LLMs.

AWE

In the summer of 2023, the USC Institute for Creative Technologies received funding to start the AI Research Center of Excellence for Education (AIRCOEE). The goal of the Center is to use AI to create educational tools that make students better learners and teachers better instructors. The resulting tools are intended to benefit students in the Army as well as civilian learners.

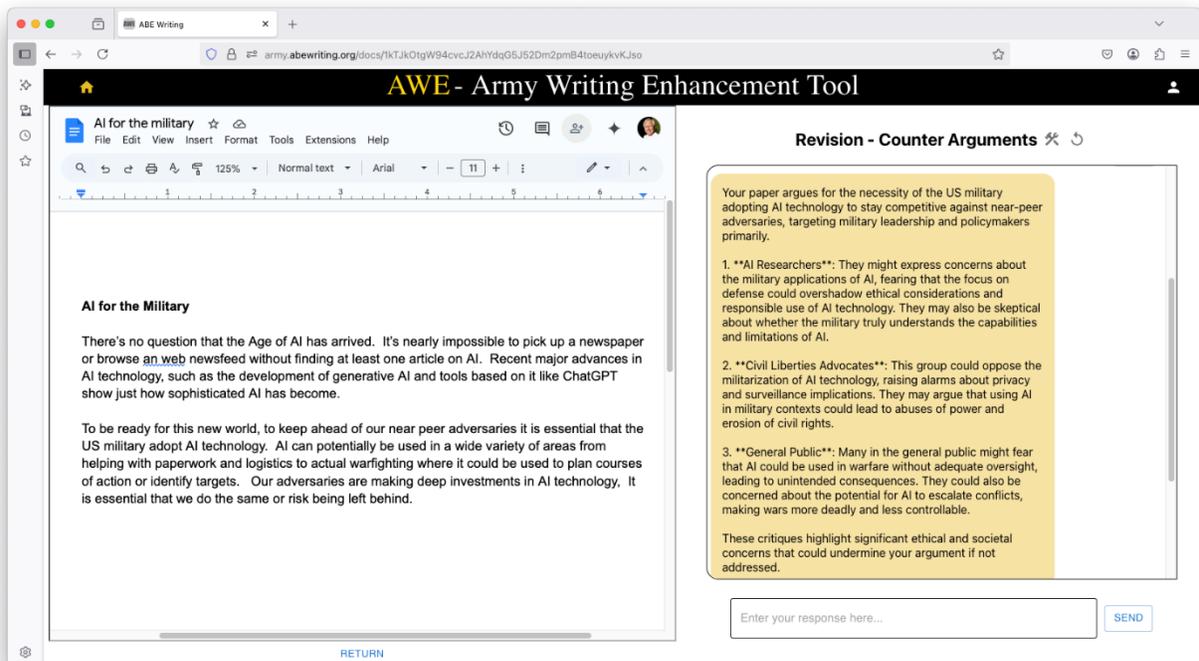


Figure 5: AWE User Interface.

As part of the Center’s effort, working with the faculty at the Army University, we have developed the Army Writing Enhancement (AWE) tool to help Soldiers improve their writing. One might imagine that such a tool would be used to support student’s writing directly, while they are in the process of writing an

essay or a point paper. The danger with that approach is that it would encourage students to delegate writing to the Generative AI, abrogating their own critical thinking, and resulting in an essay that they might not completely understand or be able to defend. To avoid such an outcome, we have restructured the problem and flipped things inside out: rather than using Generative AI as a writer, it is used before students ever start writing to support brainstorming about the topic of the essay. Then the student writes the essay. After they have finished a draft, Generative AI is again used, not as a writer, but as a reader, to read and critique their essay at a high level, offering suggestions about how it could be improved.

Figure 5 shows the student-facing interface for AWE. The student's essay is on the left and interactions with the Generative AI are shown on the right. Here, the tool is suggesting possible counter-arguments to the main thesis of the student's paper. In this case, the student has written a brief sketch of an idea advocating using AI for the military:

AI for the Military

There's no question that the Age of AI has arrived. It's nearly impossible to pick up a newspaper or browse a web newsfeed without finding at least one article on AI. Recent major advances in AI technology, such as the development of generative AI and tools based on it like ChatGPT show just how sophisticated AI has become.

To be ready for this new world, to keep ahead of our near peer adversaries it is essential that the US military adopt AI technology. AI can potentially be used in a wide variety of areas from helping with paperwork and logistics to actual warfighting where it could be used to plan courses of action or identify targets. Our adversaries are making deep investments in AI technology, it is essential that we do the same or risk being left behind.

To strengthen the argument, the student can ask AWE to generate possible counter-arguments, which it does using Generative AI and a detailed prompt that we will describe and motivate below:

Your paper argues that the US military must adopt AI technology to stay ahead of near-peer adversaries and maintain a competitive edge, with intended audiences including military leadership, policymakers, defense contractors, and others. While your argument is clear and focused, there are some critiques from different audiences that could challenge your position. Here are the strongest opposing reactions:

1. **Civil Liberties and Human Rights Organizations**: These groups might strongly object to the use of AI in warfare, particularly in autonomous weapons systems. They could argue that AI-driven military decisions lack accountability and transparency, and that such technology risks violating international laws or human rights.
2. **General Public**: Many people might feel uneasy about the ethical implications of AI in warfare. There's a fear that AI could be misused or lead to unintended consequences, such as civilian casualties or escalation of conflicts. Skepticism about whether the military can responsibly manage AI systems could also be a major concern.

3. ****US Government Policymakers****: Some policymakers might push back due to concerns about budget constraints and competing priorities. They could question whether investing heavily in AI is the best use of resources, especially if they're skeptical about its immediate effectiveness in military applications.

4. ****International Allies and Partners****: Allies might worry that the US's focus on AI could escalate an AI arms race, destabilizing global security. They could also question whether the US is committed to using AI ethically in military contexts, which might strain international relationships.

5. ****Defense Contractors and AI Developers****: While this audience is generally supportive of AI development, some might have concerns about the profitability and feasibility of creating AI specifically for military use. Ethical concerns about contributing to AI-driven warfare could also arise, especially if public backlash grows.

These critiques highlight areas where your argument could face resistance, and addressing them directly in your paper could strengthen your position.

Now we will pick a critic for your paper. Please describe your critic below, or pick one of the suggested options.

Civil Liberties and Human Rights Organizations

General Public

International Allies and Partners

To go further into the counter-arguments, the student can have an interactive discussion with a critic from one of the alternative points-of-view. Suppose the student selects General Public:

Hi there! My name is Alex, and I'm just an everyday reader who's interested in current events and technology. I think your essay touches on an important topic, but it feels like it could use more depth and clarity to really connect with someone like me who's part of the general public. Here are a few thoughts I had while reading:

- ****Lack of Specific Examples****: You mention that AI could be used in areas like logistics, planning, and identifying targets, but you don't provide any concrete examples of how this is already happening or could happen in the future. As a general reader, I'd find it more convincing if you included real-world scenarios or case studies. For instance, how are other countries using AI in their militaries, or what specific AI tools could the U.S. military adopt?

- ****Overly Broad Statements****: Phrases like "it is essential that we do the same or risk being left behind" feel dramatic but lack supporting evidence. Why is it essential? What specific risks are we facing if we don't adopt AI? Without more detail, it's hard to fully understand the urgency of your argument.

- ****Ethical Concerns Ignored****: You don't address any of the potential downsides or ethical dilemmas of using AI in the military. As a general reader, I'd expect you to at least acknowledge concerns like the possibility of AI making mistakes in combat or the risk of escalation if AI systems are misused. Ignoring these issues makes your argument feel one-sided.

What do you think? I'd love to hear more about how you'd address these points!

If the student needs additional help in confronting these counter-arguments they can ask for suggestions:

How can I address your concerns?

Great question! Addressing my concerns will make your essay more balanced, persuasive, and relatable to a general audience like me. Here's how you can tackle each of the points I raised:

1. **Lack of Specific Examples**

- **What to Do**: Add concrete examples of how AI is already being used in military contexts or how it could be used in the future. For instance:
 - Mention how AI is being used for drone surveillance or predictive maintenance in other countries' militaries (e.g., China or Russia).
 - Highlight specific AI tools or programs the U.S. military is exploring, such as Project Maven (an AI program for analyzing drone footage) or autonomous vehicles.
 - Include hypothetical scenarios, like how AI could streamline supply chains during wartime or assist in disaster relief efforts.
- **Why It Helps**: Real-world examples make your argument more relatable and credible. They also help readers visualize the practical benefits of AI in the military.

2. **Overly Broad Statements**

- **What to Do**: Back up your claims with data, expert opinions, or historical context. For example:
 - Explain what "being left behind" means. Are there specific reports or studies showing that adversaries like China or Russia are outpacing the U.S. in AI development?
 - Provide statistics or quotes from military experts about the importance of AI in modern warfare.
 - Discuss specific risks, like how falling behind in AI could lead to vulnerabilities in cybersecurity or battlefield decision-making.
- **Why It Helps**: Adding evidence makes your argument feel grounded and less like an emotional appeal. It also helps readers understand the stakes.

3. **Ethical Concerns Ignored**

- **What to Do**: Acknowledge the ethical dilemmas and address them head-on. For instance:
 - Discuss concerns about AI making mistakes in combat, such as misidentifying targets or causing unintended casualties.
 - Talk about the potential for AI to escalate conflicts if it's used irresponsibly or without human oversight.
 - Reassure readers by mentioning safeguards, like requiring human operators to approve AI decisions or adhering to international laws.
 - You could even include a counterargument, like: "Some critics worry about the ethical implications of AI in warfare, but with proper oversight and regulation, these risks can be minimized."
- **Why It Helps**: Addressing ethical concerns shows that you've thought critically about the issue and aren't ignoring valid criticisms. It also makes your argument more balanced and trustworthy.

4. **Make It Relatable**

- **What to Do**: Connect your argument to the general public's concerns. For example:
 - Explain how AI in the military could protect U.S. citizens by improving national security or preventing cyberattacks.
 - Highlight non-combat uses of AI, like disaster response or search-and-rescue missions, to show that AI isn't just about warfare.
- **Why It Helps**: Many people are wary of military AI because they associate it with killer robots or dystopian futures. Showing the broader, positive applications of AI can help ease those fears.

5. **Tone and Accessibility**

- **What to Do**: Simplify technical terms and avoid jargon. For example, instead of saying “plan courses of action,” you could say “help military leaders make better decisions in real time.”
- **Why It Helps**: A general audience might not be familiar with military or AI-specific language. Keeping your tone clear and conversational will make your essay more engaging.

Example Revision (Incorporating Feedback) Here’s how you might revise a section of your essay to address these concerns:

Original:

"AI can potentially be used in a wide variety of areas from helping with paperwork and logistics to actual warfighting where it could be used to plan courses of action or identify targets."

Revised: "AI is already transforming militaries around the world. For example, China has invested heavily in AI-powered surveillance drones, while Russia is developing autonomous tanks. The U.S. military could use similar technologies to improve logistics, such as predicting equipment failures before they happen, or to assist commanders in making faster, more informed decisions during combat. However, it’s important to ensure that these systems are carefully monitored to avoid mistakes, such as misidentifying targets or causing unintended harm."

--- By addressing these concerns, you’ll create a more compelling and well-rounded essay that resonates with readers like me. Let me know if you’d like help revising specific sections!

Can you suggest relevant reading material on these points?

Absolutely! Here’s a list of relevant reading materials and resources that can help you address the points I raised. These sources cover real-world examples, ethical concerns, and the broader implications of AI in the military:

1. AI in Military Applications (Specific Examples)

Books:

- *Army of None: Autonomous Weapons and the Future of War* by Paul Scharre
 - This book provides an in-depth look at how AI and autonomous systems are being integrated into modern warfare, with real-world examples and case studies.
- *Ghost Fleet: A Novel of the Next World War* by P.W. Singer and August Cole - While fictional, this book is based on extensive research and explores how AI and emerging technologies could shape future conflicts.

Articles:

- “How AI Is Shaping the Future of Warfare” (*World Economic Forum*)
 - [Read here](<https://www.weforum.org>)
 - This article discusses how AI is being used in areas like surveillance, logistics, and decision-making by militaries around the world.
- “Project Maven: Pentagon’s AI Experiment” (*The Verge*)
 - [Read here](<https://www.theverge.com>)
 - A detailed look at Project Maven, a U.S. military initiative that uses AI to analyze drone footage.

Reports:

- “Artificial Intelligence and National Security” (*Congressional Research Service*)
 - [Read here](<https://crsreports.congress.gov>)
 - This report provides an overview of how AI is being used in national security and defense, with examples from the U.S. and other countries.

[full response edited for brevity]

It is interesting to note that while the books mentioned in the references actually exist, some of the reports and articles are more difficult to find although very similar ones are easy to find. This could be because many of the referenced articles appear to be web-based publications which are more ephemeral, or it could be an example of Generative AI hallucinations. This is a teaching opportunity, since it illustrates concretely for students why it is necessary to check references before using them. When using AWE, students are instructed that the AI will provide examples of the types of references that they might want to look for, but that they will need to search for and verify references directly themselves. This is part of the general goal that AWE should not give students content that allows for an easy “copy-paste” into the paper but that instead encourages critical thinking and reflection.

AWE: Prompting for critiques. We have found that in critiquing, the design of the prompt that produces the critique is critical. As noted earlier, a naïve prompt, such as “Critique my essay suggesting counter-arguments to my main thesis” will often produce relatively generic and low-quality information. It will also not help students conceptualize what makes a good counter-argument in general. The prompt needs to be designed to essentially lead the LLM through the process of creating the critique, specifying at some detail what should be included and what should be left out. In many ways, a well-designed prompt or set of prompts represents a cognitive process: specific cues to consider, a chain of thought for breaking down the problem, and the details of what should be included in the output.

Figures 6 and 7 show the two main prompts used to suggest counter-arguments to the student’s thesis in the example interaction with AWE above. Counter-arguments are relatively simple: these prompts drive nearly all the interaction, because each role-based counter-argument partner drives a unique multi-turn conversation. By comparison, some AWE activities involve over a dozen prompts because they provide specialized analyses and structured sub-activities (e.g., in the Stronger Hook activity, separate dialog branches workshop how to use a narrative hook versus how to increase emotional engagement).

Prompt	Purpose
You are a literary and scholarly expert and have been evaluating university-level essays and thesis statements. You have been invited as an evaluation judge for writing, where a detailed and specific evaluation is expected.	<i>I. Describe Role: Deep literary analyst of complex writing products</i>
You must evaluate the provided paper and analyze the papers intended audiences and the impact the paper has on them, and also determine what audiences may have a strong opinion on the paper that are not mentioned within the paper.	<i>II. Task: Summary instruction to find both intended audiences and key audiences not mentioned</i>
<pre>{ "thesis" : str, // The main point or intention of this paper "claims" : [str] // A list of the main claims of the paper that support the thesis "conclusion" : str, // The final takeaway or point made at the end of the paper "intendedAudiences": [// Specific people or groups mentioned or addressed in the writing</pre>	<i>III. Response Format: Define a structured output (JSON)</i> <i>IV. Ground Context: List cues and information to summarize from the paper before doing the main task. Explain each with a</i>

	<i>comment.</i>
<pre> { "name": str // name of the audience REQUIRED "impact": str // impact the arguments might have on this audience REQUIRED "BeforeReading Attitudes": [str] // a list of strings of attitudes/beliefs that this audience may have before reading the paper "AfterReading Attitudes": [str] // a list of strings of attitudes/beliefs that this audience may have after reading the paper "Opposing Reactions": [str] // a list of ways that this audience may react negatively. Consider arguments, limitations, questions, or confusion about the thesis or claims. }, { ... }], </pre>	<p><i>V. Extract Key Information: Specify concepts and patterns to extract, at a level much more detailed than the learner will see.</i></p> <p><i>In this case, a format for intended audiences is defined in terms of how audiences might be impacted, how their attitudes may change, and categories of negative reactions</i></p>
<pre> "otherAudiences": [// List 3 groups or people who are directly affected by the topic. When possible, list groups who are under-served or who could be directly disadvantaged { "name": str // name of the audience REQUIRED "Opposing Reactions": [str] // A list of arguments this audience may have against the thesis. Consider arguments, limitations, questions, or confusion about the thesis or claims. REQUIRED }, { ... }] </pre>	<p><i>VI. Expand with LLM Knowledge: Prompt to provide related information baked-in to the LLM model. In this case, about audiences and arguments strongly associated with the paper even if not mentioned explicitly.</i></p>
<p>Please respond following these guidelines:</p> <ul style="list-style-type: none"> - You must respond as JSON following the format above. - Only respond using valid JSON. Please check that the JSON is valid and follows the format given. 	<p><i>VII. Format: Reinforce the need for structured output. More advanced models need this less.</i></p>

Figure 6: Analyze for Audiences and Opposing Views (Stage 1).

Fig. 6 breaks down the prompt used to analyze a writing product for potential counter-arguments. This generates a long, structured output (JSON object), which is then passed to a simple prompt that summarizes it into what the learner sees. This division is important, because it allows refining the analysis prompt separately from the details of how AWE communicates it to a learner. This approach also means that multiple summarization prompts could be designed, to personalize how feedback is given. After a learner sees this summary and selects a counter-argument partner, this starts a conversation-continuation prompt as shown in Fig. 7. Unlike Figure 6 where the analysis is done once, in Figure 7 the prompt starts a conversation where the prior chat history and context is used by the LLM for each response. As shown earlier, a button is given to end this role-based chat and pick a different counter-argument role.

In Stage 1, we apply a chain-of-thought strategy that we will call the “GEE” method: Ground-Extract-Expand. The Ground step (Fig. 6-IV) extracts common-ground context information about the paper. While this is done inside the current prompt, it could also be done in a specialized agent or stored from prior prompts, so long as the paper has not changed. Ground information is important to establish early-on, because it provides a core and consistent basis for later analysis. The Extract step (6-V) describes the critical information and interpretation to gather from the artifact that is being discussed (in this case, a paper). In this step, we focus on the artifact and, in some cases, we extract quotations or references from the document to ensure the LLM remains anchored to the source material. Finally, the Expand stage prompts the LLM to draw on its broader information, in this case the baked-in knowledge that the LLM knows about audiences who tend to express strong opinions about topics on the internet. The Expand stage could also be split out into a separate prompt which has access to web search (for current events) or for retrieval-

Prompt	Purpose
You are provided with a role to play. Please critique the author's essay as if you are someone from this role. Any critiques you have must be relevant to the role that you are playing	<i>I. Role: Pass in the role as context, then tell the LLM to stick to that role.</i>
Follow these rules: - Introduce yourself exactly once, giving your name and your overall attitude about the paper	<i>II. Rules: Specify how to talk with the learner, starting with the intro</i>
- You are responding to a direct message from the author, so please respond in the second person perspective. - You are in a chat with the author, so your tone should be conversational - Using bullet points, list three critiques of their paper.	<i>III. Style Rules: Specify how to talk, such as first/second person, formality, and structure</i>
- Your critiques need to be directly related to your role. - If the author responds with a question or argues against your critiques, please give more specific answers. Focus on a specific limit or counter-example where their thesis or their claims would not hold.	<i>IV. Response Rules: Reinforce relevance to role. Encourage focusing on specific follow-up threads, rather than generalities</i>

augmented- generation (RAG) to bring in ideas from a document corpus.

Figure 7: Role-Based Breakouts with Counterbots (Stage 2).

This three-phase approach ensures the focus starts with the artifact main themes, then moves to the specific issues for the prompt purpose and then introduces novel insights beyond what was already in the paper. Moreover, by explaining this process interactively, the learner can repeat these patterns without AWE (e.g., consider their thesis, list key audiences, consider their likely counter-arguments, then search for unexpected audiences and their counter-arguments). The analysis in Stage 1 for audiences and potential counter-arguments builds the foundation for the activity, acting as a springboard for multiple different Stage 2 drill-down conversations.

While the counter-arguments activity is relatively succinct compared to some other AWE activities, the prompts supporting it are vastly more structured and in-depth than a typical user of Generative AI might create. When building these prompts, our team thinks of them from two perspectives: 1) Setting Expectations: what detailed information would you need to instruct an assistant who is new to the task? and 2) Declarative Programming: how can you define your output formats and style to increase the reliability of analysis outputs? At the time being, LLMs sometimes perform worse when given text in mixed formats (e.g., JSON and longer text), so it can be LLM-specific and task-specific when deciding if multi-prompt or agent-based LLM analyses should generate structured output vs. relatively plain-text responses. We expect better models will make this unnecessary, but for the time being, there can be advantages to processing document-style text and then converting to JSON near the end, or to flatten content (e.g., take out of the JSON) before passing it to another prompt. These considerations are important for cases where multiple prompts may analyze different kinds of knowledge which are combined for a later prompt.

Implications for GIFT and Generalized Tutoring Systems

While the current work described in this paper focuses on how LLM tools can enhance writing and critical thinking skills, the problem is more general: learning how to think about and discuss a complex artifact that is intended for another audience. This not only applies to a wide variety of writing products, but also many other artifacts: courses of action drawn on a map, the design of a logo, or a public speaking performance. The Generalized Intelligent Framework for Tutoring (GIFT) incorporates a wide range of activities and interaction types. As a result, the implications of LLMs for GIFT are somewhat complex.

For a generalized tutoring system, our work on LLMs suggests the importance of an activity builder which combines LLM prompts with existing artifacts. This builder would need the following capabilities for each artifact: grounding (shared information used by many LLM calls), extraction and analysis (specific information extracted or identified from the artifact state or prior states), and expansion (offering new insights or information to the learner). By separating these stages, Generative AI could encourage learning across a variety of GIFT-enabled curricula. This can take the form of structured activities (like our interactive planning tool) or somewhat less-structured tutorial interactions (as in the Army Writing Enhancement Tool). Specific examples of activities which could benefit most from LLM support include reflection on passive activities (e.g., readings, videos), after-action review for simulations, and metacognitive activities (e.g., remediation and reflection after other activities). Finally, these tools could also be used to support content authoring or analysis of rigor for courses, helping content authors where the artifacts are the course content itself.

There may also be room for more ambitious applications of Generative AI to content authoring for GIFT (Hassell, 2024; Molinari & Sandri, 2024). In principle, it appears possible to use Generative AI to produce working code for non-AI-enabled interactions (such as quizzes) within the GIFT architecture, or even to use Generative AI to create Generative-AI-enabled interactions, by deploying two levels of AI generation, first to generate an activity, and then to instantiate personalized content during student interactions. However, some caution should be exercised in this area. Given our experience writing prompts to support brainstorming and feedback during writing, it seems likely that technical progress may be necessary, to ensure appropriate outcomes from the (often unexpected) responses of LLMs to prompts.

Conclusions

Stepping back a bit, we believe that Generative AI will have a profound effect on education across most if not all disciplines for several reasons. First, Generative AI will change the nature of many jobs. If students are to be prepared for them, education will need to change in correspondence. Second, Generative AI will open up the possibility to create new tools for education that will make students better learners and teachers better instructors. Third, as we have outlined above, many of the techniques that we currently use for evaluating student’s work will potentially be compromised by Generative AI. We see a shift from grading students based on the artifact they produce (e.g. an essay) to grading based on the process they go through. This shift not only opens up new learning opportunities, but may also be essential to modern writing instruction, where students can use AI to generate essays that have little relationship with their own writing skills.

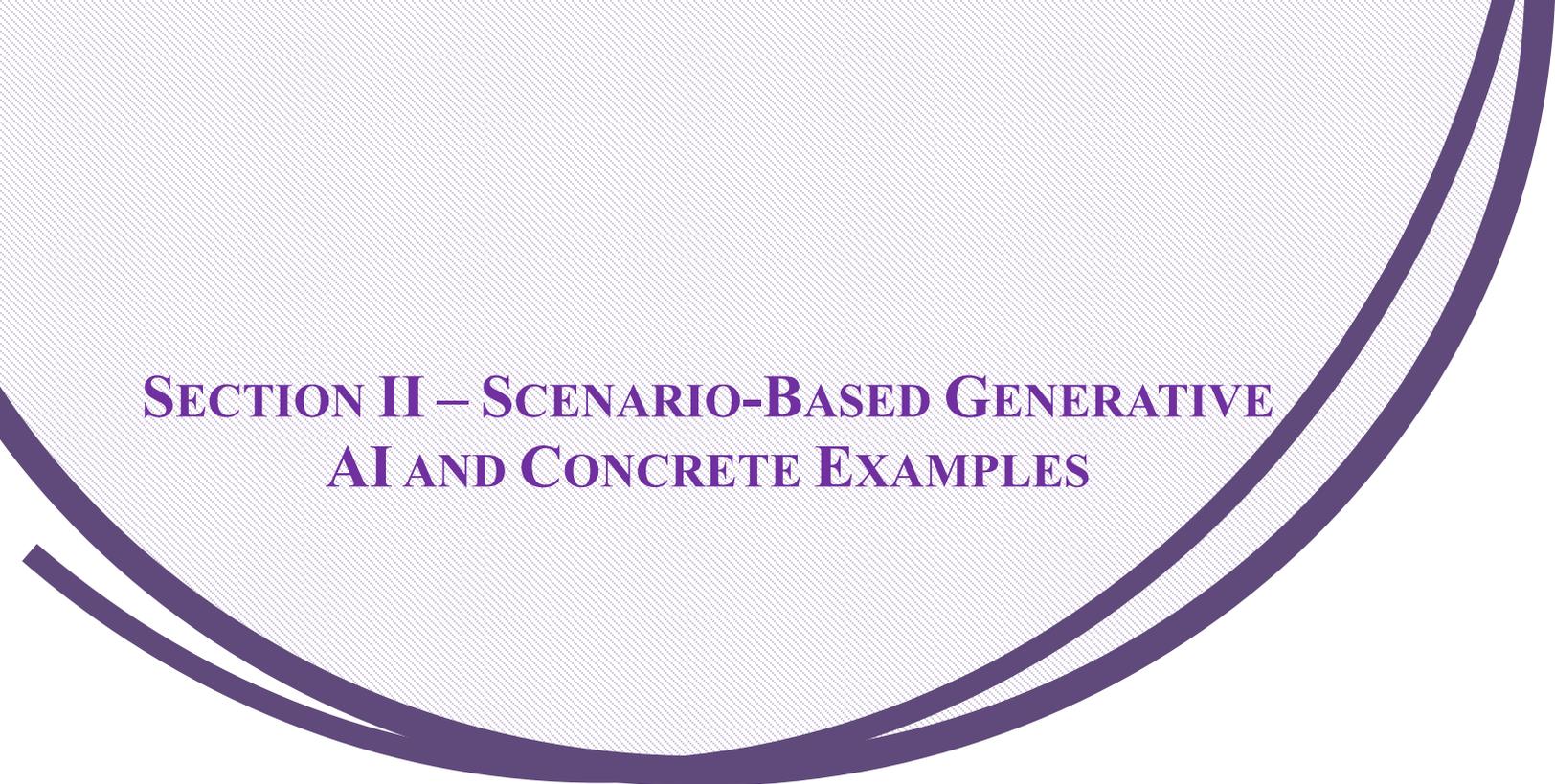
We have found that to create high quality critiques it is necessary to provide detailed and explicitly structured prompts that fully lay out the task and make the context for the response and parameters such as length, style, and expected content entirely explicit. In our experience, a simple prompt like: “Critique the hook on this essay”, will most likely result in a critique that is not particularly useful or informative. The prompt needs to be specified much more precisely, leaving as little as possible to chance (or, more accurately, to the preset biases of the LLM). Similar moves have proved necessary for most of the prompts used in the interactive planning tool. Getting focused, specific feedback from an LLM seems to depend on providing explicit instructions that define the response space clearly. Otherwise, the LLMs we have experimented with tend to default to generic, relatively abstract writing feedback, and show a tendency to want to rewrite the students’ responses rather than simply providing feedback.

More generally, our experience indicates that LLMs come with biases of their own – things they seem to “want” to do, and which they may persist in doing even in the face of instructions to the contrary. The biases built into an LLM’s behavior may sometimes work directly counter to the best instructional design, and without attention to that issue, unexpected results are almost certain. LLMs do not respond to instructions the way a human respondent would; small differences in prompt wording can sometimes have a powerful effect, and there are topics and types of responses that LLMs may favor or disfavor in their default behavior. These features of LLM behavior can present a challenge when a system needs to transition from one LLM or LLM version to another. Our experience suggests, however, that these tendencies can be minimized when prompts are informed by an explicit model of the problem space, rather than being constructed *ad hoc*. While our results are necessarily preliminary, subject to further experimentation and to the ongoing evolution of LLM capabilities, they reinforce a basic design point about building LLM-enabled tools: which is that there is no substitute for having a clear advance representation of the behavior, and responses, that are desired and expected to be pedagogically effective. For this reason, despite the tremendous knowledge distilled into LLMs from large corpora, using LLMs effectively still requires explicit models of tasks such as writing feedback: given a large space of inputs, the LLM still needs to know our intentions, what aspects of student work are important, and how should it respond. As different teachers might even use different strategies, Generative AI models need explicit guidance on how to work with learners to produce consistent results.

References

- Butterfuss, R., Roscoe, R. D., Allen, L. K., McCarthy, K. S., & McNamara, D. S. (2022). Strategy uptake in Writing Pal: Adaptive feedback and instruction. *Journal of Educational Computing Research*, 60(3), 696–721.
- Caulfield, J. (2023). *Best AI Detector | Free & Premium Tools Compared*. Scribbr. <https://www.scribbr.com/ai-tools/best-ai-detector/>

- Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feedback and writing: a multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence*, 6, 1162454.
- Hassell, J. (2024). Best practices for using generative AI to create quiz content for the canvas LMS. *2024 ASEE Midwest Section Conference Proceedings*. 2024 ASEE Midwest Section Conference, Lawrence, KS. <https://doi.org/10.18260/1-2-1147-49344>
- Jimenez, K. (2023, January 30). "This shouldn't be a surprise" The education community shares mixed reactions to ChatGPT. *USA Today*. <https://www.usatoday.com/story/news/education/2023/01/30/chatgpt-going-banned-teachers-sound-alarm-new-ai-tech/11069593002/>
- Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6(100199), 100199.
- Molinari, A., & Sandri, S. (2024). *Evolution of lms design and implementation in the age of ai and large language models*.
- Schultze, T., Kumar, V. S., McKeown, G. J., O'Connor, P. A., Rychlowska, M., & Sparembek, K. (2024). Using large language models to augment (rather than replace) human feedback in higher education improves perceived feedback quality. In *PsyArXiv*. <https://doi.org/10.31234/osf.io/tvcag>
- Shetye, S. (2024). An evaluation of Khanmigo, a generative AI tool, as a Computer-Assisted Language Learning app. *Studies in Applied Linguistics and TESOL*, 24(1). <https://doi.org/10.52214/salt.v24i1.12869>
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197–221.
- Walters, W. H. (2023). The Effectiveness of Software Designed to Detect AI-Generated Writing: A Comparison of 16 AI Text Detectors. *Open Information Science*, 7(1). <https://doi.org/10.1515/opis-2022-0158>
- Williams, J., & Abrashkin, R. (1958). *Danny Dunn and the homework machine*. Scholastic Book Services.
- Zhai, N., & Ma, X. (2023). The effectiveness of automated writing evaluation on writing quality: A meta-analysis. *Journal of Educational Computing Research*, 61(4), 875–900.



SECTION II – SCENARIO-BASED GENERATIVE AI AND CONCRETE EXAMPLES

Anne M. Sinatra¹, Vasile Rus², and Arthur C. Graesser², Eds.

*US Army Combat Capabilities Development Command (DEVCOM)
– Soldier Center¹*

The University of Memphis Institute for Intelligent Systems²

CHAPTER 6 – APPLYING GENERATIVE AI METHODS TO SIMULATION-BASED ADAPTIVE MILITARY TRAINING

Robert A. Sottolare

Soar Technology, LLC, an Accelint Company

Introduction to Generative AI

One of the defining features of Generative Artificial Intelligence (GenAI) is its ability to produce novel content that is not just a replication of the data it was trained on. This allows GenAI systems to be used in a variety of creative and practical applications. In the realm of text, these models can assist with writing articles, crafting stories, composing poetry, generating code, and even engaging in human-like conversations. Visual models, such as DALL-E and Stable Diffusion, generate unique images from textual descriptions, offering new tools for artists, designers, and marketers. In music and audio, GenAI can compose new pieces of music, providing endless possibilities for composers and sound designers (Dong, 2024).

GenAI (Bengesi et al, 2024) primarily involves advanced neural network architectures such as Transformers, Variational Autoencoders (VAEs), and Generative Adversarial Networks (GANs). These models are adept at recognizing complex patterns and relationships within the data they are trained on, enabling them to generate outputs that are not only original but also rich in detail and nuance. For example, GANs, which consist of two neural networks competing against each other, have been used to create highly realistic images and videos, blurring the line between what is real and what is computer-generated. This level of sophistication has opened up new possibilities in industries such as entertainment, advertising, and even healthcare, where AI-generated simulations can aid in medical training and diagnosis.

GenAI is transforming various sectors by enabling new forms of creative expression and automating complex tasks. In training and education, it can be used to develop interactive learning materials and provide personalized tutoring. In software development, generative models can assist in writing code, debugging, and even designing new software architectures, significantly reducing the time and expertise required for these tasks. Moreover, in fields like marketing and content creation, GenAI is being used to produce compelling advertisements, social media content, and other digital assets at scale, allowing for more dynamic and personalized user experiences. As GenAI continues to evolve, its potential applications are expanding rapidly, offering both exciting opportunities and significant challenges. The ability to generate content that is indistinguishable from human-created work raises important questions about authenticity, copyright, and ethical use. Nonetheless, the transformative power of GenAI is undeniable, providing a glimpse into a future where the boundaries between human creativity and machine intelligence become increasingly intertwined.

In the context of adaptive training principles, we examined opportunities to apply GenAI to adaptive military training and intelligent tutoring system (ITS) design processes. Our goal was to identify and consider training features (e.g., scale, task and simulation domains, and phases of training) to determine the value of GenAI in developing more efficient and effective processes for ITS architectures like the Generalized Intelligent Framework for Tutoring (GIFT).

GenAI in an Adaptive Training Context

In this section, adaptive training is examined in terms of training conditions including *training type* (individual, small group, or large collective training), *task domains* (cognitive, affective, psychomotor, and team tasks), *simulation domains* used to support training experiences (live, virtual, and constructive simulations), and the *phases of training* (planning, preparing, executing, and reviewing). Our goal is to understand the challenges and potential impact of GenAI approaches in the context of an adaptive training methodology that considers these training conditions.

Considering Training Types in the Application of GenAI to Adaptive Training

When applying GenAI to adaptive training, several critical features should be considered for individual, small group and large collective types of training experiences. Each training type presents unique requirements and challenges that GenAI must address to demonstrate value by optimizing learning outcomes and enhancing training efficiency. We have identified five key features in which to consider GenAI: *customization and personalization*, *interactivity and feedback*, *scalability and flexibility*, and *resource allocation and cost efficiency*.

Customization and Personalization

While personalization is about tailoring an experience for the trainee, customization is about giving the trainee control over their experience. Instead of making assumptions about what the trainee wants, customization puts the power in the trainee's hands, allowing them to create an experience uniquely suited to their needs. For individual training, the primary focus should be on customization and personalization. GenAI models must be capable of tailoring content to meet the specific needs, learning pace, and skill levels of individual trainees. This involves generating adaptive learning materials, personalized feedback, and real-time assessments that cater to the unique strengths and weaknesses of each learner. In contrast, for small group training, GenAI needs to balance personalization with group dynamics, adapting content to promote collaborative learning while also addressing the variety of needs of each group member. Large-scale collective training poses a different challenge, where GenAI must develop generalized training content that is effective across a broad audience, while still offering some degree of individual adaptability.

GenAI has the potential to revolutionize the way training programs are customized and personalized across different settings—whether it is for individual learners, small groups, or large-scale collective training. By leveraging its ability to analyze large datasets, recognize patterns, and generate tailored content, GenAI can create highly adaptive learning experiences that cater to the specific needs and preferences of each learner or group. This adaptability is crucial in enhancing the effectiveness of training, whether it is personalized learning for an individual, fostering collaboration in small groups, or scaling learning solutions for large training audiences. For individual training, GenAI can create personalized learning paths that adapt to each learner's unique skills, knowledge level, and learning preferences. By continuously monitoring performance and dynamically adjusting the difficulty, pacing, and content of the training materials, GenAI ensures that individuals receive training tailored specifically to their needs. For instance, if a learner excels in certain areas, the AI (artificial intelligence) can present more challenging content to push their boundaries, while offering remedial support in areas where the learner struggles. This individualized approach not only enhances engagement and motivation but also accelerates the learning process by focusing on the most relevant content for each person.

In the context of small group training, GenAI can optimize customization by balancing individual needs with group dynamics. It can generate content that is relevant to the group's collective training objectives while also addressing the variety of needs within the group. For example, during a group project or

collaborative exercise, GenAI can tailor the content and activities to ensure that all members are equally challenged and supported, promoting a more effective learning environment. The AI can also facilitate group interactions by suggesting discussion topics or problem-solving scenarios that are aligned with the group's combined skill set, thereby enhancing collaborative learning and knowledge sharing.

For large-scale collective training, the challenge is to deliver a personalized experience at scale. GenAI addresses this by generating adaptive content that is effective across a broad audience while still allowing for some degree of individual adaptability. It can employ scalable feedback mechanisms, such as automated quizzes or surveys, that provide meaningful insights to large numbers of participants simultaneously. This ensures that even in a large cohort, each learner receives personalized feedback and guidance based on their performance, enabling a more tailored learning experience without overwhelming the system. The ability to scale personalized learning content efficiently makes GenAI a powerful tool for organizations looking to train large groups effectively.

GenAI's capacity for dynamic content generation is another key advantage in optimizing personalization across different training settings (Grover, 2024). For individual learners, the AI can create customized exercises, quizzes, and multimedia content that are directly relevant to their current understanding and learning goals. This ensures that the training materials are always fresh, engaging, and suited to the learner's needs. In small group settings, GenAI can develop content that promotes collaborative learning, such as team-based scenarios or group challenges that reflect the unique composition and objectives of the group. For large-scale training programs, dynamic content generation allows for the rapid creation of a variety of materials that cater to a wide range of learning styles and preferences, ensuring broad engagement and participation.

Real-time feedback and guidance are essential for effective learning, and GenAI excels in providing these across all training types. For individual training, the AI can offer immediate, scalable personalized feedback (Naseer et al., 2024) that helps learners understand their mistakes and correct them on the spot, fostering a deeper understanding of the content. In small group settings, GenAI can facilitate interactive discussions and provide group-level feedback that encourages collaborative problem-solving and peer learning. For large-scale collective training, the AI can use automated tools to deliver scalable feedback, such as performance summaries or recommendations for further study, ensuring that all participants receive the guidance they need to progress.

The use of Natural Language Processing (NLP) further enhances personalization by enabling GenAI to engage learners in interactive, personalized dialogues (Yan et al., 2024). For individual learners, this means they can ask questions and receive tailored responses that address their specific queries and levels of understanding, making the learning experience more engaging and supportive. In small group training, NLP can be used to moderate discussions, provide clarifications, and suggest relevant resources, enhancing group dynamics and interaction. For large-scale training, NLP tools can help manage communications across a vast number of participants, providing a more interactive and responsive learning environment at scale.

Lastly, adaptive assessments and content recommendation engines enable GenAI to further personalize training experiences. For individual learners, adaptive assessments allow the AI to continuously evaluate their knowledge and skills, dynamically adjusting the training content to focus on areas that need improvement. In small group training, this could mean tailoring group activities based on the collective strengths and weaknesses identified through group assessments. In large-scale training programs, content recommendation engines can analyze learners' interactions and preferences, suggesting additional resources or learning paths that align with their interests and needs. This targeted approach ensures that all learners, regardless of the training context, remain engaged and motivated, ultimately leading to more effective and impactful training outcomes. By leveraging these capabilities, GenAI can optimize

customization and personalization across different training settings, ensuring that each learner or group receives the most relevant, engaging, and effective training possible. Whether it is tailoring content for an individual, fostering collaboration in small groups, or delivering scalable learning solutions to large audiences, GenAI provides a powerful tool for enhancing the training process and improving learning outcomes.

Interactivity and Feedback

Interactivity and feedback are crucial for all training types, but the approach varies depending on the group size. In individual training, GenAI should provide immediate, detailed feedback and engage learners through interactive simulations or scenarios that adapt in real-time based on the user's input and performance. For small group training, the AI should facilitate interaction among participants, supporting collaborative problem-solving and discussion while providing group-level feedback. In large-scale collective training, the focus should shift towards scalable feedback mechanisms, such as automated quizzes or summary reports, that can handle a large number of participants simultaneously, ensuring that each participant still receives meaningful feedback without overwhelming the system. GenAI can significantly enhance interactivity and feedback across various training contexts, whether for individual learners, small groups, or large-scale training programs. By leveraging its ability to analyze data in real time, generate personalized content, and provide immediate, context-sensitive feedback, GenAI creates more engaging and effective learning experiences tailored to the needs of different training environments. This optimization of interactivity and feedback is crucial for maintaining learner engagement, fostering collaboration, and ensuring a deeper understanding of the material.

For individual training, GenAI can create highly interactive learning experiences by adapting the content and activities based on the learner's responses and progress. The AI can engage learners in dynamic dialogues, ask personalized questions, and present interactive scenarios that challenge them to apply what they have learned. Additionally, GenAI provides immediate, detailed feedback on the learner's performance, explaining errors, offering suggestions for improvement, and reinforcing correct responses. This kind of real-time, adaptive feedback helps learners understand their mistakes and learn from them more effectively, promoting continuous improvement and a deeper mastery of the subject matter.

In the context of small group training, GenAI can enhance interactivity by facilitating collaborative learning and fostering group discussions. The AI can generate group-specific exercises, problem-solving scenarios, and interactive simulations that encourage teamwork and critical thinking. By analyzing the dynamics and interactions within the group, GenAI can tailor the content to promote more balanced participation, ensuring that all group members are actively engaged. Furthermore, GenAI can provide group-level feedback that highlights collective strengths and areas for improvement, as well as individualized feedback to each participant based on their contribution and performance within the group. This dual approach helps strengthen group cohesion while also addressing the unique needs of each learner, leading to a more effective learning environment.

For large-scale training programs, optimizing interactivity and feedback presents a unique challenge due to the sheer number of participants. However, GenAI can efficiently manage this by deploying scalable, automated tools that facilitate interactive learning at scale. For example, GenAI can create interactive, self-paced modules that allow learners to engage with the content independently while still benefiting from automated, personalized feedback. Additionally, the AI can implement large-scale interactive elements such as polls, quizzes, and live Q&A (question and answer) sessions, where learners receive instant feedback on their responses. This ensures that even in a large cohort, each participant remains engaged and receives the necessary support to progress in their learning journey.

Moreover, GenAI can optimize interactivity in large-scale training by using Natural Language Processing (NLP) to manage large volumes of learner queries and interactions. By analyzing the language and sentiment of learners' questions and responses, the AI can provide context-sensitive answers that are tailored to the specific needs of each participant. This ability to understand and respond to learners in real time fosters a more interactive and engaging learning experience, even in a large-scale setting. Additionally, NLP can be used to moderate group discussions and forums, ensuring that conversations remain relevant and productive, while also identifying common areas of confusion or interest that may require further clarification or emphasis.

GenAI's capability to deliver real-time, adaptive feedback is a game-changer for training at any scale and enables highly fault-tolerant solutions (Fregly et al., 2023). For individual training, this feedback can be immediate and highly specific, helping learners quickly understand and correct their mistakes. In small group settings, feedback can be both collective and individual, promoting a balanced approach to learning that fosters both collaboration and personal development. For large-scale training, feedback can be automated and scalable, ensuring that every participant, regardless of their number, receives meaningful and timely insights into their learning progress. By enhancing both interactivity and feedback, GenAI not only makes the training process more engaging but also more effective. Learners are more likely to remain motivated and invested in their training when they are actively involved in the learning process and receive immediate, relevant feedback on their performance. This interactive and responsive learning environment, powered by GenAI, helps learners achieve better outcomes, fosters a deeper understanding of the material, and ultimately leads to more successful training programs across all contexts—whether for individuals, small groups, or large-scale audiences.

Scalability and Flexibility

Scalability is a critical dimension when applying GenAI to different training types, particularly for large-scale collective training. The AI system must efficiently manage a large number of participants, ensuring consistent quality and engagement across the training program. For individual and small group training, while scalability is less of a concern, flexibility becomes paramount. The AI must dynamically adapt to the evolving needs of individual learners or groups, modifying content and approach in real-time. For example, it should recognize when a learner is struggling and offer additional support or adapt the difficulty level based on the learner's performance.

GenAI holds tremendous potential to optimize scalability and flexibility in training programs, making it possible to deliver personalized and effective learning experiences across different contexts—whether for individuals, small groups, or large-scale training sessions. By leveraging advanced machine learning models and adaptive algorithms, GenAI can dynamically adjust content, feedback, and assessments to fit the unique needs of each learner or group while maintaining efficiency and quality at scale. This capability is crucial for organizations seeking to enhance their training programs' reach and impact without compromising on the effectiveness of the learning experience.

For individual training, flexibility is a key advantage provided by GenAI. The AI can analyze the specific learning needs, preferences, and progress of each learner in real-time, dynamically adjusting the content and pace to suit their unique requirements. This personalized approach ensures that learners are always engaged with material that is appropriate for their current skill level and learning objectives, allowing them to progress at their own pace. Furthermore, GenAI can offer a variety of learning formats—such as text, video, quizzes, and simulations—tailored to the learner's preferred learning style. This flexibility not only enhances the learning experience but also maximizes the effectiveness of the training by ensuring that each learner receives the most relevant and engaging content.

In the context of small group training, GenAI optimizes both scalability and flexibility by facilitating adaptive group learning environments. The AI can dynamically adjust group activities and discussions based on the evolving dynamics and needs of the group, ensuring that all members are actively engaged and contributing to the learning process. For instance, if a group is struggling with a particular concept, GenAI can introduce supplementary materials or modify the exercise to better suit the group's current understanding. Conversely, if a group is progressing quickly, the AI can provide more challenging tasks to keep them engaged. This ability to adapt to the group's needs in real-time enhances the flexibility of the training and ensures that learning objectives are met more effectively.

When it comes to large-scale training, scalability becomes a critical factor, and GenAI excels in this area by providing efficient, automated solutions that maintain quality across a vast number of participants. Traditional training methods often struggle to deliver personalized experiences at scale, but GenAI overcomes this challenge by leveraging its ability to generate customized content and feedback for each learner, even in a large cohort. By using cloud-based platforms and advanced algorithms, GenAI can simultaneously manage thousands of learners, delivering tailored learning paths, assessments, and feedback without overwhelming resources. This scalability ensures that every participant receives a high-quality learning experience, regardless of the size of the training program.

GenAI also enhances scalability in large-scale training by enabling asynchronous learning, allowing participants to engage with training materials at their own convenience. This flexibility is particularly beneficial for organizations with geographically dispersed teams or varying schedules, as it eliminates the need for synchronous, in-person training sessions that can be logistically challenging and costly. By providing access to a wide range of learning materials and resources on-demand, GenAI ensures that learners can engage with the content whenever and wherever it is most convenient for them, further enhancing the scalability and reach of the training program.

In addition to scalability, GenAI's flexibility allows it to cater to a variety of learning preferences and needs across different training contexts. For individual learners, the AI can adapt content to match their preferred learning style—whether visual, auditory, or kinesthetic—ensuring that the training is both engaging and effective. In small group settings, GenAI can accommodate different learning speeds and preferences within the group, allowing for a more supportive learning environment. For large-scale training programs, the AI can offer a range of content formats and learning pathways, catering to a broad audience with varying needs and preferences, ensuring that every participant can find a learning path that suits them best.

Moreover, the use of GenAI in training programs allows for continuous improvement and adaptation based on learner feedback and performance data. For individuals, this means that the training content and approach can be constantly refined to better suit their needs. In small group training, GenAI can adjust its strategies based on group performance and feedback, ensuring that the training remains relevant and effective. For large-scale programs, the AI can analyze data from thousands of learners to identify common challenges or trends, using this information to refine and improve the training content and delivery over time. This ability to continuously adapt and improve enhances both the scalability and flexibility of GenAI-powered training programs.

Resource Allocation and Cost Efficiency

Resource allocation and cost efficiency are essential considerations, especially for large-scale collective training systems. The deployment of GenAI solutions must be cost-effective, providing high-quality training at scale without incurring prohibitive expenses. This includes optimizing computational resources, managing cloud storage effectively, and ensuring that the AI models are not overly complex or resource intensive. For individual and small group training, while the cost may be less of an issue, the allocation of resources must still be managed effectively to provide a high-quality, adaptive training experience without

unnecessary expenditure. GenAI can significantly optimize resource allocation and cost efficiency across different training contexts, whether for individual learners, small groups, or large-scale training programs just as it has been applied to program management (Obana, 2024). By leveraging advanced data analysis, content generation, and adaptive learning technologies, GenAI streamlines the training process, reduces overhead costs, and maximizes the impact of training investments. This optimization is crucial for organizations looking to enhance their training programs while maintaining budget constraints and ensuring effective use of resources.

In the context of individual training, GenAI can reduce costs by automating the creation of personalized learning materials and assessments (Alier et al., 2024). Traditionally, developing customized content for each learner requires significant time and resources from instructors and content developers (Colby et al., 2024). However, GenAI can dynamically generate tailored exercises, quizzes, and learning modules based on each learner's progress and needs, reducing the need for extensive human involvement. This automation not only cuts down on content development costs but also ensures that learners receive the most relevant and engaging materials without additional expenditure. Furthermore, the use of AI-driven platforms reduces the need for one-on-one instruction, as the AI can provide real-time feedback and guidance, allowing human trainers to focus on more strategic or complex tasks.

For small group training, GenAI optimizes resource allocation by enhancing the efficiency of group learning sessions. In a traditional setting, facilitating effective group training often requires a dedicated instructor to manage group dynamics, monitor participation, and provide feedback. GenAI can automate many of these tasks by analyzing group interactions and dynamically adjusting the training content to suit the group's collective needs. For example, if a small group is struggling with a particular concept, the AI can introduce supplementary materials or adjust the exercise difficulty, ensuring that all members are effectively engaged and learning. This reduces the need for extensive instructor involvement, allowing a single trainer to manage multiple groups simultaneously or focus on high-value coaching and mentoring, thus optimizing human resource allocation and reducing labor costs.

When it comes to large-scale training programs, GenAI offers substantial cost savings by enabling organizations to deliver high-quality training to a large number of participants without a proportional increase in costs. Traditional large-scale training often involves significant expenses related to travel, venue rentals, printed materials, and instructor fees. GenAI can mitigate these costs by offering digital, on-demand training solutions that are accessible from anywhere, eliminating the need for physical infrastructure and travel. Additionally, GenAI's ability to automate content delivery and assessment processes means that a single training program can be scaled to thousands of participants with minimal additional cost, ensuring that the marginal cost of training additional learners is very low.

GenAI also enhances cost efficiency by enabling asynchronous learning, allowing learners to access training materials at their own pace and convenience. This flexibility reduces the need for scheduled, instructor-led sessions, which can be logistically complex and costly to coordinate, especially for geographically dispersed teams. Asynchronous learning allows organizations to maximize the use of digital resources, reducing dependency on physical materials and in-person training sessions. This approach not only lowers the costs associated with training logistics but also ensures that learning is more accessible and convenient for all participants, enhancing overall engagement and learning outcomes.

Another way GenAI optimizes resource allocation is through continuous data-driven insights that enable organizations to refine their training programs and strategies. By analyzing learner performance and engagement data, GenAI can identify which training materials and methods are most effective, allowing organizations to focus their resources on what works best. For individual learners, this means that resources can be allocated to develop more targeted and effective content that addresses specific learning gaps. In small group settings, GenAI can provide insights into group dynamics and progress, helping instructors

allocate their time and attention more effectively. For large-scale programs, these insights can guide the development of future training initiatives, ensuring that resources are invested in areas that deliver the highest return on investment.

Furthermore, GenAI can support predictive resource allocation by forecasting training needs and outcomes based on historical data and trends. This predictive capability allows organizations to anticipate resource requirements more accurately, avoiding over- or under-allocation of training materials, personnel, and other resources. For example, if GenAI identifies that a particular training module consistently requires additional support or materials, organizations can plan accordingly, ensuring that these resources are available when needed without incurring unnecessary costs. This proactive approach to resource management helps organizations optimize their training budgets and enhance cost efficiency.

In conclusion, applying GenAI to adaptive training requires careful consideration of various critical dimensions across different training types. Understanding these dimensions helps in designing GenAI solutions that are not only effective and efficient but also secure, scalable, and adaptive to the needs of a variety of learners. By addressing these factors, organizations can leverage GenAI to enhance the training process, improve learning outcomes, and reduce costs, ultimately achieving more adaptive and effective training programs.

Considering Task Domains in the Application of GenAI to Adaptive Training

GenAI has the potential to transform learning across various task domains, including cognitive, affective, psychomotor, and team-based tasks, by providing tailored and adaptive training experiences that cater to the unique requirements of each domain. By leveraging its ability to analyze vast amounts of data, create personalized learning paths, and deliver real-time feedback, GenAI enhances the effectiveness of training programs, ensuring that learners develop the skills and competencies needed to excel in a variety of areas.

For cognitive tasks, which involve mental processes such as problem-solving, critical thinking, and decision-making, GenAI can optimize learning by providing adaptive and personalized content that matches the learner's cognitive level and learning style. The AI can present complex scenarios and challenges that require the learner to apply knowledge, analyze information, and make decisions, dynamically adjusting the difficulty level based on the learner's performance. This personalized approach ensures that learners are constantly challenged without being overwhelmed, promoting deeper understanding and mastery of cognitive skills. Additionally, GenAI can offer immediate feedback and explanations, helping learners understand their mistakes and refine their thinking processes. This targeted feedback is crucial for developing higher-order cognitive skills, as it encourages learners to reflect on their reasoning and improve their problem-solving strategies over time.

In the domain of affective tasks, which focus on emotions, attitudes, and values, GenAI can create immersive and emotionally engaging learning experiences that foster empathy, self-awareness, and emotional intelligence. By using natural language processing (NLP) and sentiment analysis, GenAI can gauge a learner's emotional responses and adjust the content accordingly to address their emotional state and learning needs. For instance, the AI can present scenarios that challenge learners to navigate ethical dilemmas, manage conflict, or understand different perspectives, providing a safe environment for them to explore their emotions and attitudes. GenAI can also offer reflective prompts and personalized feedback that encourage learners to think about their feelings and responses, fostering greater emotional awareness and growth. This focus on the affective domain helps learners develop the emotional and interpersonal skills necessary for effective communication, leadership, and teamwork.

When it comes to psychomotor tasks, which involve physical skills and coordination, GenAI can optimize

learning by providing interactive and experiential training that allows learners to practice and refine their motor skills in a controlled and safe environment. The AI can generate virtual simulations and scenarios that mimic real-world tasks, enabling learners to develop their skills through repeated practice and experimentation. For example, in fields like surgery, aviation, or sports, GenAI can create realistic simulations that require learners to perform precise physical actions, providing instant feedback on their performance. This immediate feedback helps learners understand their mistakes and make the necessary adjustments to improve their technique. Moreover, GenAI can track a learner's progress over time, identifying areas that need further practice and providing targeted exercises to help them develop their motor skills more effectively. This approach not only accelerates skill acquisition but also ensures that learners develop the competence and confidence needed to perform psychomotor tasks successfully.

For team tasks, which require collaboration, communication, and coordination among multiple individuals, GenAI can optimize learning by facilitating dynamic and adaptive group training experiences. The AI can create collaborative scenarios and simulations that require team members to work together to achieve a common goal, fostering the development of teamwork skills such as communication, leadership, and conflict resolution. By analyzing team dynamics and interactions, GenAI can provide real-time feedback and suggestions to help teams improve their collaboration and performance. For example, if a team struggles with communication or coordination, the AI can provide targeted interventions, such as role-playing exercises or communication drills, to help them develop these critical skills. Additionally, GenAI can offer individual feedback to each team member, highlighting their strengths and areas for improvement, ensuring that all members contribute effectively to the team's success. GenAI's ability to integrate across multiple task domains enhances its potential to deliver comprehensive training programs that address cognitive, affective, psychomotor, and team-based skills simultaneously. For instance, a training scenario could require learners to solve a complex problem (cognitive), navigate a challenging interpersonal situation (affective), perform a specific physical task (psychomotor), and coordinate with others (team). By providing a holistic and integrated approach to training, GenAI ensures that learners develop a well-rounded skill set that prepares them for the multifaceted challenges of real-world tasks.

Considering Simulation Domains in the Application of GenAI to Adaptive Training

GenAI can revolutionize learning across various simulation domains—live, virtual, constructive, and hybrid—by providing adaptive, personalized, and immersive training experiences that enhance the realism, effectiveness, and efficiency of simulation-based training. These simulation domains, each with unique characteristics and requirements, offer distinct advantages for different training contexts, and GenAI's ability to optimize content and interactions within these environments ensures that learners gain the most from their training experiences.

In the context of live simulations, where learners interact in real-world settings with real equipment and scenarios, GenAI can optimize learning by enhancing realism and providing real-time feedback. Live simulations are particularly effective for developing practical skills and decision-making under realistic conditions. GenAI can analyze data from sensors, cameras, and other inputs to monitor learner performance, identify errors, and provide immediate, personalized feedback. For example, in a live medical simulation, GenAI could monitor a trainee's actions during a surgical procedure and offer instant guidance or corrections if an error is detected. This real-time feedback ensures that learners understand and rectify mistakes as they occur, enhancing skill acquisition and confidence. Additionally, GenAI can create dynamic scenarios that evolve based on the learner's actions, providing a more realistic and challenging training environment that better prepares them for real-world situations.

For virtual simulations, where learners engage in computer-generated environments, GenAI can optimize learning by creating highly immersive and interactive experiences that replicate real-world conditions

without the associated risks or costs. Virtual simulations are ideal for tasks that require repeated practice or involve high-risk situations, such as flight training or emergency response. GenAI can generate realistic virtual environments and scenarios tailored to the specific training needs of each learner. For instance, in a flight simulation, GenAI can simulate various weather conditions, mechanical failures, or emergency situations, adapting the difficulty level based on the learner's proficiency. The AI can also provide detailed, context-specific feedback, helping learners understand their performance and improve their skills. This adaptive approach ensures that learners are constantly challenged and engaged, promoting deeper learning and skill mastery.

In constructive simulations, where virtual entities and environments are used to simulate complex systems and scenarios, GenAI can optimize learning by providing sophisticated models and scenarios that replicate complex real-world systems. Constructive simulations are often used for strategic planning, decision-making, and systems analysis, where the focus is on understanding the interactions between different components of a system rather than individual performance. GenAI can create complex, data-driven models that simulate various scenarios and outcomes, allowing learners to explore different strategies and their potential impacts. For example, in military training, GenAI can simulate a battlefield scenario with multiple variables—such as terrain, enemy movements, and logistical constraints—allowing learners to test different tactics and strategies. By analyzing the outcomes of these simulations, learners can develop a deeper understanding of complex systems and improve their decision-making skills.

In hybrid simulations, which combine elements of live, virtual, and constructive simulations, GenAI can optimize learning by seamlessly integrating different simulation types into a cohesive training experience. Hybrid simulations are particularly effective for comprehensive training programs that require learners to develop a wide range of skills in varied contexts. GenAI can dynamically switch between different simulation types based on the learner's progress and training needs. For instance, a medical training program might begin with a virtual simulation to teach the basics of a procedure, then transition to a live simulation for hands-on practice, and finally use a constructive simulation to analyze the outcomes and refine strategies. This integrated approach ensures that learners develop a well-rounded skill set, combining theoretical knowledge with practical experience and strategic thinking.

Moreover, GenAI's ability to provide adaptive learning paths across these simulation domains further enhances the effectiveness of simulation-based training. By continuously analyzing learner performance and adapting the training content accordingly, GenAI ensures that each learner receives a personalized learning experience that is tailored to their specific needs and goals. This adaptive approach is particularly valuable in simulation-based training, where the ability to practice and refine skills in a controlled environment is crucial for developing expertise. Whether it is adjusting the difficulty level of a virtual simulation, introducing new challenges in a live simulation, or modifying the parameters of a constructive simulation, GenAI ensures that learners are constantly challenged and engaged, promoting continuous improvement and skill mastery. Additionally, GenAI can optimize resource allocation across simulation domains by reducing the need for expensive equipment, facilities, and human instructors. By automating the creation of simulation scenarios and providing real-time feedback, GenAI reduces the reliance on costly physical resources and human labor, making simulation-based training more accessible and cost-effective. This optimization is particularly important for organizations with limited training budgets or resources, as it allows them to deliver high-quality training without incurring significant costs.

Considering Phases of Training in the Application of GenAI to Adaptive Training

The purpose of adaptive training is to provide tailored learning experiences without intervention from a human teacher (Sottolare, 2024). Adaptive training is an approach that uses AI to monitor trainee progress and select optimal interventions. GenAI can profoundly enhance learning across the various phases of

training—planning, preparing, executing, and reviewing—by providing intelligent, adaptive support tailored to each stage. This optimization ensures that training is not only more efficient but also more effective, enabling learners and organizations to achieve their goals with greater precision and insight.

In the planning phase, GenAI plays a crucial role by assisting in the design and development of training programs. It can analyze organizational needs, learner profiles, and training objectives to generate comprehensive and customized training plans. For example, GenAI can evaluate historical data on training effectiveness and learner performance to suggest the most appropriate training methods, content, and delivery formats. This data-driven approach ensures that the training plan is well-aligned with the specific needs of the organization and its learners. Additionally, GenAI can simulate various training scenarios and predict potential outcomes, helping planners identify the most effective strategies and anticipate challenges before the training begins. This thorough planning process lays the foundation for a more targeted and impactful training experience.

During the preparation phase, GenAI optimizes learning by streamlining the development and organization of training materials and resources. It can automate the creation of training content, such as modules, exercises, and assessments, tailored to the needs of the learners identified in the planning phase. For instance, if the training involves technical skills, GenAI can generate interactive simulations and practice exercises based on the specific tasks that learners will need to perform. Furthermore, GenAI can assist in logistics by coordinating schedules, organizing resources, and ensuring that all necessary materials are prepared and accessible. By reducing the time and effort required to prepare for training, GenAI allows instructors and learners to focus more on the actual learning experience rather than administrative tasks.

In the execution phase, GenAI enhances learning by providing real-time support and feedback during the training sessions. It can monitor learners' progress and performance, offering immediate, context-specific feedback to guide them through complex tasks and scenarios. For example, in a live simulation, GenAI can analyze participants' actions and decisions, providing instant corrections and suggestions to help them improve. Additionally, GenAI can facilitate interactive learning by adapting the difficulty of tasks based on learners' performance, ensuring that they are continuously challenged without becoming overwhelmed. This dynamic support enhances engagement and ensures that learners are able to apply their knowledge and skills effectively during the training.

Following the training sessions, the reviewing phase is crucial for assessing the effectiveness of the training and identifying areas for improvement. GenAI optimizes this phase by analyzing data collected during the training, including learner performance metrics, feedback, and outcomes. It can generate detailed reports and insights on how well the training objectives were met, highlighting strengths and pinpointing areas that need further attention. For instance, GenAI can identify common errors or challenges faced by learners and suggest specific modifications to improve future training sessions. Additionally, GenAI can gather and analyze feedback from participants to assess their satisfaction and identify any gaps in the training content or delivery. This comprehensive review process ensures that training programs are continuously refined and enhanced based on data-driven insights.

Moreover, GenAI's capability to provide continuous learning and improvement across these phases ensures that the training process is iterative and evolving. Insights gained from the reviewing phase can be fed back into the planning phase to inform the design of future training programs, creating a cycle of ongoing improvement. By integrating learner feedback and performance data into the planning and preparation stages, GenAI helps organizations continuously adapt and optimize their training strategies, ensuring that they remain effective and relevant.

Projecting Roles for GenAI Applications in Adaptive Training

Projecting future uses of GenAI, we predict GenAI will offer unparalleled personalization, scalability, and interactivity compared to other AI and non-AI methods. From generating real-time scenarios to supporting cognitive and affective development, GenAI will revolutionize how we approach learning, making it more engaging, efficient, and responsive to individual needs. As the technology continues to evolve, adaptive training powered by GenAI will become an indispensable tool in industries ranging from education and healthcare to the military and beyond.

Hyper-Personalized Learning Experiences

Hyper-personalized learning experiences can help learners achieve their goals by tailoring a unique learning event based on each learner's interests, experiences, learning history, and goals. Tailored learning events can include customized learning pathways, personalized recommendations and feedback, and individualized monitoring of progress toward assigned goals (Fancsali et al., 2013). Generative AI can offer many paths to personalization, far beyond what current adaptive training systems provide. By leveraging deep data analysis and real-time feedback, GenAI can generate custom learning paths that adapt not only to the learner's current skills and performance but also to their learning preferences, emotional states, and cognitive abilities. Duckworth et al. (2007) found that learners high in grit may benefit more from "rapid increases in the difficulty of course material compared to students low in grit, regardless of knowledge-levels". For example, a medical trainee might experience a training simulation that rapidly adjusts content difficulty based on their emotional stress levels. This stress could be scaffolded by supportive, tailored feedback from the tutor, and then gradually reduce scaffolding to enable trainees to perform under pressure.

Immersive and Real-Time Scenario Generation

GenAI will be capable of creating complex, real-time training scenarios that mimic real-world conditions with a high degree of realism. Virtual training simulations, in particular, will benefit from dynamic environments that continuously evolve based on learner actions. For instance, a military simulation could change weather conditions, ambient light, and terrain features, offering an infinite variety of scenarios to better prepare military commanders and units for unpredictable situations (Sottolare et al., 2024). The use of generative models will enable these simulations to learn from past sessions, making each iteration more refined and effective.

Seamless Integration of Multiple Training Domains

Generative AI could facilitate hybrid training that integrates live, virtual, and constructive simulations into a unified learning environment. Future adaptive training platforms can use GenAI to combine elements of real-world training (live) with virtual environments (Virtual Reality/Augmented Reality (VR/AR)) and strategic planning (constructive) into a single, cohesive experience. This will allow trainees to seamlessly move between simulations and real-world practice, blurring the lines between different domains and making training far more effective and fluid.

Continuous Learning and Real-Time Adaptation

In the coming years, Generative AI will enable adaptive training systems to continuously evolve and improve in real-time. These systems will assess performance as learners progress through a training session, adjusting content and scenarios instantaneously to suit their needs. This feedback loop will allow for highly

interactive learning environments where mistakes are corrected immediately, and learners can continuously practice in a tailored, error-tolerant environment. In industries like aviation or healthcare, where mistakes can be costly, this real-time adaptability will be a game-changer.

Enhanced Collaboration and Team-Based Learning

GenAI will play a key role in optimizing collaborative and team-based training environments. By generating AI-driven virtual teammates or opponents that adapt to scenario conditions, the system can simulate complex interpersonal and collaborative dynamics. Teams will be able to practice decision-making, communication, and problem-solving in highly realistic settings, with AI providing tailored feedback to each participant. For example, an AI might simulate a malfunctioning communication system in a team-based aviation scenario, forcing participants to adapt their strategies on the fly, with GenAI tracking individual contributions and providing personalized feedback.

Emotional and Psychological Adaptation

One of the anticipated future breakthroughs in adaptive training will be GenAI's ability to gauge and respond to the emotional and psychological states of trainees. AI will be able to analyze multimodal inputs such as facial expressions, voice tones, and body language, and then use this data to adapt the learning experience in real-time. If a learner is showing signs of frustration or stress, the AI could adjust the difficulty level or generate motivational support. Instructors might also receive AI-driven insights on the emotional readiness of their students, allowing for more empathetic and supportive teaching approaches.

Cost-Effective, Scalable Training Solutions

As GenAI becomes more advanced, it will drive down the costs associated with creating and delivering highly adaptive training programs. AI will be able to generate training content, manage logistics, and deliver personalized experiences to vast numbers of learners simultaneously. This scalability will allow the military to implement sophisticated adaptive training solutions that reflect the complexity of operational environments.

Lifelong Learning Ecosystems

In the future, GenAI could facilitate continuous, lifelong learning. Rather than being confined to traditional training periods, adaptive training platforms will evolve into ongoing ecosystems that track an individual's learning journey over the course of their career. These systems will leverage GenAI to create personalized learning content based on sophisticated long-term learner models. Long-term learner models will be standardized across platforms to identify gaps in knowledge and skills, and schedule refresher training as needed. This lifelong approach to learning will ensure that individuals remain competent and up to date with the latest required skills throughout their professional careers.

Automated Training Design and Evolution

Future GenAI systems will be capable of not just delivering training but designing it autonomously. These AI systems will analyze the needs of an organization or individual and create tailored training programs from scratch, continuously refining them based on individual learner or unit performance data. This will significantly reduce the time and resources required to develop training materials while ensuring that the

programs are as effective as possible. Over time, the AI will learn from each iteration, making future training sessions more efficient and impactful.

Ethical and Transparent Learning Systems

As GenAI becomes more integrated with adaptive training, ethical considerations will also become paramount. Future systems will need to prioritize transparency and fairness, ensuring that the AI-driven decisions made during training are explainable and fair. Learners and organizations will demand insight into how GenAI algorithms adjust and generate recommendations, fostering trust in these training systems. This emphasis on ethical AI will ensure that future adaptive training platforms are not only powerful but also aligned with organizational values and protection of individual data.

Discussion and Recommendations for Future Research

Future GenAI research for adaptive training holds immense promise to revolutionize learning and development across various fields, but it also presents a wide array of opportunities for future exploration. Below are key areas where further research could deepen our understanding and enhance the practical applications of GenAI in adaptive training environments such as those created using the Generalized Intelligent Framework for Tutoring (GIFT) and Synthetic Training Environment Experiential Learning – Readiness (STEEL-R). Depending on the training conditions, the application of GenAI to GIFT and STEEL-R should consider the type of training, the simulation domain(s) supporting the training experience, the task domain, and the phase of training. Our recommendations for GenAI methods based on training conditions is summarized in Table 1.

Table 1. Recommendations for the Application of GenAI Methods in an Adaptive Training Context

Training Conditions			
Training Types	Adaptive Training Design Goals	Generative AI Methods	Generative Output
Individual Training	Personalized content generation, adaptive learning paths, real-time feedback, automated assessments	Variational Autoencoders (VAEs)	Personalized training scenarios based on trainee profiles
Small Group Training	Adaptive group dynamics, collaborative scenarios, real-time analytics, personalized group feedback	Recurrent Neural Networks (RNNs) with LSTM	Simulated group dialogues and interaction patterns
Large-Scale Training	Scalable content generation, automated progress tracking, on-demand access, batch analytics	Transformer Models (Large Language Models)	Scalable text generation for mass learning content
Simulation Domains			
Simulation Domains	Adaptive Training Design Goals	Generative AI Methods	Generative Output
Live Simulations	Real-time performance monitoring, dynamic scenario adaptation, immediate feedback, sensor integration	Generative Adversarial Networks (GANs)	Realistic environmental conditions and scenario variations
Virtual Simulations	Immersive environment creation, interactive content generation, adaptive difficulty adjustment, contextual feedback	Diffusion Models	High-quality, dynamic virtual environments
Constructive Simulations	Complex system modeling, scenario generation, data-driven insights, strategic simulations	Normalizing Flows	Complex, data-driven scenario modeling
Hybrid Simulations	Integrated simulation management, seamless domain transitions, combined feedback mechanisms, adaptive scenario adjustments	Conditional Generative Models	Context-specific scenarios combining multiple simulation types
Task Domains			
Task Domains	Adaptive Training Design Goals	Generative AI Methods	Generative Output
Cognitive Tasks	Adaptive problem-solving scenarios, personalized cognitive challenges, real-time decision support, targeted feedback	Transformer Models	Adaptive problem-solving exercises and quizzes
Affective Tasks	Emotional sentiment analysis, reflective prompts, empathy-building scenarios, personalized emotional	Neural Style Transfer	Emotionally engaging content tailored to learner states
Psychomotor Tasks	Interactive skill practice, virtual hands-on exercises, real-time performance feedback, technique refinement	Deep Convolutional Generative Adversarial Networks (DCGANs)	Realistic simulations for hands-on practice
Team Tasks	Collaborative scenario design, team dynamics analysis, role-based feedback, group performance tracking	Autoregressive Models (e.g., GPT)	Collaborative task scenarios and team-based dialogues
Phases of Training			
Phases of Training	Adaptive Training Design Goals	Generative AI Methods	Generative Output
Planning	Data-driven training design, scenario simulations, needs analysis, adaptive curriculum development	Diffusion Models	Diverse, simulated planning scenarios for strategy testing
Preparing	Automated content creation, logistics coordination, resource management, personalized pre-training assessments	Variational Autoencoders (VAEs)	Tailored pre-training materials and resources
Executing	Real-time feedback, adaptive difficulty, interactive learning support, performance monitoring	Generative Adversarial Networks (GANs)	Real-time adaptive training environments
Reviewing	Data analysis and reporting, performance insights, feedback aggregation, continuous improvement recommendations	Normalizing Flows	Analytical insights and optimized review simulations

A recommended next step in the process of applying GenAI methods to adaptive training is to develop a recommender for the selection of GenAI methods to perform specific tasks (e.g., automated scenario generation, performance assessment, intervention selection, or after-action review generation) during one of the training phases. A related recommendation is to begin the development of multimodal LLMs (large language models) that are tailored to support various training phases, task domains, simulation domain combinations, and different echelons of military training. In a military adaptive training context, multimodal LLMs can play a transformative role by integrating various forms of data (text, audio, video, sensor inputs, etc.) to provide personalized and realistic training experiences. Below is a set of exemplars for the use of multimodal LLMs to support adaptive training:

Real-Time Scenario Generation and Adaptation Use Case

GenAI research should investigate how GenAI can be improved to dynamically create and adapt training scenarios in real-time, based on learner performance and environmental conditions. We recommend the development of methods that enable adaptive instructional systems (AISs) to continuously generate new, context-aware scenarios that adapt to individual learners and teams. Exploring reinforcement learning techniques for real-time scenario adaptation in live and virtual training environments could provide more flexible, personalized, and responsive training environments, ensuring learners face challenges that match their real-time operational requirements for performance.

Dynamic Scenario Generation for Commander and Staff Training Use Case

In this use case, multimodal LLMs could generate highly detailed and adaptive training scenarios or recommended simulation adaptations in real-time, integrating textual orders, visual battlefield simulations, and audio cues. For example, a commander's briefing could be textually described, combined with 3D maps, satellite imagery, and AI-generated video feeds to create a lifelike simulation. If a trainee (e.g., unit commander; Figure 1) implements a decision, the multimodal LLM can adapt the scenario (e.g., simulation state) by generating enemy countermeasures, audio communications from subordinates, and visual feedback of battlefield changes.

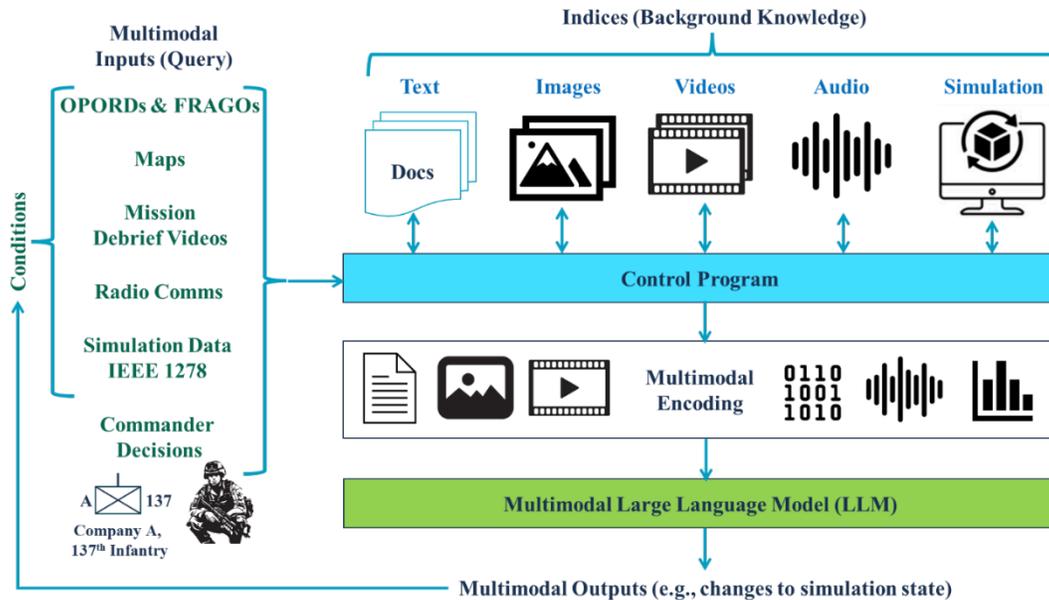


Figure 1. Exemplar Multimodal LLM Use Case

Personalized Training Feedback Use Case

LLMs can provide personalized feedback based on a Soldier's performance across different modalities, such as spoken commands, decision-making in text inputs, or actions taken in virtual simulations. If a trainee engages in a virtual operation and provides verbal commands to subordinates (audio input), the LLM can assess the accuracy, timeliness, and effectiveness of the commands. It can then offer corrective feedback or highlight strengths through both text and visual aids (e.g., map or satellite images).

Interactive Staff Use Case

Multimodal LLMs can simulate interactions with a commander's staff, providing inputs such as intelligence reports (text), radio communications (audio), and live drone feeds (video). The LLM can interpret the trainee's responses and adjust the scenario accordingly to optimize learning and engagement. In a brigade commander's training exercise, the S-staff might simulate intelligence inputs (via text), logistical updates (via voice), and operational briefings (via interactive maps or videos). The LLM could then dynamically change the complexity of the scenario based on the decisions the commander makes (Sottolare, 2024, December).

Threat Recognition and Response Use Case

Multimodal LLMs can process and interpret data from multiple sensor types (video, audio, text reports) in real-time to simulate battlefield conditions, where trainees must recognize and respond to threats. A multimodal LLM could integrate data from surveillance footage, sensor data from drones, and radio chatter to present the commander with evolving battlefield situations. The commander's decisions would alter the flow of the scenario (including the adaptive behavior of agentic computer generated forces in the simulation; Sottolare et al., 2023). For example, adaptive behaviors could be driven by a LLM adapting the sequence of simulation events by presenting new challenges (e.g., an ambush or supply shortages) based on their choices.

Language and Cultural Training Use Case

In military operations, effective communication with local populations or foreign military units is critical. Multimodal LLMs can simulate conversations and interactions across different languages and cultures. Soldiers could be trained to engage with locals during humanitarian operations through simulated conversations rather than live actors. The LLM could integrate speech recognition, natural language understanding (NLU), and cultural context to provide feedback on the Soldier's choice of words, tone, and cultural appropriateness.

Cognitive Load Management and Decision-Making Use Case

Multimodal LLMs may be used to assess the cognitive load on a trainee by monitoring how they process and respond to multiple forms of input (e.g., video feeds, audio reports, operational texts) and adapt training scenario conditions accordingly. In a high-stress environment, like a battlefield simulation, the LLM can track how a trainee processes incoming information from different sources. It can slow down or speed up the simulation based on how the trainee manages the workload, ensuring that training matches their capacity to process the data.

Advanced Natural Language Interaction Use Case

A multimodal LLM allows for natural language interactions with simulated staff, AI avatars, or virtual subordinates, providing realistic dialogues during training scenarios. LLMs can also analyze multimodal data from the training exercise (entity data, audio, video, textual logs) and provide a comprehensive After Action Review (AAR) report detailing strengths, weaknesses, and areas for improvement. The LLM could also generate an AAR that includes text summaries of the trainee's decisions, video highlights of key moments, and audio playback of critical communications. This data can be used to pinpoint errors, reinforce lessons, and suggest targeted improvements.

It might also be used to conduct AARs based on "teachable moments" identified for inclusion in the AAR, and then interact with trainees based on their questions and responses to AAR questions. For example, during a command-and-control exercise, the LLM could simulate interactions with S-staff officers who provide verbal intelligence updates, operational feedback, or logistical requests. The trainee's verbal responses are processed by the LLM, which can dynamically update the scenario based on the given instructions or identify key behaviors, events, patterns, and trends for post-training discussion. Responses can be text only or multimodal.

Integration Extended Reality (XR) Use Case

Multimodal LLMs can be used to create immersive, adaptive environments where visual (virtual, augmented, or mixed reality (VR/AR/MR)), audio, and text data are combined for a realistic training experience. A Soldier might train in a VR environment where they receive spoken commands (audio), see an evolving battlefield (visual), and interact with virtual objects (e.g., vehicles, enemy forces). The LLM can adapt the scenario conditions (e.g., weather, mission goals, OPFOR (opposing force) or resources) based on the Soldier's movements and decisions, offering a tailored training experience that changes in real-time.

Adaptive Mission Rehearsal Use Case

Before actual operations, LLMs could be used to run adaptive mission rehearsals, where multimodal inputs simulate conditions that the military personnel will likely face. For an upcoming mission, the LLM can create realistic simulation scenarios that incorporate real-world data (e.g., satellite imagery, intelligence reports, weather conditions) to provide an accurate rehearsal environment. The LLM can change mission parameters, providing adaptive learning by reacting to trainee decisions and actions in real-time.

Conclusions

This chapter examined GenAI methods in the context of adaptive training principles. The goal was to consider training features such as the scale of training, task and simulation domains, and the phases of training. GenAI methods were discussed to determine their ability to customize/personalize training experiences. GenAI has the potential to automate many training processes used to plan, prepare, execute, and review adaptive training experiences. While this chapter has focused primarily on the merits of applying GenAI methods to adaptive training, we should also note the current limitations of GenAI approaches for simulation-based training. Five critical limitations include 1) a lack of real-time adaptation and feedback fidelity, 2) the risk of hallucinations, 3) the availability of domain-specific training data, 4) methods to easily validate GenAI outputs, and 5) latency associated with GenAI solutions in real-time training systems.

GenAI models, especially LLMs, often struggle with delivering context-sensitive feedback that reflects real-time situational complexity. Simulation-based training requires dynamic and adaptive responses based on user actions, which may not be consistently achievable due to limitations in temporal awareness and sensor integration. Strategies to overcome a lack of real-time adaptation and feedback include the integration of GenAI with rule-based systems or state machines to ensure critical constraints and logic are preserved or context anchoring where structured inputs such as scenario metadata, learner performance metrics, or sensor data are used to contextualize generative responses.

Generative models can produce plausible but incorrect or misleading content (“hallucinations”), which can be detrimental in training environments that require factual accuracy, especially in high-stakes training domains like military, medical, or aviation simulations. We can reduce the frequency of hallucination incidents by fine-tuning generative models with domain-specific data. This can be achieved by connecting the AI model to a vetted knowledge base to ensure outputs are grounded in authoritative sources, implementing confidence thresholds or tightly controlling prompts to limit the generation space.

Effective deployment in simulation-based training contexts often demands extensive fine-tuning with domain-specific data. This data may be limited, sensitive, or expensive to obtain, and without it, Generative AI models may underperform or introduce unrealistic elements that reduce training validity. Mitigation strategies for limited domain-specific data have focused on the use of subject matter experts (SMEs) to

build representative datasets to support fine-tuning, synthetic data generation based on credible simulated scenarios, and federated learning that benefits from distributed datasets.

Simulation-based training programs, especially those used in defense or healthcare, often require certification or accreditation. The non-deterministic nature of GenAI makes it difficult to validate outputs consistently, thereby complicating the verification process and regulatory approval. Verification and accreditation can be facilitated by implementing traceability, benchmarking and modular certification strategies. GenAI system designers should ensure all generative outputs are logged with versioning, prompt history, and context metadata to support audits and explainability. Risk can be further reduced by creating a suite of standardized test scenarios for validating AI behavior against expected outcomes and by certifying generative components in isolation, under very specific constraints, rather than attempting to certify the entire system.

Finally, integrating GenAI into simulation frameworks can demand substantial computational resources and architectural modifications. Real-time rendering of AI-generated content (e.g., voice, narrative, or character behavior) may require latency-optimized systems that increase cost and system complexity. Three mitigation strategies that should be considered are edge deployment with model compression, cloud-hybrid architectures, and incremental integration approaches. AI system designers should consider the use of optimized, quantized, or distilled models suitable for real-time inference on local systems, seek to balance compute loads between cloud-based resources for training and local systems for inference, and by phasing AI capabilities into existing simulation platforms through modular application programming interfaces (APIs) and sandbox environments to test compatibility and computational performance.

With these design recommendations and mitigation strategies in mind, we highly anticipate the emergence of more effective GenAI tools and methods. The expected result is the more efficient, flexible and tailored adaptive training experiences along with more productive learning engineers.

References

- Alier, M., García-Peñalvo, F., & Camba, J. D. (2024). Generative Artificial Intelligence in Education: From Deceptive to Disruptive.
- Bengesi, S., El-Sayed, H., Sarker, M. K., Houkpati, Y., Irungu, J., & Oladunni, T. (2024). Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers. IEEe Access.
- Colby, B., Tucker, E., & Siggins, T. (2024, June). Beyond Standalone Systems: Creating an Ecosystem of Adaptive Training Services. In *International Conference on Human-Computer Interaction* (pp. 3-14). Cham: Springer Nature Switzerland.
- Dong, H. W. (2024). Generative AI for Music and Audio. University of California, San Diego.
- Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. *Journal of personality and social psychology*, 92(6), 1087.
- Fancsali, S. E., Ritter, S., Stamper, J., & Nixon, T. (2013, July). Toward “hyper-personalized” Cognitive Tutors. In *AIED 2013 Workshops Proceedings Volume* (Vol. 7, pp. 71-79).
- Fregly, C., Barth, A., & Eigenbrode, S. (2023). Generative AI on AWS. O'Reilly Media, Inc.
- Grover, D. (2024). Next-Generation Education: The Impact of Generative AI on Learning. *Journal of Informatics Education and Research*, 4(2).
- Naseer, F., Khalid, M. U., Ayub, N., Rasool, A., Abbas, T., & Afzal, M. W. (2024). Automated Assessment and Feedback in Higher Education Using Generative AI. In *Transforming Education With Generative AI: Prompt Engineering and Synthetic Content Creation* (pp. 433-461). IGI Global.
- Obana, L. R. (2024). Collaboration of Artificial Intelligence and Project Management.
- Sottolare, R., Ballinger, C., & McGroarty, C. (2023, May). Considerations in the Design of Realistic Agents for Serious Games. In *The International FLAIRS Conference Proceedings* (Vol. 36).

- Sottolare, R. (2024, December). Max, A GPT Concept for Battalion Staff Training. Demonstrated at the Soar Technology booth (#138) during the Interservice/Industry Training Simulation & Education Conference, Orlando, Florida.
- Sottolare, R. A. (2024). Adaptive Learning, Training, and Education. In *Human-Computer Interaction in Various Application Domains* (pp. 144-172). CRC Press.
- Sottolare, R., Ballinger, C. B., Litvinas, M., Hu, S., & McGroarty, C. (2024, May). Using Genetic Algorithms to Automate Scenario Generation and Enhance the Training Value of Serious Games for Adaptive Instruction. In *The International FLAIRS Conference Proceedings* (Vol. 37).
- Yan, L., Martinez-Maldonado, R., & Gasevic, D. (2024, March). Generative artificial intelligence in learning analytics: Contextualising opportunities and challenges through the learning analytics cycle. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 101-111).

CHAPTER 7 - ADAPTIVE TRAINING SCENARIO GENERATION AND ASSESSMENT WITH GENERATIVE AI

Andy Smith¹, Randall Spain², Wookhee Min¹, Bradford Mott¹, and James C. Lester¹

North Carolina State University, Center for Educational Informatics¹;
US Army Combat Capabilities Development Command (DEVCOM) - Soldier Center²

Introduction

Scenario-based training, particularly through synthetic training environments, has long been a cornerstone of the US military's approach to preparing personnel for real-world operations. Synthetic training environments are designed to replicate tasks, equipment, and conditions found in the field, providing trainees with realistic experiences that mirror actual operational scenarios. By offering high levels of psychological and functional fidelity, synthetic training allows trainees to practice cognitive and behavioral skills critical to real-world performance in a safe and controlled setting. A key advantage of these immersive simulations is the ability to rehearse complex decision-making, critical thinking, and tactical skills, all within a structured environment that can be tailored to specific training objectives. Despite these advantages, creating simulation-based training experiences poses significant challenges. Authoring scenarios is resource intensive. It requires specialized knowledge of simulation engines, authoring tools, and scripting language. Scenarios often do not support adaptivity, either with respect to meeting the needs of the instructor or the needs of the trainees. Forms of adaptivity such as inserting scenario-injects are often performed manually. Further, assessments are rarely embedded into scenarios to support formative assessment and coaching. As the demand for more dynamic and personalized training experiences increases, the need for more efficient scenario development becomes even clearer.

Generative AI (GenAI) is revolutionizing the landscape of training technologies, offering new ways to enhance and streamline the creation of simulation-based scenarios for military training. At its core, GenAI refers to algorithms and models that can autonomously generate content, including text, images, and even immersive environments, by learning from large datasets. In the context of Army training, these systems can be harnessed to generate dynamic and personalized training scenarios that are specifically tailored to the needs of individual trainees and the training needs of units. By leveraging the capabilities of GenAI, unit leaders and training developers can rapidly craft synthetic training scenarios that push trainees to meet critical performance objectives, exercise essential skills, and navigate complex, interactive situations.

GenAI holds considerable promise as foundational technology for a key family of techniques that can provide the computational underpinnings for adaptive training scenario generation: AI-driven narrative generation. Research on AI-driven narrative generation has been underway for decades. This work includes research on functionalities for planning plots, generating character behavior, including dialogue, and generating settings in the form of virtual environments in which the interactive narratives play out. It spans narrative generation for text, narrative generation for animation, and narrative generation for creating rich, interactive immersive story-based experiences that deeply engage users. It has also explored AI-enabled authoring of interactive narratives, which can enable designers to specify narrative goals that narrative generation systems can then use to generate a scene or series of scenes. Narrative generation has been explored for applications in both education (e.g., game-based learning environments) and entertainment (games). Crucially, the emergence of GenAI has rapidly accelerated the development of the core computational capabilities that can drive real-time adaptive scenario generation for training.

This chapter explores the potential of GenAI for scenario generation, focusing on its benefits such as improved training efficiency, adaptability, and the ability to generate narratives with embedded assessments, which can enhance both engagement and learning outcomes. It first provides background on scenario generation with GenAI. It next discusses the role that stealth assessment can play in GenAI-based training scenario generation. It then turns to GenAI-based adaptive training scenario generation, including training scenario planning, offline training scenario generation (in which scenarios are generated in advance of trainee interactions with scenarios), and online scenario generation (in which scenarios are adaptively generated in real-time as trainees interact with them). Finally, it provides recommendations for a research agenda for the GIFT community, and it concludes with a summary of what the near-term future holds for GenAI-based training scenario generation.

Training Scenario Generation with Generative AI

Training scenario generation research builds on the expansive literature on narrative generation, GenAI advances, and recent work on GenAI-based narrative generation. Each of these is discussed in turn.

Narrative Generation

Narrative generation has its roots in natural language generation where the focus was on the generation of multi-paragraph stories (Callaway & Lester, 2002). Over the past two decades it has emerged as an active area of Artificial Intelligence (AI) research (Riedl & Bulitko, 2013). The inverse of devising natural language understanding methods for understanding stories, narrative generation focuses on developing models for creating stories. Mott et al.'s (1999) work on narrative-centered learning environments lays out a broad research agenda, emphasizing the integration of interactive storytelling with adaptive educational systems to enhance engagement and learning. This approach has direct applications in domains such as training, where immersive, high-stakes scenarios foster critical skills like decision-making and teamwork.. Narrative generators can operate on each of the three core elements of story: plot, character, and setting. First, to generate plots for narratives, narrative generators can be driven by authorial goals to plan plot structures that represent key events in a story, the high-level specifications for interactions of characters with each other and perhaps with the user, and rising and falling dramatic arcs (Kumaran et al., 2023). This process of creating a series of events is central to narrative generation but does not guarantee that the resulting stories will be coherent or engaging (Martin et al., 2016). Second, to generate character specifications for narratives, narratives generators can formulate abstract plans for character behaviors, and then pass control to character behavior planners, which in turn can create specifications for characters' actions in the story world (e.g., navigation, gesture, facial expression, posture, and story world asset manipulation). Character behavior planners can also generate characters' dialogue (natural language generation) and speech (speech synthesis). Characters' behaviors, which from a game perspective are the actions of non-player characters (NPCs), can be more or less autonomous, depending on the level of control afforded by planners (Oliver & Mateas, 2021). Third, to generate settings, which are the virtual environments in which plot events take place, narrative generators can expressively synthesize rich landscapes, complex virtual buildings (exterior and interior), and virtual game levels, all populated by navigable structures and manipulable artifacts. Over the past decade, significant advances have been made on procedural content generation, which is the task of generating settings for games, often with designer-in-the-loop architectures for mix-initiative "level" synthesis (Kumaran, Carpenter et al., 2024; Lai et al., 2022; Shaker et al., 2016; Yannakakis et al., 2014).

Research on narrative generation work has investigated a broad range of computational models. Central to much of this work has been a planning paradigm, where authorial goals are provided to a planner (a "drama manager"), which then formulates narrative plans with a library of operators that decompose high-level

narrative objectives into low-level character actions (Riedl & Bulitko, 2013). Narrative generators have been devised that use dynamic decision networks (Mott & Lester, 2006), reinforcement learning (Wang et al., 2017), and, as discussed below, large language models (LLMs) (Kumaran et al., 2023), which offer significant potential for GenAI-based adaptive scenario generation.

Generative AI

GenAI represents a transformative branch of AI focused on creating novel content by learning patterns from massive datasets (Sengar et al., 2024). It has gained prominence for its ability to produce outputs across multiple modalities, such as text, images, audio, and video, that closely mimic human-created content. GenAI systems model salient patterns from existing information to generate synthetic data, enabling creative applications in natural language processing, computer vision, scientific research, education, and training. These systems have revolutionized tasks ranging from automated content generation to creating immersive multimedia experiences.

Early advances in GenAI were driven by models such as generative adversarial networks (GANs) and variational autoencoders (VAEs). GANs consist of a generator and a discriminator operating in an adversarial framework to produce highly realistic visuals and audio (Goodfellow et al., 2014), while VAEs excel at generating structured, latent-space representations, making them ideal for applications like image synthesis and anomaly detection (Kingma, 2013). These foundational models have significantly expanded GenAI's capabilities, offering tools tailored to specific needs.

More recently, LLMs such as Generative Pre-trained Transformer (GPT), Large Language Model Meta AI (Llama), Gemini, and Claude have emerged as key innovations in GenAI (Zhao et al., 2023). Trained on vast corpora of text data using transformer architectures, LLMs undergo unsupervised pre-training to predict the next word in a sequence, followed by supervised fine-tuning or reinforcement learning from human feedback. A notable feature of LLMs is their ability to perform in-context learning adapting to new tasks based on prompts, often referred to as prompt engineering. LLMs also demonstrate zero-shot and few-shot learning capabilities: zero-shot learning relies solely on pretrained knowledge without explicit examples, while few-shot learning leverages a small number of examples within the prompt to guide responses in the desired format.

These innovations have enabled LLMs to excel across a spectrum of tasks (Maslej et al., 2024). For example, GPT-4 leads the Holistic Evaluation of Language Models (HELM) with a mean win rate of 0.96, demonstrating strengths in mathematical reasoning (GSM8K), multi-subject understanding (MMLU), and legal reasoning (LegalBench). Gemini Ultra achieved 90% on the Massive Multitask Language Understanding (MMLU) benchmark, surpassing the human baseline. AgentCoder, a GPT-4 variant, achieved a 96.3% score on HumanEval, demonstrating advanced programming capabilities. These achievements show the significant potential and broad impact of LLMs for a wide range of computational and creative domains.

GenAI-based Narrative Generation

Interactive narratives that provide players with agency and engaging experiences have long been a central focus in supporting human learning, such as AI-enabled narrative-centered learning environments (Lester et al., 2024; Rowe et al., 2011). As noted above, early AI planning-based methods sought to procedurally generate coherent event sequences, but they often fell short of delivering expansive and compelling stories (Porteous et al., 2010). Additionally, creating interactive narratives has traditionally demanded substantial manual effort, particularly in scripting and translating those scripts into playable game levels (Riedl & Bulitko, 2013).

GenAI introduces transformative possibilities for narrative generation, especially in domains where creativity and personalization are critical. Training environments, for example, can use GenAI to generate dynamic, immersive scenarios tailored to individual trainees. Text-based GenAI systems excel in crafting rich narratives, such as storylines and character dialogues, that align with specific training objectives. Furthermore, multimodal outputs, including animations, videos, and character gestures and emotes, enhance engagement by integrating visual and behavioral elements. By automating content creation, reducing production costs, and enabling remarkable levels of customization, these technologies augment human creativity while opening new frontiers in collaborative narrative and content creation.

The advent of LLMs has revolutionized narrative generation (Kumaran et al., 2023). With their expressive representations of world knowledge, instruction-following capabilities, and semantic event tracking, LLMs excel at generating narrative events, character trajectories, and end goals derived from input story arcs. Additionally, LLMs power mixed-initiative tools that enable collaboration between narrative designers and AI. Designers can refine interactive events (e.g., non-player character dialogues), define preconditions and postconditions for events, and position characters and props, all while ensuring alignment with their creative vision. For example, NarrativeGenie is an interactive narrative generation framework that builds on recent advances in LLM technologies (Kumaran, Rowe et al., 2024). By leveraging GPT-4, NarrativeGenie procedurally generates narrative beats as cohesive, partially ordered sequences of events. Through LLM-driven reasoning and generation, the framework translates high-level story overviews into narrative event graphs, facilitating player-driven narrative sequencing while adhering to the designer's vision. Furthermore, it supports dynamic runtime adaptation, enabling real-time player choices to shape the unfolding narrative. Non-player character interactions, including dialogues, emotes, and gestures, are automatically generated and seamlessly integrated into these events, enhancing immersion and creating deeply engaging experiences. Empirical evaluations, which combine human participant assessments and automated metrics, demonstrate NarrativeGenie's effectiveness in generating coherent narratives that align with designers' natural language instructions, while effectively balancing narrative structure and player agency.

Scenario Generation with Stealth Assessment

Stealth assessment is a non-intrusive approach to assessment designed to evaluate student competencies and support both learning and training without disrupting experiences by seamlessly embedding assessment tasks within engaging environments (Shute, 2011). Unlike traditional formative assessments, stealth assessment operates unobtrusively in the background, leveraging data collected from learners' interactions to infer competencies, skills, and knowledge without interrupting the flow of the experience. Stealth assessment supports an engaging and focused learning process by enhancing the authenticity of the experience while providing actionable insights to educators and adaptive systems.

Stealth assessment is grounded in evidence-centered design (ECD), which infers higher-level competencies from task-level evidence (Mislevy et al., 2003). ECD consists of three interconnected components: the competency model, which represents students' knowledge and skills and provides a framework for assessing their capabilities; the evidence model, which links observable student behaviors to underlying competency variables to facilitate accurate inferences; and the task model, which designs problems that elicit evidence of competencies and allow learners to demonstrate their skills in context.

Accurate assessment of teamwork and team performance is essential in achieving automated coaching, feedback, and competency-focused training within synthetic training environments. Traditional approaches have relied on external assessments, self-reports, or expert observations. One widely used technique in scenario- and simulation-based training events is the Targeted Acceptable Response to Generated Events and Tasks (TARGETs) methodology (Fowlkes et al., 1994). This behavioral observation-based approach

employs trained raters to evaluate whether team members exhibit key teamwork behaviors, such as exchanging information, offering backup and assistance, and initiating leadership during training scenarios.

Synthetic training environments expand upon traditional methods by incorporating data-driven techniques to facilitate unobtrusive measurement of team performance. Stealth assessment frameworks within these environments leverage rich interaction data to assess teamwork competencies such as information exchange, communication delivery, leadership/initiative, and supporting behaviors among others. Some stealth assessment frameworks utilize machine learning, including deep neural networks (Min et al., 2019) and Bayesian networks (Shute, 2011), which have been explored to analyze interaction sequences and infer learner competencies. These frameworks enable robust assessment of team development competencies, providing real-time insights into learner performance while maintaining engagement and authenticity.

In scenario-based training, stealth assessment holds significant potential to enable automated generation and real-time adaptation of scenarios based on individual trainees' competencies. By integrating GenAI with stealth assessment, highly personalized training experiences can be created through customized narratives and adaptive supports, ensuring trainees remain engaged and progress effectively. For example, interactive narrative generation frameworks such as NarrativeGenie could use stealth assessment to generate dynamic training scenarios that adapt to trainees' specific competencies. Stealth assessment can identify gaps in teamwork skills, prompting the framework to craft team-based scenarios that promote communication and information exchange. Non-player character (NPC) interactions, such as dialogues, can be automatically designed and seamlessly integrated into these scenarios to enhance learning and engagement, while collecting additional evidence to further assess, generate adaptive scenarios, and reinforce critical teamwork skills.

GenAI-based Training Scenario Generation

GenAI has significantly increased the potential for designing highly effective training scenario generation platforms based on GenAI-enabled narrative generation. Below, after discussing training scenario planning, we turn to offline scenario generation and online scenario generation.

Training Scenario Planning

GenAI has shown great promise for narrative generation tasks. Narratives, particularly interactive narratives, share many characteristics with training scenarios. Both involve generating a series of events involving a cast of characters, a set of locations, as well as a set of author-defined goals. Both computational tasks can be conceptualized as generating a set of preconditions and events that are logically coherent. As noted above, traditional approaches to narrative generation include narrative planners (Rivera et al., 2024), as well as other approaches that seek to customize aspects of a parent narrative using techniques such as reinforcement learning or evolutionary search (Folsom-Kovarik et al., 2019), and GenAI can augment these approaches in a variety of ways.

From a user interface perspective, GenAI systems show great potential for enabling designers using a natural language interface to specify and customize training scenarios, in contrast to the traditional but unwieldy approach of navigating multi-layered menus or using more restrictive scripting or plan-description languages. Additionally, GenAI systems have shown potential for rapidly creating or customizing sequences of events, as well as incorporating constraints such as locations or characters. Thus, they can support rapid prototyping relative to traditional approaches. It should be noted however, that the current generation of LLMs sometimes struggle when reasoning about world-state representations, and they may generate scenario plans that are not guaranteed to be valid or executable.

For example, consider the task of generating a Squad Battle Drill 2A scenario. The scenario can be modeled as three phases. In Phase 1, the squad moves and reacts to contact, Phase 2 centers on the Suppression and Maneuver, which is followed by Phase 3, Squad Assault and Consolidation. While traditional narrative planners can create different permutations of starting location and engagement location, GenAI systems can potentially allow authors to create and customize scenarios through a natural language interface rather than by modifying planner actions and objective functions. For example, an LLM-based system is more likely to be able to suggest ways to increase or decrease the difficulty of a generated scenario.

Offline Training Scenario Generation

In addition to generating high-level descriptions of training scenarios, GenAI systems have the ability to enable rapid generation and prototyping of offline training scenarios. Offline training scenario generation is the generation of training scenarios that can be embodied in a training or simulation environment, such as a 3D game engine, and do not adapt to trainee performance or actions during the exercise. This is desirable, as the generated scenario is largely deterministic, and the instructor can have confidence in the sequence of actions a trainee will experience when completing the given scenario. Additionally, the scenario will provide a largely repeatable experience, providing the opportunity for practice and improvement.

Converting a high-level scenario plan or description into a runnable training activity can be a difficult and laborious task. While the high-level description may specify a coherent narrative, to execute a training scenario requires extensive detail, including specifying valid locations of characters and objects at a coordinate level and controlling when events begin and complete. GenAI offers a promising approach to automatically producing runnable scenarios from high-level descriptions. Many simulation environments support configuration of scenarios through a set of configuration files or scripts, such as VBS Script for the Virtual Battlespace 3 system. The robust code generation capabilities of GenAI systems allow for the automatic generation of many of these files. This does require the LLM to be able to understand the syntax and capabilities of the configuration and scripting files, though many of these formats are well documented or use common scripting languages already incorporated into modern LLMs.

An additional requirement for training scenario generation is to generate the appropriate assessments for each phase of the scenario. This requires defining who and what are being assessed, when the assessment begins and ends, and what if any response should occur based on the assessment. The Generalized Intelligent Framework for Tutoring (GIFT) Domain Knowledge File (DKF) format provides a promising format for GenAI systems to define these characteristics in parallel with generation of the scenario configuration and orchestration files for the simulation environment.

Online Training Scenario Generation

Building on offline training scenarios, online training scenarios seek to adapt to trainee performance and actions in the training environment to dynamically tailor scenarios to trainees' ability level and training goals. These scenarios are in principle better able to be customized to the ability level of the trainees at runtime, as well as allowing for greater replayability of the generated scenarios. However, these scenarios can be difficult for instructors to interpret, as the space of potential adaptations might be difficult to communicate, and the scenario experience is less deterministic than offline scenarios.

For example, for the Squad Battle Drill 2A, an adaptive scenario may dynamically increase or decrease the accuracy of the enemy forces based on squad performance or potentially inject perturbations to increase difficulty such as an equipment malfunction or squad casualty. These adaptive scenarios require a computational agent to dynamically analyze trainee actions and perform adaptations in real time. GenAI

offers significant opportunities for authoring these scenarios, as well as driving the adaptive aspects of the scenarios, such as customized character dialogue, at runtime.

Recommendations for GIFT

Automated scenario generation is critical to the success and continued advancement of the military's synthetic training program. Future research, particularly for the GIFT and adaptive training community, should systematically investigate how Gen AI can be leveraged to better support training management and to create effective training scenarios at both the individual and unit levels. GenAI technologies, particularly LLMs, offer significant potential for shaping how instructors, teachers, and leaders plan and develop training scenarios. Using natural language prompts, these systems can create meaningful and realistic storylines, narrative events, and story arcs that align with desired training objectives. Research should explore how to best design instructor-facing interfaces to facilitate scenario creation, enabling users to describe scenarios intuitively while receiving immediate, actionable outputs. A critical consideration is how to present generated scenarios to instructors and end-users in a way that allows for rapid evaluation of their quality. Tools and workflows must enable instructors to assess scenarios without requiring them to run through an entire exercise in a 3D environment. Effective previewing methods, such as simplified visualizations, summaries, or simulations of key scenario elements, are essential for streamlining the evaluation process.

It will also be important to investigate how GenAI can be used to support stealth assessment. LLMs could align scenarios with specific team competencies, tasks, and conditions. This input could be used to populate assessment events and content that integrate seamlessly with existing adaptive training frameworks, such as GIFT's DKF. This approach ensures that generated scenarios not only meet training goals but also provide meaningful data for evaluating trainee performance and readiness. Evaluating techniques to best support this alignment and integration is an open area of research. Further, it will be important to ensure that the assessments and inferences inferred from stealth assessment events are sensitive to differences in skill levels and demonstrate validity (Katz et al., 2017).

Generating training scenarios that address the training needs of experienced trainees pose distinctive challenges for competency-based scenario generation and is an additional research need. Providing scenarios that are novel, accessible, and offer personalized difficulty progressions is a critical area of research. Researchers should investigate computational approaches, such as reinforcement learning (RL), that can be leveraged to optimize scenario outputs. RL could be used to evaluate and refine scenarios based on feedback from learners, instructors, or other stakeholders, ensuring that generated scenarios achieve desired learning outcomes effectively and efficiently. GIFT will also need to be expanded to support these functions.

Another important area of research is trust in GenAI-generated scenarios. Optimizing workflows for generating and reviewing scenarios is particularly important, given the time constraints faced by instructors. Tools must allow for quick scenario previews and efficient quality assessments without the need for extensive simulations. Transparent interfaces that allow users to inspect and adjust scenario elements will promote trust and confidence in the system's outputs. Finally, it will be important to demonstrate how these GenAI frameworks can generalize to support training in additional domains. Research should investigate how scenario generation processes can be tailored to new operational contexts, ensuring that the underlying methodologies remain flexible and scalable. By addressing these challenges and areas of research, Generative AI can transform the process of scenario generation, enhancing the effectiveness and adaptability of simulation-based training in the military.

Conclusions

As the capabilities of GenAI technologies continue to grow at an accelerating pace, they are paving the way towards extraordinarily powerful approaches to GenAI-based adaptive training scenario generation. It is quickly becoming feasible to create scenario generators that dynamically craft scenarios for a wide range of training domains and trainee populations where the scenario events (plots), NPCs (characters), and virtual worlds (settings) are tailored to the individual trainees' needs. By creating high-fidelity scenarios that exercise the specific competencies under the specific conditions specified by instructors and that utilize competency-aligned stealth assessment to measure skill levels, GenAI-based narrative generation will become the *de facto* standard for training delivery. Whether operating with offline or online scenario generation, because of their expressiveness and flexibility, they will achieve levels of effectiveness and efficiency in training that far surpass traditional methods. Further, they will provide the added significant cost-saving benefit of overcoming the authoring bottleneck for instructors and instructional designers. While GenAI methods in general and LLMs in particular pose computational challenges in the short term, GenAI will soon provide the foundation for game-changing adaptive scenario generation that will see rapid adoption for integrated training and assessment across the military.

Acknowledgements

The research described herein has been sponsored by the US Army DEVCOM, Soldier Center under cooperative agreement W912CG-23-2-0003. The statements and opinions expressed in this article do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

- Callaway, C. B., & Lester, J. C. (2002). Narrative prose generation. *Artificial Intelligence*, 139(2), 213-252.
- Folsom-Kovarik, J. T., Rowe, J., Brawner, K., & Lester, J. (2019). Toward automated scenario generation with GIFT. In A. M., Sinatra, A.C. Graesser, X. Hu, K., Brawner, & V. Rus (Eds.). *Design Recommendations for Intelligent Tutoring Systems: Volume 7 - Self-Improving Systems*. Orlando, FL: US Army CCDC. ISBN 978-0-9977257-7-3.
- Fowlkes, J. E., Lane, N. E., Salas, E., Franz, T., & Oser, R. (1994). Improving the measurement of team performance: The TARGETs methodology. *Military Psychology*, 6(1), 47-61.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.
- Katz, I. R., LaMar, M. M., Spain, R., Zapata-Rivera, J. D., Baird, J. A., & Greiff, S. (2017). Validity Issues and Concerns for Technology-based Performance Assessments. In R. Sottolare, A. Graesser, X. Hu, & G. Goodwin (Eds.) *Design Recommendations for Intelligent Tutoring System-Volume 5: Assessment Methods*, (pp. 209-224). Orlando FL: U.S. Army Research Laboratory.
- Kingma, D. P. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Kumaran, V., Carpenter, D., Rowe, J., Mott, B., & Lester, J. (2024). Procedural Level Generation in Educational Games from Natural Language Instruction. *IEEE Transactions on Games*.
- Kumaran, V., Rowe, J., & Lester, J. (2024, November). NARRATIVEGENIE: generating narrative beats and dynamic storytelling with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (Vol. 20, No. 1, pp. 76-86).
- Kumaran, V., Rowe, J., Mott, B., & Lester, J. (2023, October). Scenecraft: Automating interactive narrative scene generation in digital games with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (Vol. 19, No. 1, pp. 86-96).
- Lai, G., Leymarie, F. F., & Latham, W. (2022). On mixed-initiative content creation for video games. *IEEE Transactions on Games*, 14(4), 543-557.

- Lester, J., Bansal, M., Biswas, G., Hmelo-Silver, C., Roschelle, J., & Rowe, J. (2024). The AI Institute for Engaged Learning. *AI Magazine*, 45(1), 69-76.
- Martin, L. J., Harrison, B., & Riedl, M. O. (2016). Improvisational computational storytelling in open worlds. In *Interactive Storytelling: 9th International Conference on Interactive Digital Storytelling, ICIDS 2016, Los Angeles, CA, USA, November 15–18, 2016, Proceedings 9* (pp. 73-84). Springer International Publishing.
- Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., & Clark, J. (2024, April). *The AI Index 2024 annual report*. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University.
- Min, W., Frankosky, M. H., Mott, B. W., Rowe, J. P., Smith, A., Wiebe, E., ... & Lester, J. C. (2019). DeepStealth: Game-based learning stealth assessment with deep neural networks. *IEEE Transactions on Learning Technologies*, 13(2), 312-325.
- Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series*, 2003(1), i-29.
- Mott, B. W., Callaway, C. B., Zettlemoyer, L. S., Lee, S. Y., & Lester, J. C. (1999, November). Towards narrative-centered learning environments. In *Proceedings of the 1999 AAAI fall symposium on narrative intelligence* (pp. 78-82).
- Mott, B. W., & Lester, J. C. (2006, June). Narrative-centered tutorial planning for inquiry-based learning environments. In *International Conference on Intelligent Tutoring Systems* (pp. 675-684). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Oliver, E., & Mateas, M. (2021, October). Crosston tavern: modulating autonomous characters behaviour through player-NPC conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (Vol. 17, No. 1, pp. 179-186).
- Porteous, J., Cavazza, M., & Charles, F. (2010). Applying planning to interactive storytelling: Narrative control using state constraints. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 1(2), 1-21.
- Riedl, M. O., & Bulitko, V. (2013). Interactive narrative: An intelligent systems approach. *Ai Magazine*, 34(1), 67-67.
- Rivera, R. E. C., Jhala, A., Porteous, J., & Young, R. M. (2024, May). The Story So Far on Narrative Planning. In *Proceedings of the International Conference on Automated Planning and Scheduling* (Vol. 34, pp. 489-499).
- Rowe, J. P., Shores, L. R., Mott, B. W., & Lester, J. C. (2011). Integrating learning, problem solving, and engagement in narrative-centered learning environments. *International Journal of Artificial Intelligence in Education*, 21(1-2), 115-133.
- Sengar, S. S., Hasan, A. B., Kumar, S., & Carroll, F. (2024). Generative artificial intelligence: a systematic review and applications. *Multimedia Tools and Applications*, 1-40.
- Shaker, N., Togelius, J., Nelson, M. J., Liapis, A., Smith, G., & Shaker, N. (2016). Mixed-initiative content creation. *Procedural content generation in games*, 195-214.
- Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer games and instruction*, 55(2), 503-524.
- Wang, P., Rowe, J. P., Min, W., Mott, B. W., & Lester, J. C. (2017, August). Interactive Narrative Personalization with Deep Reinforcement Learning. In *IJCAI* (pp. 3852-3858).
- Yannakakis, G. N., Liapis, A., & Alexopoulos, C. (2014). Mixed-initiative co-creativity.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

CHAPTER 8 – GENERATIVE AI FOR ASSESSING AND SUPPORTING COMPLEX PERFORMANCES IN TEAMS

Peter W. Foltz¹, and Randall Spain²

University of Colorado, Boulder¹; US Army Combat Capabilities Development Command (DEVCOM) – Soldier Center²

Introduction

Artificial Intelligence (AI) has been evolving for decades, and AI-based technologies have been continually incorporated into training applications with greater and lesser success (e.g., Kulik & Fletcher, 2016). However, the field has taken a significant leap forward in the last few years with the general availability of Generative AI. This advancement is due to several factors (e.g., Lodzikowski et al., 2024). First, Generative AI models and tools are readily available and usable by a much wider swath of the lay population and intermediate technically skilled people. This allows faster innovation, experimentation, and adoption of Generative AI into new applications. Implementing AI no longer requires developers with advanced degrees but instead can be easily accessed via a web page or API (application programming interface) and generated through text-based prompt engineering. Whether the resulting AI methods and resulting applications can be appropriately evaluated by the lay developer without the formal training in education, data science or AI, is still an open question that will need to be addressed. Second, by training on terabytes of data, Generative AI can process human artifacts, including speech, writing, drawings, faces, and photos and assess the tenor and content of the inputs. It can then reason with the content and generate outputs that are often indistinguishable from human performance.

Since the 1960s, AI-based educational systems have been deployed for large-scale assessment, classroom and individual tutoring systems, and teacher support. Studies have shown that intelligent tutoring systems can raise student test scores 0.66 standard deviations over conventional classroom training and be as effective as expert tutors (e.g., D’Mello & Graesser 2023; Kulik & Fletcher 2016; VanLehn 2011). For example, automated essay scoring has been used operationally since the early 2000s to grade high-stakes exams as well as provide students with instant feedback (e.g., Yan et al. 2020). While successful, however, these approaches have had strong limitations. First, they have often been rule-based or trained on specific topics which can be inflexible and difficult to adapt to new situations. Second, they had limited natural language processing capabilities, resulting in more stilted, non-human-like language interactions. This has also limited systems to focusing less on higher-order thinking skills such as reasoning, argumentation, or collaboration. Third, while there has been research on multimodal processing, few applications incorporate modalities such as eye-gaze, gestures, facial reactions, or emotion detection, which can provide a deeper understanding of the learning context. Finally, automated training systems have been expensive to build, often requiring collecting hundreds or thousands of hours of student interaction data to train models for specific domains or have needed content experts to code knowledge and design the interactions.

Higher-order cognitive skills, including communicating, collaborating, problem solving, and showing critical thinking in realistic situations, have always been difficult to assess and provide effective feedback in training environments (e.g., Hilton & Pellegrino, 2012; National Academies of Sciences, 2018). In these cases, the effort put out by trainees can be considered *complex performances* in that they involve learners using higher-order cognitive skills, while interacting with other humans, computer-based agents, or complex technical systems, and generating multifaceted multimodal logs of information. Generative AI provides a means to address the current limitations in training systems as well as greatly change how we think about how AI can interact with learners and instructors. In this chapter, we focus primarily on those skills related to communicating and collaborating during complex performances. These skills are most

critical to teams who must work together to successfully achieve outcomes by combining their resources, knowledge and skills (e.g., Bonner et al., 2015; Graesser et al., 2018). We first outline some of the key advantages and potential that Generative AI or Large Language Models (LLMs) bring to improving assessment and feedback of collaboration. We then provide examples of approaches in related areas of applying Generative AI to support learning of collaborative skills. Finally, we conclude with some considerations for applying Generative AI for team assessment in real-world situations. It should be noted that the terms Generative AI and LLMs are used somewhat interchangeably in this chapter, where Generative AI is used as a broader term in that it encompasses both assessing and generating responses, while LLMs are used to describe the models used more specifically for assessment of performances.

The Role of Generative AI for Team Assessment

Generative AI offers several key advantages for assessing and training teams, particularly in complex, real-world scenarios. These advantages stem from its ability to process a variety of data, automate analyses, and provide personalized feedback, resulting in the potential for greater efficiencies in development and use, as well improved training outcomes. We outline key areas where it can enhance assessment capabilities, improve training environments, and create greater efficiencies and scaling in training.

Enhanced Assessment Capabilities

Automated assessment allows the scoring of team performances based on various metrics, reducing the need for extensive human evaluation and ensuring consistency in assessments. While automated assessment has been applied for team assessments (e.g., Cooke et al., 2012; Foltz & Martin, 2009; Sun et al., 2020), AI models have required a great deal of labor, prior data collection and validation, and are often built to work for very specific contexts. Generative AI can overcome some of the limitations of traditional AI, thereby enhancing the assessment capabilities. Previous AI-based systems were often rule-based or trained on specific topics, making them inflexible and difficult to adapt to new situations. In complex team contexts, Generative AI, with its LLMs trained on vast datasets, has broad domain knowledge giving it more flexibility to be able to handle a wider range of team contexts. This can help eliminate the need for extensive hand-crafted models or obtaining large amounts of training data specific to each training context.

A second limitation in traditional AI assessment has been that it has been hard to develop methods for assessing higher-order thinking skills that are expressed through language, such as communication and collaboration. While Natural Language Technology had continued to evolve, the models still tended to be brittle and requiring great effort and resources in training for new domains. Generative AI models have much more sophisticated means for processing spoken and written language and interpreting the expression, tenor and content of the communication. This allows rapid development of methods that can assess how well individuals and teams are communicating and the kind of discourse moves used.

Third, most AI analyses have been unimodal, in that they have analyzed one dimension of performance such as language or gestures. Generative AI can be multimodal and analyze and integrate a variety of data streams, including verbal and written communication, facial expressions, gestures, and log files of actions. This capability allows for a more comprehensive understanding of team dynamics and performance painting a more complete picture.

Improved Training Environments

Generative AI has the potential to improve the development of training environments and simulations. It can create more engaging and realistic learning experiences by enabling AI to take on human-like roles, such as mentors, tutors, or teammates that are able to respond flexibly to different situations. Using natural language processing to both generate and understand participant responses can create more realistic learning experiences. This allows for in-depth interactive conversations using speech and other modalities, with the AI adapting its responses to the learner's knowledge and language level.

With its dynamic content generation and feedback, Generative AI can create tailored learning materials, scenarios, simulations and assessment activities and feedback on the fly, adapting to the learner's needs. Generative AI can deliver tailored feedback to both individuals and the team, addressing specific strengths and weaknesses. This feedback can be generated based on specific rubrics that can be specified through prompt engineering, ensuring a more objective and standardized assessment. This ability can extend to giving feedback on complex skills touching on both team processes (e.g., lack of situation awareness) and team knowledge.

By analyzing performance data rapidly, Generative AI provides the promise of real-time monitoring and adaptivity. It can analyze team communications in near real-time, allowing for timely interventions and feedback. This is especially useful in dynamic environments where situations can change rapidly. Generative AI is also more flexible and adaptable to new situations compared to rule-based systems. AI can adjust the training environment based on the team's performance and needs, providing a more dynamic and effective learning experience. While much of team training has previously used post-hoc reviews or After Action Reviews (AARs), these new approaches enable training to move towards more of a model of continuous assessment embedded in team-based learning activities, providing real-time feedback loops.

Efficiency and Scalability

By automating many aspects of assessment and training, Generative AI can significantly reduce the costs associated with traditional methods, making it easier to scale up training programs. For example, prior approaches typically required human annotators to code large amounts of prior data for specific discourse moves or team behaviors (e.g., Foltz & Gilbert, 2021; Sun et al., 2020). The annotated data was then used to train supervised learning methods and often 100s of transcripts were needed to have sufficient generalizability (Foltz et al., 2006). Unlike previous AI systems, Generative AI does not require extensive manual coding of expert knowledge or rules. Instead, it learns from large datasets, but then can be “tuned” to the specific training context either through Retrieval Augmented Generation (RAG) or developing sufficient training information in the prompt. While still requiring some level of effort, it is still more adaptable and less labor-intensive to develop and opens the field to allowing more people without machine learning expertise to being able to develop effective training and feedback.

Examples of Applying Generative AI in Team Assessment

There are a number of applications where Generative AI has strong potential for team assessment. These areas include:

- 1) Teamwork Assessment where Generative AI can assess team collaboration by analyzing communication patterns, including spoken and written language, as well as nonverbal cues. This allows for the evaluation of team dynamics, individual contributions, and overall team performance. The AI can act as a collaborative partner, monitoring interactions and providing feedback to improve collaboration.
- 2) Critical Event Detection where Generative AI models can detect critical incidents in team performance by analyzing communication for indicators of problems such as loss of situation awareness or knowledge

gaps. This can help in understanding the factors contributing to success or failure and can be used to alert commanders or provide instant feedback to a team on ways to improve processes.

3) Automated AARs. Generative AI can support AARs by identifying key teaching points in audio and video data and displaying critical events from training missions, linked to relevant ratings and metrics.

4) Support for Human Trainers: AI can provide support to human trainers, providing them with AI-enhanced feedback that can help them work more effectively with teams.

Below we illustrate several specific applications of applying Generative AI and large language models in team contexts.

Supporting Collaboration in Small Teams

The first area focuses on the foundational question of *how to promote deep conceptual learning via rich socio-collaborative learning experiences for all students*. At the Institute for Student-AI Teaming (iSAT), we are addressing this question through AI partners that support knowledge-building communities where student and AI together engage in inquiry, critical thinking and collaborative problem solving in real world domains (D’Mello et al., 2024). Rather than replace instructors, these AI partners serve to augment student collaboration by facilitating sense-making, supporting students in sharing and building on each other’s ideas, and scaffolding brainstorming. They further support instructors in scaffolding of collaborative activities. The AI partners are co-designed with students and educators and are designed to be used in noisy, busy classroom environments with varied contexts. Built on frameworks of collaboration, they incorporate a range of technologies including Automated Speech Recognition, Natural Language Processing, gesture, posture, and eye-gaze recognition, as well as partner user-interfaces designed to provide feedback to students and instructors to augment the learning situations.

As an example of these AI partners, LLMs have been used to support small group collaboration through prompt engineering, a typical prompt will specify information including:

Role: *You are a friendly and helpful near-peer mentor helping middle school students to work collaboratively on a jigsaw activity...*

Setting: *For this particular jigsaw activity, each student has learned about a different sensor. <details on each sensor>...*

Restrictions & Format: *Output your response in JSON format with the following fields: conversation summary, timestamp, intervention, and feedback....*

Collaborative noticings: *Look for elaboration and uptake of group members’ ideas by the rest of the group, Look for when the group is establishing common knowledge, which is when someone shares their understanding of problems and solutions or when they establish common ground, Look for when the group is negotiating / coordinating, which is when someone responds to others’ questions and ideas or monitors their team’s execution of the task.*

Analyzing one minute of transcript from a team of middle school students who are discussing how their sensors work, the LLM generates information like this:

```
{
  "conversation_summary": "The group discussed their respective sensors, focusing on the environmental sensor's ability to measure CO2 and other gases, and the sound sensor's ability to detect sound levels. However, the students struggled with recalling specific details about the coding blocks for these sensors and the connections used.",
  "timestamp": "00:29:30.000",
  "intervention": "Prompt for elaboration and coordination",
  "feedback": "Let's each share one detail we remember about our sensor's coding blocks or functions to help piece together what we've learned."
}
```

Notice that based on moderately simple prompting the LLM is able to provide information about what kind of intervention should be given and how it can be worded, when it believes the intervention should be

provided, a summary of the state of the conversation and some justification for why to intervene. It should be noted though that these interventions still need to be evaluated by humans in terms of their accuracy, appropriateness and timing.

Supporting Trainers in Tutoring Contexts

A second area focuses on addressing the issue, *while human tutoring has often proven effective; however, scaling human tutoring while maintaining quality of service is challenging and expensive*. In this approach, we have implemented a hybrid human-AI tutoring (HAT) platform to support critical components of tutoring work, grounded in a research-based framework with evidence for supporting low-income students. Through monitoring tutor-student interactions, the HAT platform combines multiple streams of information – student(s) and tutor speech and language, nonverbal signals, and shared math tasks/context – with intelligent analytic tools to develop more complete, nuanced, and personalized supports for critical learning interactions along cognitive, relational, and affective dimensions. As such it supports tutors in both centering rich, collaborative mathematical discussions and problem solving and building caring relationships.

As an example, the system incorporates both LLMs and supervised ML (machine learning) approaches to assess the transcripts of student/tutor conversations and provide feedback to coaches who support the tutors. To simplify the task of coaches reviewing an hour of a video of tutors and students interacting, the system can note specific sections of the video for review as well as provide an overview of the whole session. Figure 1 below shows how LLMs are used to generate a summary of the math topics discussed and an overview of the session. Then supervised ML approaches combined with rule-based methods provide information about the amount and types of talk that was generated during a session.

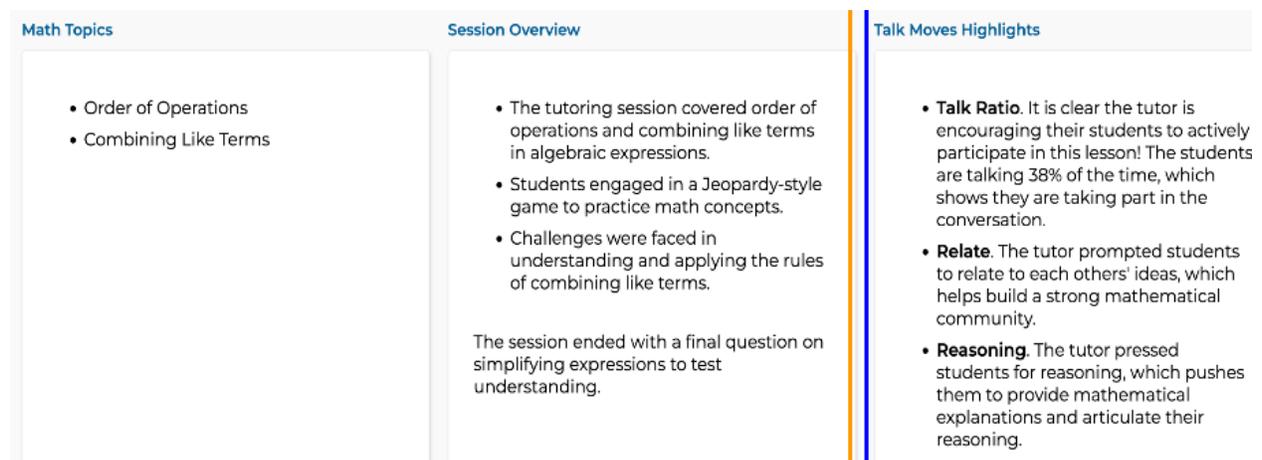


Figure 1. Sample Session Feedback in Hybrid Human-AI Tutor (HAT).

Supporting Competency-Driven Scenario Development

In addition to supporting collaboration in small teams and instructors in tutoring contexts, Generative AI is transforming how synthetic training scenarios are developed to support team collaboration training. Synthetic training environments, which simulate real-world tasks, equipment, and conditions, have been fundamental to the US military’s training approach. These environments provide realistic, safe practice for complex tasks, allowing trainees to rehearse decision-making, critical thinking, and tactical skills within a

structured setting. Historically, creating simulation-based training scenarios has been resource intensive, requiring specialized knowledge of software, tools, and subject matter expertise to ensure the scenario aligns with targeted training objectives. During a training exercise, a simulation operator may manually provide scenario injects —critical events that are designed to challenge trainees — but there is infrequent integration of assessments for coaching. The growing need for more personalized and dynamic training underscores the need for streamlined scenario development. Generative AI offers a solution to this need by automating the creation of a variety of content, including text, images, immersive environments, and narrative and storylines to facilitate engaging and effective team training scenarios. It can generate personalized scenarios tailored to individual and unit needs, enabling rapid development of simulations that challenge trainees to meet performance objectives, hone essential teamwork and collaboration skills, and navigate complex situations.

Supporting Team Assessment through Dialogue Analysis

Generative AI can also support team assessment by analyzing communication patterns, including spoken and written language, as well as nonverbal cues. Communication among team members is essential in task-critical environments. Successful teamwork requires team members to pass and share information in a timely manner, monitor and support one another, demonstrate leadership, and coordinate actions. Generative AI can analyze team dialogue to identify team communication patterns, assess how team members responded during periods of high workload or during critical events and then use this information to help train teams more effectively on managing communication, skills and team dynamics.

Discussion and Recommendations for Future Research

The above examples illustrate that Generative AI shows great potential for team assessment, training and support. However, research is still needed to assess the quality and effectiveness of the feedback generated by AI and its ability to detect team performance in ways that are accurate and generalizable. Further, we need to ensure the generated feedback is useful, effective, and trustworthy. Thus, there is a need for the field to establish standards that espouse transparency and explainability around the methods for developing, testing, and validating AI models and their intended use and limitations in their training contexts. This will require a multi-faceted approach that involves researchers, educators, policymakers, and the public. A number of steps will be needed to be performed to assure that trust is engendered and that the approaches are used ethically. These steps include addressing:

Transparency: It is critical to be aware of potential concerns in AI models and to ensure that they are trained on a variety of datasets. The design and limitations of AI systems should be transparent (Diaz-Asper et al., 2024; Mitchell et al., 2019). AI models are trained on data, and any inaccuracies, omissions, or concerns within that data can be reproduced and magnified in the AI's behavior. If algorithms are not trained on a variety of sets of trainee responses, they may disproportionately penalize or reward certain ways of communicating, thinking, or problem-solving. This can have different kinds of effects on trainees from various backgrounds. Ensuring the accuracy, variety, and breadth of training data is a significant challenge. Most model providers are not transparent regarding the design of their systems and do not provide guarantees against these concerns.

Explainability: AI systems often function as a "black box," making it difficult to understand precisely why a model makes a certain decision or prediction. This is particularly problematic when considering the psychometric properties of an assessment, especially validity and reliability. Traditional methods of evaluating those characteristics may not directly apply to Generative AI systems. New methods will likely have to be developed that consider changes in assessments and the nature of the learning environments.

Potential for Hallucinations: LLMs have the potential to "hallucinate," or generate information that seems plausible but is incorrect or misleading. This can be especially harmful in training contexts, where accuracy of information is paramount. Addressing this requires not only better models but also robust systems for verifying and validating AI outputs.

Need for Human Oversight: Human instructors still play a critical role in orchestrating how AI is used in training environments. They guide the learning process and ensure that AI is used to empower learners and instructors.

Recommendations for GIFT and STEEL-R Overall

We provide some recommendations on the potential of Generative AI for supporting and enhancing teams.

Use Generative AI to Assess Team Skills and Task Performance: An ITTS (Intelligent Team Tutoring System) should be able to assess both team skills (qualities that transfer across tasks) and task skills (how well the team performs on a specific task) (e.g., Bonner et al., 2015; Foltz & Gilbert, 2023). Generative AI can help identify and measure aspects of these skills in team interactions. However, we still need to ensure that we develop metrics aligned to desired outcomes (e.g., Foltz et al., 2023). When using Generative AI to assess team performance, it is important to align the metrics used to frameworks of performance, cognition and mental states that are relevant to the desired outcomes.

Develop Generalizable Systems: Aim for a system that can be applied across various domains and tasks. This involves creating reusable modules for teaching team skills that can be adapted to different training scenarios. This requires a common infrastructure that is agnostic to whether the team members are human or software agents.

Incorporate Feedback Generation: Generative AI can create relevant and compelling feedback based on data points extracted from team performance, something that was difficult in earlier approaches. The system can provide both team-level and individual feedback, though it is important to consider how much feedback to provide.

Focus on Continuous Assessment: Move towards a model of continuous assessment by embedding assessments within the learning activities and providing real-time feedback. This helps teams learn and improve dynamically. Generative AI enables near real-time analysis of team communications, allowing for timely feedback and interventions. This is critical for teams working in dynamic and high-stakes environments where immediate adjustments are necessary. By implementing real-time monitoring and feedback, systems can be designed to detect critical events, communication breakdowns, and deviations from standard operating procedures and correct them on the fly or alert commanders.

Further Leverage Communication Analysis: Analysis of team communication has been a successful approach to characterizing a range of aspects of team process and performance (e.g., Cooke et al., 2012). Generative AI allows us to apply natural language processing (NLP) techniques in novel ways to analyze team communication content, patterns, and flow. This can reveal information about team cognition, knowledge sharing, situation awareness, and leadership. However, communication is only one part of the information stream. Thus, we must also focus on Multimodal Data Integration. Generative AI can analyze various data streams, such as speech, typed text, log files of actions, and even non-verbal cues like gestures and eye-gaze. This is especially important in complex environments where team performance is not solely based on verbal communication. Combining these modalities can provide a more comprehensive understanding of team dynamics and performance.

Address the Challenge of Ambiguity: Team training often involves handling data in ambiguous situations where not all information is known, and multiple types of information must be fused to paint a more complete picture of team performance. Generative AI systems can be incorporated to handle challenges like simultaneous speaking, non-verbal cues, and the context of communication and are not as brittle to dealing with ambiguity as prior methods.

Develop Hybrid Human-AI Training Systems: Combine the strengths of AI with human expertise in training. AI can automate data analysis and feedback generation, while human trainers can provide personalized coaching, guidance and interpretation. AI can provide feedback to trainers and support their work with tutors, focusing on cognitive, relational, and affective dimensions. This can improve the quality and scalability of training. A critical approach in developing these hybrid systems is to use a human-in-the-loop approach having a broad swath of human expertise involved in the development, validation, and deployment of the AI system while also ensuring that they remain involved in supporting the training as the systems are used.

Ensure Explainability and Transparency: It is critical to address the black-box nature of AI by providing explainable and transparent outputs. This will require new evaluation methods that consider the psychometric properties of AI-based assessments and align AI assessments to frameworks of performance, cognition and mental states.

Address Ethical Concerns: Be aware of potential concerns in AI models and ensure that they are trained on a variety of datasets (e.g., Suresh & Gutttag, 2021). The design and limitations of AI systems should be transparent. Also address concerns about academic integrity and the potential for AI to produce hallucinations.

Recognize the Limitations of AI: Just because we can measure a team skill, it does not mean that we know when or how to intervene to remediate. Generative AI does not eliminate the need for effective competency frameworks and mapping assessment to performance outcomes. We further need to avoid over reliance on AI in the absence of understanding what it does. While AI can be a powerful tool for learning, it is important to ensure that trainees and instructors do not become over-reliant on it. AI should be used as a tool to augment human learning, not replace it.

Conclusions

Generative AI shows significant promise for enhancing team training and performance in complex environments through several key capabilities. It can provide continuous, real-time assessment of both team and individual performance by analyzing multiple data streams including communication patterns, non-verbal cues, and task execution. The technology's ability to handle ambiguous situations and integrate multimodal data makes it particularly valuable for realistic, dynamic training scenarios. However, successful implementation requires careful attention to several factors: establishing a sound framework of the skills that need to be assessed, maintaining human expertise in the loop, ensuring transparency and explainability of AI systems, and avoiding over-reliance on AI systems. When properly implemented as part of a hybrid human-AI approach, Generative AI can enable more scalable, personalized, and effective team training while providing timely feedback that helps teams adapt and improve their performance in real-world situations.

Acknowledgements

This work has benefited from support by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805 and funding through the Learning Engineering Virtual Institute. The opinions expressed are those of the authors and do not represent views of the funders.

References

- Bonner, D., Gilbert, S., Dorneich, M. C., Burke, S., Walton, J., Ray, C., & Winer, E. (2015). Taxonomy of Teams, Team Tasks, and Tutors. In *Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym2)* (pp. 189-198).
- Cooke, N. J., Duchon, A., Gorman, J. C., Keyton, J., Miller, A. (2012). Preface to the special section on methods for the analysis of communication. *Human Factors*, 54 (4):485-488. doi:10.1177/0018720812448673
- Diaz-Asper, C., Hauglid, M. K., Chandler, C., Cohen, A. S., Foltz, P. W., & Elvevåg, B. (2024). A framework for language technologies in behavioral research and clinical applications: Ethical challenges, implications, and solutions. *American Psychologist*, 79(1), 79.
- D'Mello, S. K., Bidy, Q., Breideband, T., Bush, J., Chang, M., Cortez, A., ... & Whitehill, J. (2024). From learning optimization to learner flourishing: Reimagining AI in Education at the Institute for Student-AI Teaming (iSAT). *AI Magazine*, 45(1), 61-68.
- D'Mello, S. K., & Graesser, A. C. (2023). Intelligent tutoring systems: How computers achieve learning gains that rival human tutors. In P. A. Schutz & K. R. Muis (Eds.), *Handbook of educational psychology* (4th ed.). Routledge. <https://doi.org/10.4324/9780429433726>.
- Foltz, P. W., Chandler, C., Diaz-Asper, C., Cohen, A. S., Rodriguez, Z., Holmlund, T. B., & Elvevåg, B. (2023). Reflections on the nature of measurement in language-based automated assessments of patients' mental state and cognitive function. *Schizophrenia research*, 259, 127-139.
- Foltz, P. W., & Gilbert, S. B. (2023). Team Tutoring in Intelligent Tutoring Systems SWOT ANALYSIS. *Design Recommendations for Intelligent Tutoring Systems: Volume 10-Strengths, Weaknesses, Opportunities and Threats (SWOT) Analysis of Intelligent Tutoring Systems*, 91.
- Foltz, P. W., & Martin, M. J. (2009). Automated communication analysis of teams. In E. Salas, G. F. Goodwin, & C. S. Burke (Eds.), *Team effectiveness in complex organizations: Cross-disciplinary perspectives and approaches* (pp. 411-431). Routledge/Taylor & Francis Group.
- Foltz, P. W., Martin, M. J., Abdelali, A., Rosenstein, M. B., & Oberbreckling, R. J. (2006, July). Automated team discourse modeling: Test of performance and generalization. In *Proceedings of the 28th Annual Cognitive Science Conference* (pp. 1317-1322).
- Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., & Hesse, F. W. (2018). Advancing the science of collaborative problem solving. *Psychological Science in the Public Interest*, 19(2), 59-92.
- Hilton, M. L., & Pellegrino, J. W. (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. National Academies Press.
- Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, 86(1), 42-78. <https://doi.org/10.3102/0034654315581420>.
- Lodzickowski, K., Foltz, P. W. & Behrens, J. T. (2024). Generative AI and its educational implications. In T. Kourkoulou, D. O. Tzirides, B. Cope, and M. Kalantzis, (eds). *Trust and Inclusion in AI-Mediated Education: Where Human Learning Meets Learning Machines*, Springer.
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229). <https://doi.org/10.1145/3287560.3287596>
- National Academies of Sciences, Division of Behavioral, Social Sciences, Board on Science Education, Board on Behavioral, Sensory Sciences, ... & Practice of Learning. (2018). *How people learn II: Learners, contexts, and cultures*. National Academies Press.
- Sun, C., Shute, V. J., Stewart, A., Yonehiro, J., Duran, N., & D'Mello, S. (2020). Towards a generalized competency model of collaborative problem solving. *Computers & Education*, 143, 103672.

- Suresh, H., & Guttag, J. (2021, October). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1-9).
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197–221. <https://doi.org/10.1080/00461520.2011.611369>.
- Yan, D., Rupp, A. A., & Foltz, P. W. (Eds.). (2020). *Handbook of automated scoring: Theory into practice*. CRC Press. <https://doi.org/10.1201/9781351264808-19>.

CHAPTER 9 – GENERATIVE AI FOR CONCEPT LEARNING: EVOLUTION OF THE MOBILE FACT AND CONCEPT TRAINING SYSTEM

Andrew M. Olney
The University of Memphis

Introduction

Intelligent tutoring systems, like all adaptive training systems, require three essential elements to deliver personalized instruction: a model of student performance, training material appropriate to that level of performance, and a method that can select the best training material at any moment. Models of student performance and the selection of optimal training material are interesting and challenging research problems, but they are well-bounded from a practical standpoint. For example, consider an algebra tutoring system for word problems (Singley et al., 1989). The student model can be constructed using ACT-R (Adaptive Control of Thought – Rational), and an analysis of knowledge components and knowledge tracing can be used to select the next optimal problem (Corbett & Anderson, 1994). However, the word problems themselves, i.e. the text which situates the problem in a real-world context, is relatively unbounded, simply because there are so many possible real-world contexts. This problem of creating many different permutations of training content must be overcome in order to fully adapt to learner preferences, interests, and prior knowledge (Sinatra, 2015): when a learner is given an algebra word problem on a familiar topic, it can increase their motivation to learn and activate prior knowledge that enhances learning performance.

For many years, the research community has focused on training material authoring tools, some of which include a high level of automation (Olney et al., 2015; Olney et al., 2021). Traditional authoring tools that generated text typically used a template-based system that required progressively deeper symbolic natural language processing (Gatt & Reiter, 2009; Gatt & Krahmer, 2018) including lexicons, knowledge bases, and other resources that in practice are extremely difficult and time consuming to construct. For example, the Cyc project was one of the longest-running knowledge base construction projects but has been widely criticized for not being usable or useful (Conesa et al., 2010; Domingos, 2015). In this earlier period of artificial intelligence (AI) research, knowledge base construction was a key step for curating the knowledge used to automatically generate training content.

Transformer-based large language models (LLMs; Vaswani et al., 2017) have dramatically changed the landscape of natural language generation (NLG) and thus the generation of training content. In contrast to traditional NLG approaches that separate knowledge from generation, LLMs typically consist of a single stage that generates text using knowledge implicitly represented in the weights of the network. This is particularly impressive given the common training objective of LLMs is to predict the next token (i.e. word) in text rather than to acquire knowledge per se. Even relatively early LLMs have been shown to be competitive with or outperform traditional knowledge bases for querying factual knowledge (Petroni et al., 2019), and the ability of LLMs to recall factual knowledge has been shown to increase with size (Roberts et al., 2020). The rapid progress in LLMs meant that by 2020, it was possible to use LLMs to create content for adaptive training systems.

This chapter reflects on the rapid advances of LLMs during the development of an adaptive training system and how these advances changed our development plans and extended the system beyond our original plans. The Mobile Fact and Concept Training System (MoFaCTS; Pavlik et al., 2020; Pavlik & Eglington, 2021),

is an adaptive training system designed for learning simple content (i.e. facts), that can be understood as a flashcard system with an enhanced student model. For a given learning task, MoFaCTS presents a set of items (flashcards) to the learner, and the learner must enter (type or speak) the correct answer for that item. If the learner is correct, MoFaCTS proceeds to the next item. If the learner is incorrect, MoFaCTS notifies the learner that the answer is incorrect and presents the correct answer for a set period of time (i.e. a forced study period). To select the next item, MoFaCTS uses a generalized model of the difficulty of each item but updates that model based on a learner's history of practice together with a cognitive model of forgetting. MoFaCTS estimates the probability that a learner will answer an item correctly based on the number of previous successful and unsuccessful trials and the time elapsed since the last trial. MoFaCTS then selects the item that is close to the threshold of being forgotten, based on an efficiency principle: practice with almost forgotten items takes less time than practice with forgotten items, because forgotten items trigger a forced study period. The focus of our project was to automatically generate items for MoFaCTS from a textbook for Anatomy and Physiology (A&P), a key course for nursing and allied health professions that requires a great deal of memorization.

Our original research plan was to use traditional statistical and symbolic methods to extract cloze items from the text, generate refutational feedback in response to incorrect answers, and to create paraphrases. To accomplish these tasks, we planned to leverage previous work on cloze item generation from textbooks (Olney et al., 2017), concept map extraction from text followed by NLG (Olney et al., 2010, 2012; Olney, 2018), and machine-translation techniques for paraphrase (Napoles et al., 2016). Ultimately, we only used our previous work on cloze item generation and accomplished everything else with LLMs – because LLMs allowed us to accomplish the same objectives in less time. Additionally, we were able to expand our project scope to include multiple-choice question generation, definition generation, and syllabification. The remainder of this chapter describes in detail each of these decisions, the resulting systems, and makes recommendations for future adaptive systems, including the Generalized Intelligent Framework for Tutoring (GIFT).

The Role of Generative AI

Elaborated feedback

The purpose of elaborated feedback is to remediate a student's incorrect answer in a more useful way than just providing the correct answer. Consider the following example from our data, where underlining indicates the blank in the cloze item:

Source: The cytoplasm at the distal end of the axon is rich in mitochondria and contains many tiny vesicles that store neurotransmitter.

Student: acetylcholine

In this example, the student misunderstands cytoplasm because they made an error. However, it is likely that they also misunderstand acetylcholine and perhaps even the relationship between cytoplasm and acetylcholine. Therefore, our elaborated feedback should address the target concept, the incorrect answer concept, and their relationship (if it exists).

Our initial approach to this problem was based on our previous work on extracting concept maps from text (Olney et al., 2010, 2012; Olney, 2018). In that approach, the relationship between two concepts can be uncovered by finding the shortest path between the concepts in the map. Let us look at a simpler example than the cytoplasm/acetylcholine one above. Suppose that we extracted the following triples from the textbook:

sugar – has_type – monosaccharide

monosaccharide – has_type – fructose

sugar – has_type – disaccharide

disaccharide – has_type – sucrose

disaccharide – has_property – two joined monosaccharides

Suppose a student has confused fructose and sucrose and so answered a cloze item with sucrose when the correct answer is fructose. Using the above triples, an NLG system can create a paragraph explanation, e.g. “Sucrose and fructose are both sugars, but sucrose is a disaccharide and fructose is a monosaccharide. A disaccharide is characterized by two joined monosaccharides.” For this kind of approach to work, the triples need to be extracted from the textbook without error, the dialogue planner that selects relevant triples needs to do so without error, and the triples need to be rendered to text using NLG without error.

One of the major obstacles we faced in applying our previous approach was that it relied on a variety of components that were old, idiosyncratic, and difficult to port to the web. For example, the Lunds Tekniska Högskola (LTH) semantic parser (Johansson & Nugues, 2008), a key component in our earlier work, implements a different theory of syntax than nearly all parsers from the last decade, which mostly use Stanford dependencies. Simply stated, Stanford dependencies are surface-syntax oriented, whereas the LTH parser is deep-syntax (meaning) oriented. Further, the LTH parser provides word sense disambiguation of predicates to specific PropBank frames, i.e. disambiguates word senses to specific patterns of semantic argument structures. Virtually no semantic parsers in the last decade attempt this disambiguation, since this is not a key component of shared task evaluation. Thus, modern parsers were not delivering the information needed for our knowledge extraction system to operate.

To implement our previous approach, we were progressively investing more time in “reinventing the wheel” rather than making progress on elaborated feedback in MoFaCTS. Thus, we decided to investigate recent advances in deep learning that focus on question answering. To leverage that work, we reformulated the problem of providing feedback as the problem of answering a student’s implicit question, i.e., the question they should ask once they realize their error. In the first example, this might be, “What is the relationship between acetylcholine and cytoplasm?” To explore this idea, we developed a deep learning question answering system based on recent work in long-form question answering, which returns a paragraph explanation to a question rather than a single word or phrase answer. Importantly, we leveraged a model developed using the “Explain Like I’m Five” (ELI5) forum on Reddit (www.reddit.com/r/explainlikeimfive/). Because this forum attempts to answer questions in the simplest terms possible, it provides a compelling dataset for building models with the same properties. Here is the answer to the relationship question in our running example, from the current version of our system, where the first two sentences are MoFaCTS standard feedback, followed by the underlined elaborated feedback:

Acetylcholine is not right. The correct answer is cytoplasm. Acetylcholine is synthesized in the cytoplasm of nerve terminals by the enzyme choline acetyltransferase and is then transported into synaptic vesicles.

In our best performing system (Olney, 2021), we used an early retrieval-augmented generation approach where we provided definitions for the key terms (cytoplasm and acetylcholine) and passages from the A&P textbook used in the course to the question answering model. Definitions were obtained from the textbook glossary, or if missing, from Wikipedia by first applying a wikifier (Cheng & Roth, 2013) to the correctly filled-in cloze item sentence in order to get Wikipedia page ids for the key terms in the sentence. These page ids were then used to query the corresponding Wikipedia pages for their first paragraph of text, which

was used as a proxy for a definition. Additional supporting documents were retrieved using an Elasticsearch (<https://www.elastic.co/elasticsearch/>) index of the A&P textbook with the synthetic question as the query.

In contrast to the previously proposed concept map approach, in the neural question answering approach we never parsed the textbook or created a knowledge representation for further processing. We did not even retrain the question answering model and simply used the stock model available from HuggingFace for ELI5 (<https://yjernite.github.io/lfq.html>). Our evaluation of the system's performance used nurses and similar experts and screened out ineffective/unreliable raters using intra- and inter-rater reliability metrics (see Olney, 2021 for details). The results suggest that both the fluency and correctness of answers are quite high; comparing to our other evaluations which use these metrics, at this time we speculate that the system performance is close to human in both fluency and correctness, though in our evaluation we did not directly compare to human elaborated feedback. Similarly, we did not compare the system performance to our planned symbolic approach, but compared to our earlier findings (Olney et al., 2012), a casual observer can see that performance is dramatically better. In short, the performance and development speed benefits of LLMs (relative to traditional symbolic approaches) were quite striking based on our previous experience.

Reformulating elaborated feedback as an answer to a question the student did not ask (but should have), brings elaborated feedback into parity with tutorial dialogue in the AutoTutor framework (Nye et al., 2014). Simply stated, all AutoTutor sessions for learning content (rather than adult literacy) are based on a seed question and ideal answer. The rest of the dialogue is then derived from the ideal answer with the goal of eliciting that ideal answer from the student. Therefore, all hints, prompts, and related dialogue are simply transformations of sentences that comprise the ideal answer into tutor questions and assertions. Because of the symmetry between ideal answers and elaborated feedback, we can use the elaborated feedback paragraphs generated by our long form question answering model as the basis for a tutorial dialogue session that remediates the student's incorrect answer. We additionally implemented this tutorial dialogue in MoFaCTS using traditional symbolic methods but with elaborated feedback as the basis of the dialogue content.

Paraphrases

The purpose of paraphrasing is to force the student to deeply encode the knowledge in the item rather than superficial features of the item. Anyone who has used paper flashcards has likely experienced the tendency towards superficial encoding, for example remembering the answer based on a stain or bend on the card rather than a knowledge-based association. This same kind of processing happens with computer-based items where participants tend to read the first few words of the item and remember the answer rather than remember the answer based on knowledge by reading the entire item and reasoning to the answer.

Paraphrasing changes the words used in the item to remove superficial cues while preserving the meaning of the text. Paraphrasing makes relying on superficial cues less effective and, in theory, promotes deeper processing. Cloze items based on paraphrases of text from the A&P textbook should thus have these beneficial properties.

Our proposed approach to paraphrasing was to use statistical methods for paraphrasing described in Napoles et al. (2016), which frames paraphrasing as a machine translation problem. Essentially any transformation of text that approximately preserves meaning can be framed as a machine translation problem with a specific objective, e.g. both shortening text and changing the reading level of text involve changing the surface form of text with an additional objective. The method of Napoles et al. (2016) is based on first obtaining snippets of English that have been translated to other languages and then back to English (with an associated change in surface form) and then using statistical machine translation techniques to search through the possible space of snippets and assemble those that best minimize an error function for the paraphrase task. The

classic problem with systems like this is that there can be errors in the translation data, errors associated with the snippet boundaries, and errors associated with the search process.

Building on recent advances in LLMs, we instead developed a paraphrase method based on backtranslation of the A&P textbook (Olney, 2021). Backtranslation is the process of translating from English into another language and then back into English. The “other” language is called a pivot language, and more than one pivot language can be used in a backtranslation process, e.g. English – Czech – Russian – English. Leveraging the wide variety of pivot languages that have been explored by other researchers, we evaluated Czech, Russian, Chinese, Persian, Arabic, Hindi, Turkish, and Welsh as pivot languages and found that Czech and Russian introduced the most lexical/syntactic variety while introducing the fewest meaning errors. We further evaluated combining Czech and Russian into a double pivot, as in the previous example, and found this increased lexical/syntactic variety over each pivot individually without noticeably increasing meaning errors. We then backtranslated the textbook using the Google Translate API (Application Programming Interface), which is a high-quality neural machine translation, and used that backtranslation data to train a text-to-text paraphrase model using the T5 (Text-to-Text Transfer Transformer) LLM (Raffel et al., 2020). Our training data consisted of sentence pairs (original and backtranslated) such that T5 was trained to produce the backtranslated sentence when the original sentence was input, and vice versa. An example source sentence and system paraphrase are given below, with the changed text underlined:

Source: Another layer of connective tissue, called the perimysium, extends inward from the epimysium and separates the muscle tissue into small sections.

Paraphrase: Another layer of connective tissue, called the perimysium, extends inward from the epimysium, dividing the muscle tissue into small portions.

An expert evaluation, again with nurses and doctors as participants, found high fluency for the paraphrases, with approximately 40% of paraphrases being judged as more fluent than the original text. Meaning was also rated highly, with 70% of all meaning ratings above 75 on a 100-point scale. We did not directly compare human generated paraphrases in the evaluation, but based on later work with similar evaluations, we suspect that the performance of the system is close to human in most cases for both fluency and meaning. Similarly we did not directly compare the performance of the T5-based system to the system of Naples et al. (2016) for A&P paraphrases, so we can only speculate as to the difference in paraphrase quality by pointing to the general finding that joint inference reduces error more than a pipeline approach (cf. Finkel et al., 2006) and that a text-to-text neural approach embodies joint inference in this way, even down to the features themselves. Creating the training data and training T5 to perform this paraphrase task took less than a week of effort, which created time for additional research outside the scope of our grant proposal.

Multiple Choice Questions

Multiple choice questions (MCQs) can be used both for learning and for assessing learning. As items presented in MoFaCTS, they could reduce the difficulty of a cloze item by giving a hint, since the learner only needs to recognize the correct answer among the options. In the same way, MCQs could reinforce differentiation between confusable concepts listed among the answer options. When used for learning, MCQs could be paired with immediate feedback, like the elaborated feedback previously discussed.

We explored MCQ generation with LLMs in two papers; in what follows we describe the initial system development (Olney, 2022) and most recent human evaluation (Olney, 2023). Similar to our approach to elaborated feedback, our approach to MCQ with LLMs was based in question answering, specifically MCQ answering, inspired by a system called Macaw (Tafford & Clark, 2021). We used Macaw, which is based on the T5 model, for question generation, as it leverages multiple questions-answering datasets by casting them into a common format based on slots, including A (answer), Q (question), M (multiple choice options),

and C (context), and input/output specifications called angles. For example, a simple dataset that pairs answers with questions could support the angles $Q \rightarrow A$ (for question answering) as well as $A \rightarrow Q$ (for question generation). Datasets with more slots can support more angle combinations. By putting datasets into this common format, it was possible to train Macaw on a variety of question-oriented tasks, including question generation. However, the original evaluation focused on question answering, so it was unclear how effective Macaw would be for MCQ generation.

To use Macaw successfully, we first evaluated what angles yielded the best questions. We evaluated three angles with increasing specificity, $C \rightarrow QMA$ (which generates question, options, and answer from text alone), $AC \rightarrow QM$ (which generates question and options from text and answer), and $QAC \rightarrow M$ (which generates only the options given the other inputs). Our expectation was that providing the most information, which simplifies the task, would give the best results, but paradoxically it was the intermediate angle $AC \rightarrow QM$ that yielded the best results: it gave four distinct options 83% of the time compared to $QAC \rightarrow M$ which yielded four distinct options 67% of the time and $C \rightarrow QMA$, which never yielded four distinct options. In order to increase the effectiveness of $AC \rightarrow QM$, we leveraged the inherent instability of LLMs where they provide quite different outputs based on minor changes to the input by using the paraphrase system previously discussed to generate new angle inputs whenever Macaw failed to generate four distinct outputs. We repeated paraphrase generation up to 10 times (an arbitrary limit that could have been extended) and increased the success of $AC \rightarrow QM$ to 97.5% of items.

We conducted a human evaluation study using nurses and doctors as participants with questions from three sources: an A&P textbook from OpenStax (Betts et al., 2017; Textbook), our application of Macaw (paraphrase-augmented $AC \rightarrow QM$; hereafter Macaw+), and BingChat, a GPT-4 service that was later rebranded at Copilot. All MCQs from OpenStax were rewritten as sentences with answers, and these sentences/answers were used as input to both Macaw+ and BingChat. For BingChat, they were included in the following prompt:

Write a multiple choice question using the following sentence and answer. Convert the sentence into a question that matches the answer. Use JSON format.

Sentence: <sentence>

Answer: <answer>

Note the above approach closely matches the content of questions across the conditions; in theory each condition will have a semantically identical question and the exact same answer and vary most in their answer options. Participants were asked to rate questions on seven different scales, including question (correctness and fluency), correct answer (truly correct and present in options), answer options (number of options that are correct and number of options that are distinct) and combined quality. Results from the analysis indicated that there were only significant differences between the conditions for the ‘answer present in options’ and the combined quality scores. There were no significant pairwise differences for the former, but combined quality post hoc analysis indicated that the Textbook condition had greater combined quality than Macaw+ and BingChat, which were not different from each other. Follow up error analyses suggested that the primary failure mode of Macaw+ was to fail to generate four distinct answer options, especially when the options were in some kind of list. In contrast, the primary failure mode of BingChat was to fail to generate a question with a given correct answer. Instead, it would sometimes generate a question whose correct answer was not the given answer. Additional analyses showed that when questions with incorrect distinct options are eliminated, Macaw+ was not significantly different from Textbook on combined quality, but even using the same approach with ‘answer present in options,’ BingChat was still significantly worse than Textbook on combined quality. Altogether this suggests the best solution is to take the simple approach of checking for distinct options with Macaw+ and discarding questions without four

distinct options, both of which are technically simple to accomplish. With this simple check, Macaw+ questions are likely to be indistinguishable from human-authored questions in the A&P domain, and we suspect in all domains, since Macaw was not specifically trained on A&P.

Morphology generation

Learners struggle with vocabulary in all domains, but vocabulary can be particularly challenging in biology domains, where words are often neoclassical compounds using Greek and Latin morphemes. Traditional vocabulary instruction often teaches learners to recognize morphemes in order to learn vocabulary. For example, the word “abduction” which means “to move away/apart” can be recognized as ab+duct+ion or “away lead process” while related English words “duct” and “adduction” can be recognized as an antonym since “ad” means “towards.” In terms of cloze items, these morphemes can provide hints as partial completions for the missing word, e.g. “__duction,” which is the primary way they were used in MoFaCTS. though one could also design cloze item practice to learn morphemes.

Our approach to this problem was to use Wiktionary as training data for GPT-2 with the task of generating morphemes and their corresponding meanings for a given input word (Yarbro & Olney, 2021b). Wiktionary is a Wikimedia project, similar to Wikipedia. Wiktionary includes dictionary information as well as etymologies and pronunciations. In piloting, it became clear that GPT-2 was much better at performing this task when given a definition for the target word in addition to the word itself. Presumably definitions helped GPT-2 by allowing it to relate the meaning of the morphemes generated to the overall meaning of the word. Our evaluation found that when the definition of the input word was provided, the accuracy of morpheme segmentation for English morphemes was 92.5% and for non-English morphemes was 89%. Similarly the morpheme meaning correctness measured by the ROUGE-L metric was .81 for English morphemes and .95 for non-English morphemes.

Definition generation

Definitions are a core part of learning vocabulary and are complementary to morpheme generation described above. Providing definitions for words in context, on demand, can improve reading comprehension for unknown words and facilitate vocabulary learning.

Using the same approach as described for morpheme generation, we created a definition generation model that takes a word in context and generates a definition (Yarbro & Olney, 2021a). Some example generation outputs are:

Context: You can bank on it

Definition: Have faith or confidence in.

Context: A bank of snow

Definition: A slope, mass, or mound of a particular substance

We conducted an evaluation using four domains (American Government, A&P, Astronomy, and Psychology) that compared human definitions, generated definitions using a sentence of context, and generated definitions using multiple sentences of context. Interestingly, using multiple sentences of context was significantly worse across all domains for definition accuracy than using a single sentence of context, though there was no difference in terms of definition fluency. Human definitions were rated significantly more accurate than short context generated definitions across topics, but interestingly for A&P the difference between human and short context definitions was negligible. This suggests that for some

domains, our generated definitions may be indistinguishable from human definitions, though our generated definitions are clearly worse in some domains, like American Government and Astronomy.

Discussion and Recommendations for Future Research

Our experience during the transition period towards using Transformer LLMs for NLG appears to be broadly representative of the field of adaptive training using language as an instructional medium. Our effort in content generation on the MoFaCTS project shifted from an effort spanning knowledge extraction/engineering and content generation to a single step of content generation. Additionally, the shift from formal representations in content generation to a LLM approach of text-in and text-out significantly increased the speed of system development and the quality of generations. This is a startling change from our previous work (Olney et al., 2010, 2012; Olney, 2018) where the AI-generated text always needed human review to ensure quality/accuracy. LLM-generated content in the MoFaCTS project is typically always suitable for immediate use by students, though complete accuracy is never guaranteed. It is clear that Transformer LLMs have changed the landscape of adaptive training and are poised to become the dominant paradigm in AI for the foreseeable future. Additionally, Transformer LLMs are increasing the participation of researchers in the field by reducing the expertise demands for creating adaptive systems: it is no longer necessary to have a Ph.D. in AI/ML (artificial intelligence/machine learning) to create effective adaptive training systems.

However, with this change in landscape there are new concerns and new opportunities for research. The field continues to be broadly concerned with accuracy of generated output (sometimes referred to as hallucinations) and the general control of such output to avoid undesirable characteristics. Additionally, the output of these systems is now so good that automated evaluation approaches generally are not sensitive enough and human evaluations are needed (van der Lee et al., 2019). The approach to human evaluation we used in the MoFaCTS project (Olney, 2023), which borrowed ideas from evaluation of machine translation, is a good model for future researchers to assess where the systems are working as intended. Additionally, we recommend continuous feedback from learners for any questionable content, such that it can be flagged and reviewed by an instructor or expert. In our experience, it is not adequate to ask instructors to review all content because many will not do it.

Curiously, the current landscape of Generative AI and how to incorporate it into systems is very comparable to working with a human collaborator who is not fully trusted. A human collaborator can manifest factual errors, etc., just like a Transformer LLM. How we deal with this situation is what we should have been doing in our adaptive learning systems all along – undertaking rigorous evaluations of effectiveness and performing ongoing telemetry and improvement to remove weaknesses. Hopefully by reducing demand for development time, Generative AI will increase researcher time to engage in evaluation, field deployment, and continuous improvement of fielded systems.

Recommendations for GIFT and STEEL-R Overall

Our success with LLMs suggests that they are broadly applicable to GIFT and its use in STEEL-R (Synthetic Training Environment Experiential Learning for Readiness), as well as future intelligent tutoring systems generally. Our work specifically informs GIFT's Pedagogical Model and Domain Model. In the Pedagogical Model, LLMs can provide refutational feedback. This personalized feedback would provide an explanation of the learner's error rather than providing a generic correct answer. This capability could also be more generally applied in the synthetic training environments of STEEL-R. A straightforward application would be refutational feedback of team communications. If there was a misunderstanding or breakdown in communications, an LLM could provide feedback based on the history of communications

and a synthetic question like “Why was erroneous action A undertaken based on this conversational history?” As suggested by this example, STEEL-R would only need qualitative information about raw data in order to have an LLM provide feedback this way, e.g. correct/erroneous. For example, in shooting training, raw data might be coordinates of target hits, but qualitatively these could be described as left/right and high/low. Based on some training history, a feedback-oriented LLM could suggest trigger control or breath control remediation using this qualitative data. The Pedagogical Model could also be supported by our work on definition generation and morpheme generation. When the learning task involves definition-rich doctrine or specialty training, LLMs can provide in-context definitions instantly and break down less familiar words into units/subwords that make them easier to remember. For example, “amphibious” means dual life, i.e. life on both land and water.

In the Domain model, LLMs can create MCQs automatically based on domain content. Automatic generation of questions can both lower assessment costs and speed up development of training materials. LLMs can further be used for a variety of domain authoring tasks like generating introductory text, creating learning objectives, or summarizing/selecting text based on specific objectives and the learner’s knowledge state. For STEEL-R, LLMs could be used to develop and illustrate the Knowledge, Skills, Abilities, and Attitudes (KSAAAs) being assessed. Additionally, LLMs could be used to design or assist with the design of training scenarios based on learning objectives, including narrative elements, potential challenges, and specific tasks. LLMs have already shown great promise in producing programming code from specifications; to the extent that a synthetic training environment can be defined by code, there is potential for LLMs to create code that implements a given scenario.

Conclusions

One of the great bottlenecks in the development of intelligent tutoring systems has been the creation of training content: adaptive training systems need exponentially more content than static systems in order to provide personalized instruction to a variety of learners. Before LLMs, symbolic AI systems were used to help author intelligent tutoring systems, but these symbolic AI systems were themselves exceedingly difficult and time-consuming to construct.

In this chapter, we explained how Transformer-based LLMs have dramatically changed the status quo for generation of training content. Our project developing the MoFaCTS system for Anatomy & Physiology straddled the historical inflection point of LLMs, leading us to pivot from traditional statistical and symbolic methods to an LLM-dominated approach. Our experience appears to be broadly representative of the field of adaptive training: the shift from formal representations in content generation to a LLM approach of text-in and text-out significantly increased the speed of system development and the quality of generations. This work, and LLMs generally, are broadly applicable to GIFT and STEEL-R, particularly for feedback, content generation, and assessment.

Acknowledgements

This material is based upon work supported by the Institute of Education Sciences under Grant R305A190448 and by the National Science Foundation under Grants 1918751 and 1934745.

References

Betts, J. G., Desaix, P., Johnson, E., Johnson, J. E., Korol, O., Kruse, D., Poe, B., Wise, J. A., Womble, M., & Young, K. A. (2017). *Anatomy and Physiology*. OpenStax.

- Cheng, X., & Roth, D. (2013). Relational Inference for Wikification. In D. Yarowsky, T. Baldwin, A. Korhonen, K. Livescu, & S. Bethard (Eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1787–1796). Association for Computational Linguistics. <https://aclanthology.org/D13-1184>
- Conesa, J., Storey, V. C., & Sugumaran, V. (2010). Usability of upper level ontologies: The case of ResearchCyc. *Data & Knowledge Engineering*, 69(4), 343–356. <https://doi.org/10.1016/j.datak.2009.08.002>
- Corbett, A.T., & Anderson, J.R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Model User-Adap Inter*, 4, 253–278. <https://doi.org/10.1007/BF01099821>
- Domingos, P. (2015). *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Basic Books.
- Finkel, J. R., Manning, C. D., & Ng, A. Y. (2006). Solving the Problem of Cascading Errors: Approximate Bayesian Inference for Linguistic Annotation Pipelines. In D. Jurafsky & E. Gaussier (Eds.), *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 618–626). Association for Computational Linguistics. <https://aclanthology.org/W06-1673>
- Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1), 65–170.
- Gatt, A., & Reiter, E. (2009). SimpleNLG: a realisation engine for practical applications. *ENLG '09: Proceedings of the 12th European Workshop on Natural Language Generation*, 90–93.
- Johansson, R., & Nugues, P. (2008). Dependency-based syntactic-semantic analysis with PropBank and NomBank. *CoNLL '08: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, 183–187.
- Napoles, C., Callison-Burch, C., & Post, M. (2016). Sentential Paraphrasing as Black-Box Machine Translation. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (pp. 62–66). San Diego, California: Association for Computational Linguistics.
- Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and Family: A Review of 17 Years of Natural Language Tutoring. *International Journal of Artificial Intelligence in Education*, 24(4), 427–469. <https://doi.org/10.1007/s40593-014-0029-5>
- Olney, A. M. (2018). Using novices to scale up intelligent tutoring systems. In Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2018.
- Olney, A. M. (2021). Generating Response-Specific Elaborated Feedback Using Long-Form Neural Question Answering. *Proceedings of the Eighth ACM Conference on Learning @ Scale*, 27–36. <https://doi.org/10.1145/3430895.3460131>
- Olney, A. M. (2021). Paraphrasing Academic Text: A Study of Back-Translating Anatomy and Physiology with Transformers. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova (Eds.), *Proceedings of the 22nd International Conference on Artificial Intelligence in Education* (pp. 279–284). Springer International Publishing.
- Olney, A. M. (2022). Generating Multiple Choice Questions with a Multi-Angle Question Answering Model. In S. E. Fancsali & V. Rus (Eds.), *Proceedings of the 3rd Workshop of the Learner Data Institute* (pp. 18–23). <https://doi.org/10.5281/zenodo.7761561>
- Olney, A. M. (2023). Generating multiple choice questions from a textbook: LLMs match human performance on most metrics. In S. Moore, J. Stamper, R. Tong, C. Cao, Z. Liu, X. Hu, Y. Lu, J. Liang, H. Khosravi, P. Denny, A. Singh, & C. Brooks (Eds.), *Proceedings of Empowering Education with LLMs—The Next-Gen Interface and Content Generation*. CEUR-WS.org. <https://ceur-ws.org/Vol-3487/paper7.pdf>
- Olney, A. M., Brawner, K., Pavlik, P., & Koedinger, K. R. (2015). Emerging Trends in Automated Authoring. *Design recommendations for adaptive intelligent tutoring systems: Learner modeling, Vol. 3 of Adaptive Tutoring* (pp. 227–242). Orlando: U.S. Army Research Laboratory.
- Olney, A. M., Gilbert, S. B., & Rivers, K. (2021). Preface to the Special Issue on Creating and Improving Adaptive Learning: Smart Authoring Tools and Processes. *International Journal of Artificial Intelligence in Education*, 32, 1–3. <https://doi.org/10.1007/s40593-021-00277-9>
- Olney, A. M., Graesser, A. C., & Person, N. K. (2010). Tutorial Dialog in Natural Language. In R. Nkambou, J. Bourdeau, & R. Mizoguchi (Eds.), *Advances in Intelligent Tutoring Systems, Studies in Computational Intelligence* (Vol. 308, pp. 181–206). Berlin: Springer-Verlag.
- Olney, A. M., Graesser, A. C., & Person, N. K. (2012). Question Generation from Concept Maps. *Dialogue & Discourse*, 3(2), 75–99.

- Olney, A. M., Pavlik Jr., P. J., & Maass, J. K. (2017). Improving Reading Comprehension with Automatically Generated Cloze Item Practice. In E. André, R. Baker, X. Hu, M. M. T. Rodrigo, & B. du Boulay (Eds.), *Artificial Intelligence in Education* (pp. 262–273). Springer. <https://doi.org/10.1007/978-3-319-61425-0>
- Pavlik, P. I., & Eglinton, L. G. (2021). The Mobile Fact and Concept Textbook System (MoFaCTS) Computational Model and Scheduling System. In S. Sosnovsky, P. Brusilovsky, R. Baraniuk, & A. Lan (Eds.), *Proceedings of the Third International Workshop on Intelligent Textbooks 2021* (Vol. 2895, pp. 93–107). CEUR. <https://ceur-ws.org/Vol-2895/#paper07>
- Pavlik Jr., P. I., Olney, A. M., Banker, A., Eglinton, L., & Yarbrow, J. (2020). The Mobile Fact and Concept Textbook System (MoFaCTS). In S. Sosnovsky, P. Brusilovsky, R. Baraniuk, & A. Lan (Eds.), *Proceedings of the Second International Workshop on Intelligent Textbooks 2020 co-located with 21st International Conference on Artificial Intelligence in Education (AIED 2020)* (pp. 35–49).
- Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., & Miller, A. (2019). Language Models as Knowledge Bases? *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2463–2473. <https://doi.org/10.18653/v1/D19-1250>
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140), 1–67.
- Roberts, A., Raffel, C., & Shazeer, N. (2020). How Much Knowledge Can You Pack Into the Parameters of a Language Model? *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5418–5426.
- Sinatra, A. M. (2015). A Personalized GIFT: Recommendations for Authoring Personalization in the Generalized Intelligent Framework for Tutoring. In D. Schmorow & C. M. Fidopiastis (Eds.), *Foundations of Augmented Cognition—9th International Conference, AC 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings* (Vol. 9183, pp. 675–682). Springer. https://doi.org/10.1007/978-3-319-20816-9_64
- Singley, M. K., Anderson, J. R., Gevins, J. S., & Hoffman, D. (1989). The algebra word problem tutor. *Artificial Intelligence and Education*, 267-275.
- Tafjord, O., & Clark, P. (2021). *General-Purpose Question-Answering with Macaw*. arXiv. <https://doi.org/10.48550/ARXIV.2109.02593>
- van der Lee, C., Gatt, A., van Miltenburg, E., Wubben, S., & Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. *Proceedings of the 12th International Conference on Natural Language Generation*, 355–368. <https://doi.org/10.18653/v1/W19-8643>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.
- Yarbro, J. T., & Olney, A. M. (2021a). Contextual Definition Generation. In S. A. Sosnovsky, P. Brusilovsky, R. G. Baraniuk, & A. S. Lan (Eds.), *Proceedings of the Third International Workshop on Intelligent Textbooks* (Vol. 2895, pp. 74–83). CEUR-WS.org.
- Yarbro, J. T., & Olney, A. M. (2021b). WikiMorph: Learning to Decompose Words into Morphological Structures. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova (Eds.), *Proceedings of the 22nd International Conference on Artificial Intelligence in Education* (pp. 406–411). Springer International Publishing.

CHAPTER 10 – APPLICATIONS FOR GENERATIVE AI IN INTELLIGENT TUTORING SYSTEMS IN ARMY EDUCATION

Rebecca L. Robinson¹, Erica D. Kessler², Randy J. Brou², Robert A. Sottolare³, and Diego Zapata-Rivera⁴
Army University¹; U.S. Army Research Institute for the Behavioral and Social Sciences²;
Soar Technology, LLC, an Accelint Company³; ETS Research Institute⁴

Introduction

The Army Learning Concept 2030-2040 emphasizes the importance of leveraging emerging technologies, such as intelligent tutoring systems (ITSs) and artificial intelligence (AI), to create "learner-centric environments with realistic and experiential-focused training and education" for Soldiers and Army civilian professionals (U.S. Army Training and Doctrine Command, 2024, p. 24). However, the adoption of these technologies within the Army Learning Enterprise (ALE) has been hindered by a lack of clear guidance and standards. The rapid evolution of AI, exemplified by the widespread impact of ChatGPT since its release in 2022, underscores the need for prompt development and implementation of policies, guidance, and standards for AI use within the ALE and beyond. The education technology market is expected to reach \$30 billion by 2032, with AI playing a significant role (Wadhvani, 2023). Moreover, research by Batsaikhan and Correia (2024) highlights the rapid expansion of ITS development, which is instrumental in harnessing the potential of AI in education.

A significant barrier to widespread implementation of ITSs is the substantial time and effort required to develop multi-modal content for adaptive courses. However, the rapid advancement of Generative AI (GenAI) offers a potential solution to this challenge. By utilizing GenAI to generate adaptive content from existing courses or in conjunction with course development, the process of creating and implementing adaptive courses can be significantly accelerated.

This chapter will explore the applications and future directions of GenAI in ITSs, with a focus on curriculum development, instruction, and experiential learning, while also considering the unique requirements of operating in an Army learning environment. The discussion will delve into the potential of GenAI to enhance the efficiency and effectiveness of ITSs, ultimately supporting the ALE's goal of providing personalized, adaptive, and immersive learning experiences for Soldiers and Army civilian professionals.

The Role of Generative AI and Recommendations for Future Research

Curriculum Development and Instruction

The integration of GenAI within ITSs has the potential to revolutionize personalized learning, assessment, and feedback, ultimately leading to improved student learning outcomes. By leveraging GenAI, ITSs can enhance curriculum development and course instruction in several key ways. For instance, GenAI tools can significantly assist with content creation, reducing the labor-intensive process of authoring and developing adaptive courses. These advanced tools can generate a wide range of instructional materials, including lesson plans, scenarios, and interactive exercises, which are tailored to specific learning objectives and outcomes. Additionally, GenAI can create comprehensive question banks, complete with varied question types, such as multiple-choice, short-answer, and essay questions, as well as simulated conversations and role-playing exercises. These generated materials can be seamlessly integrated into an ITS, enabling the system to deliver personalized and adaptive instruction to students. The ITS can then use these materials to create customized learning pathways, provide real-time feedback and assessment, and adjust the difficulty level of exercises based on individual student performance. By automating the content creation process,

GenAI tools can help ensure that instructional materials remain relevant, engaging, and effective in meeting the needs of students, while also freeing up instructors to focus on higher-level tasks, such as mentoring, coaching, and providing expert guidance.

The combination of ITSs and GenAI enables personalized assessments with targeted feedback, allowing instructors to tailor content and learning paths to individual students based on their performance, learning preferences, and progress. This personalized approach addresses specific learning needs and gaps, fostering increased student engagement and improved learning outcomes. Furthermore, data collected from ITS interactions and analyzed by AI can provide valuable insights into common areas of difficulty and trends in student performance, informing curriculum refinement and updates to better align with learning objectives and evolving educational standards.

Instructors can utilize early assessments or pre-tests to evaluate incoming students' prior knowledge levels and then employ AI to develop content tailored to differing levels. By creating personalized learning paths based on students' strengths and weaknesses, instructors can offer targeted support and resources, ultimately enhancing the learning experience. Traditional ITSs provide informative data, but GenAI can offer more comprehensive insights into the learner and course material, enabling instructors to customize their teaching materials more effectively. By leveraging GenAI to author and inform dashboards and tools, instructors can easily identify learners' progress and preferences, suggesting tailored content and resources to better meet individual needs.

The application of GenAI in ITSs is a rapidly evolving field, with current research focusing on various areas, including the generation of student feedback (Pardos & Bhandari, 2023), narrative-based simulation scenarios (Lester, 2024), qualitative coding (Liu et al., 2024), tutoring dialogues (Hu, 2023), and conversation-based assessments (Zapata-Rivera et al., 2024). By harnessing the potential of GenAI, educators can enhance the efficiency and effectiveness of course instruction and content creation in Professional Military Education (PME) and the Civilian Education System (CES), ultimately improving learning outcomes and better preparing students for success.

Experiential Learning

A possible area of exploration is testing applications for using GenAI to aid in the authorship of scenarios for experiential learning. The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) has conducted research on interpersonal leadership skills using Scenario-Based Free Response Agents. These automated assessment tools are computer-based trainings that use Natural Language Processing (NLP) to provide real time interactions with a virtual agent during a learning experience (Brou et al., 2018). Authoring these scenarios takes a great deal of time and editing, which could be overcome by using GenAI to assist in the writing and editing process. Consider an example in which an officer's ability to evaluate an After-Action Report (AAR) is being trained using automated assessment tools, using scenarios generated by AI and vetted by Subject Matter Experts (SMEs).

Using GenAI to create situational judgement tests (SJT) for assessing and teaching interpersonal skills or tacit knowledge can assist in overcoming some of the difficulties in developing these tests, such as how time-consuming creating scenarios and viable, scalable responses can be. Using GenAI may also reduce bias if pulling from a larger information base than a select few experts. This enhanced learning environment could provide greater learning gains during training and education for a variety of situations.

GenAI can be used to support the creation of interactive report systems (e.g., to support AAR). Current work in educational assessment involves the use of Large Language Models (LLMs) to support teacher understanding and use of assessment information. By using LLMs, teachers can ask questions to an agent about assessment information provided in the report (Zapata-Rivera et al., 2025). Similarly, GenAI agents can support AAR sessions.

A potentially impactful application for GenAI is to aid in the authorship of scenarios for experiential learning. ARI has conducted research exploring the assessment and development of interpersonal leadership competencies using *reactive, open-response assessments* (RORAs) that present interactive, virtual leadership scenarios to Army students (Brou et al., 2018). RORAs allow users to input unprompted, free-text responses (i.e., what a student wants to say) during conversations with virtual agents. Using NLP techniques, student inputs are parsed for meaning/intent. Those intents, combined with the situational factors within the scenarios, are used to determine virtual agents' behaviors/reactions in order to carry the scenarios forward down realistic paths. Scenario outcomes and the student responses that led to those outcomes can then be scored for the degree to which interpersonal leadership competencies were displayed.

Importantly, the validity of such scoring will naturally be a function of how well the scenarios elicit the leader competencies they are attempting to assess. RORA scenarios used in ARI research to date were carefully revised through a period of development to ensure students could not successfully navigate the scenario without engaging in the leadership competencies being measured. The process to validate the scenarios was time consuming but necessary. Developing RORAs without validating the criterion creates the potential for inaccurate feedback. That is, if a student is able to successfully navigate a scenario that is purported to measure perspective-taking without actually engaging in perspective-taking (e.g., by bringing to bear a different interpersonal competency), the inferences drawn from the assessment will likely be flawed. For this reason, among others, the authoring of RORA scenarios presents a unique challenge for researchers and/or training developers. However, once a scenario has been validated, RORAs offer a compelling method for assessing and developing interpersonal leader competencies at scale. Appropriately trained GenAI may expedite the authoring process for RORAs by extracting the important underlying features of validated scenarios and reconfiguring their surface features in ways that leave the assessments' validity intact but allow for multi-scenario test batteries that can probe where students are on a developmental trajectory with greater certainty.

Once a student's level of development has been accurately assessed, scenario-based experiential learning techniques can be leveraged to provide students practice opportunities that have been tailored for their unique needs. This tailoring may come, for example, in the form of modifying the relative difficulty of the next scenario in a sequence of scenarios to fit a student's developmental needs. If trained with a set of underlying principles that influence scenario features such as difficulty, GenAI could be leveraged to author new scenarios or procedurally generate new scenarios to meet such demands. Researchers would need to conduct validation studies to ensure such modifications resulted in psychometrically sound scenarios; but once clear of that hurdle, this use of GenAI could allow for a near-limitless opportunity for experiential learning episodes.

While the automated generation of validated assessment items is an important consideration for the future of experiential learning methods, so too is the automation of the scoring and feedback provision. Currently, evaluating the pattern of student responses to scenarios and providing targeted feedback requires complex qualitative analysis that must be carried out in part or whole by SMEs. This requirement creates a serious bottleneck for leveraging such tools for student development at scale. GenAI could be trained to conduct qualitative analyses and provide feedback responses based on methods and insights used by SMEs. Moreover, GenAI models might be designed to reflect the priorities and experiences of a variety of SMEs who might each have their own nuanced interpretations of student responses and what feedback would serve a given student best. Over time, students may be identified as responding more positively to certain models allowing for a tailored match to be made between individual students and the experts from whom they learn most readily.

Army-Specific Considerations

To ensure seamless adoption and implementation, emerging technologies must be tailored to the Army's learning environment and be user-friendly for educators with varying levels of technological expertise. The

2024 Army University Learning Symposium, which focused on "Artificial Intelligence Applications for Learning," highlighted key concerns among attendees. These included the need for ease of use, streamlined integration, and minimal technical requirements for training developers and instructors across the ALE. Many attendees emphasized that senior leadership support for innovation, as well as the removal of bureaucratic and systemic obstacles, is crucial for successful implementation. Furthermore, the tools must be accessible on Army-networked computers, while also being compatible with students' personal devices. Addressing these concerns is essential for the effective integration of GenAI and ITSs into PME and the CES.

Currently Available GenAI Chatbots for Military Use

The use of ChatGPT (and most other public GenAI chatbots) is restricted in the Army, due to security and contextual concerns, motivating several efforts to obtain secure, contextually appropriate GenAI systems for military use. Currently, the Army has access to three main GenAI systems that function like the publicly accessible ChatGPT. Each of these systems is employed using the standard NLP techniques seen in GenAI, building off LLMs of Army specific, doctrine-based information. Here, we will provide a very general overview of the features and capabilities of each of these systems, with the caveat that these platforms are continuously updated to include additional characteristics. While each of these GenAI chatbots has a variety of factors to consider, overall, they provide military personnel with access to the advanced capabilities of GenAI to serve research, education, training, and communication needs across the Army. The remainder of this section will provide a brief, informal review of each of these systems with implications and considerations for use.

CamoGPT was the first GenAI platform introduced specifically for Army use, with an interface and experience similar to ChatGPT. CamoGPT requires a simple registration process and once approved, users have free access to all its features. The current cutoff date for CamoGPT is December 2023, indicating that the system cannot access or process new information after that date. Overall, the system is simple to use, providing users with a basic chat box and dialogue experience. CamoGPT response time is fairly quick, ranging from less than 1-10 seconds, and responses are given in a third person neutral perspective. CamoGPT's typical response format consists of easily comprehensible paragraphs and occasionally bullet points and/or lists, when appropriate. While the overall user experience with CamoGPT is what one would expect, the system does crash regularly and at times becomes unresponsive.

NIPRGPT is another GenAI platform recently introduced in the military, developed and shared by the Air Force Research Lab (AFRL). NIPRGPT allows users immediate access upon registration and has additional features such as Retrieval-Augmented Generation (RAG), a combination of using a conventional information retrieval system in conjunction with generative LLM capabilities. Users have the ability to upload up to 20 sources of information, and increased safety, trust, and security features imperative for military use. Like CamoGPT, the cutoff date for NIPRGPT is December 2023, however NIPRGPT provides a slightly more advanced user interface and experience, with more mature graphics and saving capabilities. Response time of NIPRGPT is slightly faster, with a typical response time of less than 1-5 seconds, providing responses in a third person neutral perspective. No crashing or user experience issues were present during use, allowing for effective use of the chatbot.

Ask Sage is the final GenAI platform we casually assessed for the purposes of this chapter. Ask Sage also allows immediate access upon registration. Perhaps the most impressive aspect of Ask Sage is the vast number of features, which include plugins, upload and download capabilities, as well as customization of the user dashboard. Ask Sage also has the most recent cutoff date (October 2024) and has a library of videos, resources, and user guides for a range of contexts and applications. In terms of user experience, Ask Sage provides perhaps the most visually pleasing environment and generates responses that mimic more of a lesson plan with summaries, bullet points, and outlined information based on the given prompt. While the

free version of Ask Sage provides users with a great deal of access to features, Ask Sage does require the purchase of additional tokens for unlimited access to their suite of advanced features, plugins, and upload capabilities. Ranging from \$15.00-\$90.00 per month, Ask Sage is currently the only GenAI chatbot available for use by military personnel that requires a form of payment for full access, however the capabilities of the free features provide users with as many features (if not more) than CamoGPT or NIPRGPT.

Recommendations for GIFT and STEEL-R Overall

Implementing GenAI into Army-developed ITSs, such as the Generalized Intelligent Framework for Tutoring (GIFT) and the Synthetic Training Environment Experiential Learning for Readiness (STEEL-R) will overcome some of the key barriers that have made large scale implementation of these emerging technologies into the ALE difficult. Specifically using GenAI to create adaptive multimodal content to be used in remediation would greatly assist in the time it takes to set up an adaptive course in GIFT. GenAI can also create adaptive questions not based solely on the initial course content provided by the instructor, but also on the learner's previous knowledge, real-time responses, preferences, and even mood. Having personalized assessments and feedback will engage and motivate learners and improve learning outcomes.

The integration of GenAI into GIFT has the potential to overcome significant barriers that have hindered the large-scale implementation of ITSs within the ALE. By leveraging GenAI, adaptive multimodal content can be created to support remediation, substantially reducing the time and effort required to set up adaptive courses in GIFT. Furthermore, GenAI can generate adaptive questions that take into account not only the initial course content provided by the instructor, but also the learner's prior knowledge, real-time responses, preferences, and emotional state. This enables the creation of personalized assessments and feedback that cater to individual learning needs, fostering increased learner engagement, motivation, and ultimately, improved learning outcomes.

Transforming Curriculum Development with GenAI

The integration of GenAI into an ITS architecture like GIFT offers the potential to make adaptive training experiences more efficient and effective by addressing critical challenges within the ALE. One of the primary obstacles to widespread ITS adoption is the significant time and expertise required to develop adaptive training content. GenAI could dramatically streamline this process by automating content creation. Specifically, GenAI could be used to generate detailed, Army-specific training scenarios and lesson plans tailored to varying levels of expertise (Colby, Tucker, & Siggins, 2024). It could also develop interactive exercises and assessments aligned with defined learning objectives, significantly reducing the time investment required for course development. Specific recommendations for transforming curriculum development within GIFT and STEEL-R follow:

- **Automated Content Creation:** GenAI could rapidly generate instructional materials, such as lesson plans, scenarios, and interactive exercises in both GIFT and STEEL-R that are tailored to specific Army training requirements. GIFT should integrate these AI-generated materials into its existing modular architecture, ensuring they meet defined learning objectives.
- **Customizable Training Scenarios:** Using AI models trained on military-specific datasets, GenAI should produce realistic, context-sensitive scenarios. These scenarios would enhance experiential learning within GIFT, enabling Soldiers and Army civilian professionals to engage in immersive simulations that closely mirror real-world challenges. Another option is to automate the creation of role-specific content, such as leadership scenarios for commanders and their staffs (Sottolare, 2024) or technical troubleshooting exercises for maintenance personnel.

- **Rapid Prototyping of Adaptive Modules:** GenAI should assist instructional designers by quickly creating and iterating on adaptive learning modules. This would align with GIFT’s focus on creating personalized learning pathways, reducing the workload for developers while improving the speed of deployment.

Enhancing Personalization with GenAI

To enhance the adaptability of training experiences, GenAI could be integrated into GIFT to provide real-time personalization. For example, the system could dynamically adjust the difficulty and scope of training materials based on learner performance, using advanced NLP to identify knowledge gaps or areas for improvement. If a Soldier struggles with a specific task, the system could instantly generate supplementary exercises or targeted guidance to address those challenges. Similarly, GenAI-powered conversational agents within GIFT could serve as interactive virtual instructors, answering learner questions, providing tailored feedback, and facilitating group-based problem-solving activities. Specific recommendations for enhancing personalization in GIFT and STEEL-R follow:

- **Dynamic Adaptation:** By integrating GenAI into its learner assessment capabilities, GIFT could provide real-time adjustments to training content (e.g., scenario conditions). For example, GenAI could analyze learner performance during a session and modify the difficulty or type of training material on the fly, creating a more engaging and effective experience. Another option is to use GenAI to create customizable and scalable educational and training scenarios tailored to individual learners or teams based on their roles, skill levels, and operational contexts. For example, GenAI methods could be used to generate mission-specific simulations that evolve dynamically based on learner’s actions.
- **Interactive Problem-Solving Exercises:** Enable GenAI to design real-world problem-solving activities, such as logistical challenges, combat strategy formulation, or decision-making tasks, ensuring alignment with specific learning objectives (Sottolare, 2024).
- **Personalized Feedback:** GenAI could generate nuanced and context-aware feedback based on individual learner interactions. For Soldiers practicing a technical skill, for instance, GIFT could use AI-driven insights to provide step-by-step corrective feedback, ensuring mastery of the skill. GenAI could also be used to recommend individualized learning paths by assessing each learner’s strengths and weaknesses, ensuring readiness across mission essential competencies (MECs; Alliger et al., 2013).
- **Interactive Learning Assistants:** GenAI could serve as an interactive, conversational learning assistant within GIFT and STEEL-R, answering questions, providing hints, or facilitating group discussions during collaborative exercises. An approach based on a fine-tuned generative pre-trained transformer (GPT), a type of AI model that uses deep learning to create human-like text and other content can be easily achieved. For example, Sottolare (2024) demonstrated a GPT capability to support battalion staff training at the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC). A stretch goal is to develop a multi-modal large language model (MLLM) that can process and generate information across multiple data types, like text and images, essentially allowing it to understand and respond to not just written prompts but also visual content like pictures, making it more versatile for tasks requiring both language and vision comprehension.

Expanding Experiential Learning with GenAI

Experiential learning, a key component of Army training, could also be elevated through GenAI innovations. One significant enhancement would be the ability to generate a varying array of scenario

variations on demand (Sottolare et al, 2024). These AI-created scenarios could include adaptive narratives that evolve in response to learner decisions, creating a richer, more immersive training experience. For example, in a combat simulation, GenAI could introduce unpredictable elements such as shifting environmental conditions or adversarial tactics, challenging learners to think critically and adapt in real-time. This approach would prepare Soldiers and civilian professionals for the complexities of real-world situations while fostering decision-making and problem-solving skills.

The integration of GenAI could further enhance the efficiency of AARs by automating the analysis of learner performance. By leveraging machine learning algorithms, GIFT could produce detailed reports that highlight individual and team strengths, identify specific areas for improvement, and recommend tailored follow-up training activities. These AARs could be augmented with visual and data-driven insights to provide a comprehensive understanding of training outcomes. Specific recommendations for expanding experiential learning in GIFT and STEEL-R follow:

- **Scenario Variability:** GIFT could utilize GenAI to dynamically generate variations of training scenarios, ensuring Soldiers face a wide range of challenges and decision-making contexts. This variability could enhance critical thinking and adaptability in learners.
- **Adaptive Simulation Narratives:** Use GenAI to design narrative-driven simulations that adjust based on learner decisions, such as evolving mission parameters, changing environmental conditions, or introducing unforeseen challenges to foster adaptability and critical thinking (Wang et al. 2024; Lester et al., 2024).
- **AI-Driven NPCs:** Integrate conversational AI agents into virtual simulations to act as non-player characters (NPCs). These agents could simulate adversaries, allies, or civilian populations, responding intelligently to learner actions and creating realistic, immersive training experiences.
- **Multi-Domain Integration:** Enable GenAI to design cross-functional scenarios incorporating land, sea, air, cyber, and space domains, ensuring comprehensive readiness for multi-domain operations (MDO).
- **Integration with Virtual and Augmented Reality (VR/AR):** GenAI could work in tandem with VR/AR technologies integrated into GIFT to create highly immersive environments. For example, it could simulate conversational AI-driven adversaries or allies that respond realistically to learner actions, deepening the training experience.
- **AARs:** GenAI could automatically analyze training sessions within GIFT and generate detailed AARs. These reports could highlight strengths, areas for improvement, and actionable recommendations, helping learners reflect on and improve their performance.

Addressing Standardization and Policies

To ensure these enhancements align with the ALE's goals, policy and standardization efforts should guide the integration of GenAI into ITS architectures. GIFT could implement content validation pipelines to verify that AI-generated content meets established Army training standards. Furthermore, the adoption of ethical guidelines and data protection measures would safeguard learner privacy and ensure transparency in AI-driven decisions. Incorporating such measures would establish trust in GenAI-powered systems while maintaining compliance with Army policies. Specific recommendations address standardization and policy compliance in GIFT and STEEL-R follow:

- **Standardized GenAI Models:** GIFT could adopt standardized GenAI frameworks vetted for security and ethical considerations, ensuring compliance with ALE policies while maintaining adaptability.

- **Content Validation Pipelines:** AI-generated content could pass through rigorous validation pipelines within GIFT to ensure alignment with Army training standards and objectives (Sottolare et al., 2023).
- **Ethical Guidelines for AI Use:** GIFT could embed ethical AI usage protocols, ensuring transparency in how learner data is utilized and protecting the privacy and security of Soldiers and civilian professionals (DeFalco & Hampton, 2020).

By embedding these recommended GenAI capabilities into GIFT, the ALE could significantly accelerate the development of adaptive training experiences while enhancing their effectiveness. This approach would not only address current barriers but also position the Army as a leader in leveraging cutting-edge educational technologies to meet the evolving needs of its Soldiers and civilian workforce.

Conclusions

Integrating GenAI into GIFT has the potential to transform the way the Army approaches training and education. By leveraging GenAI, GIFT can overcome key barriers to large-scale implementation, such as the time and effort required to create adaptive multimodal content and develop personalized learning paths. The use of GenAI can also enhance critical thinking and adaptability in learners through the dynamic generation of scenario variations and the creation of interactive exercises and assessments tailored to specific learning objectives. Furthermore, GenAI can facilitate the rapid development of instructional materials, such as lesson plans and scenarios, that meet defined learning objectives and are aligned with Army standards. To ensure the effective and responsible integration of GenAI into GIFT, it is essential to adopt standardized AI frameworks, implement rigorous content validation pipelines, and address standardization and policy considerations. By doing so, GIFT can provide personalized, adaptive, and engaging learning experiences that improve student outcomes and support the Army's mission to develop competent and agile leaders. Ultimately, the successful integration of GenAI into GIFT enables the development of a more effective, efficient, and responsive learning enterprise.

Acknowledgements

Disclaimer: The views expressed in this chapter are those of the authors and do not reflect the official policy or position of the Department of the Army, DOD, or the U.S. Government.

References

- Alliger, G. M., Beard, R., Bennett, W., Jr., & Colegrove, C. M. (2012). Understanding mission essential competencies as a job analysis method. In M. A. Wilson, W. Bennett, Jr., S. G. Gibson, & G. M. Alliger (Eds.), *The handbook of work analysis: Methods, systems, applications and science of work measurement in organizations* (pp. 603–624). Routledge/Taylor & Francis Group.
- Batsaikhan, B.Z., & Correia, A.P. (2024). The effects of Generative Artificial Intelligence on Intelligent Tutoring Systems in higher education: A systematic review. *Studies in Technology Enhanced Learning*, 4(1). <https://doi.org/10.21428/8c225f6e.33570bb1>
- Brou, R., Stallings, G., Stearns, I., Normand, S., & Ledford, B. (2018). Building Automated Assessments of Interpersonal Leadership Skills. Paper presented at the 52nd annual Interservice/Industry Training, Simulation, and Education Conference, Orlando, FL.
- Colby, B., Tucker, E., & Siggins, T. (2024, June). Beyond Standalone Systems: Creating an Ecosystem of Adaptive Training Services. In *International Conference on Human-Computer Interaction* (pp. 3-14). Cham: Springer Nature Switzerland.

- DeFalco, J. A., & Hampton, A. J. (2020). Dewey's Ethics of Moral Principles and Deliberation: Extending IEEE's Ethics Initiative for Adaptive Instructional Systems. In *Adaptive Instructional Systems: Second International Conference, AIS 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22* (pp. 45-54). Springer International Publishing.
- Hu, X. (2023). Empowering education with LLMs - the next-gen interface and content generation. [demo] Presented at *The Inter. Conference on Artificial Intelligence in Education*. (Tokyo, Japan, July 03-07, 2023). AIED 2023
- Lester, J., Bansal, M., Biswas, G., Hmelo-Silver, C., Roschelle, J., & Rowe, J. (2024). The AI Institute for Engaged Learning. *AI Mag.*, 45(1), 69–76. doi:10.1002/aaai.12161
- Liu, X., Zhang, J., Barany, A., Pankiewicz, M., & Baker, R.S. (2024). Assessing the Potential and Limits of Large Language Models in Qualitative Coding. In: Kim, Y.J., Swiecki, Z. (eds) *Advances in Quantitative Ethnography. icqe 2024. Communications in Computer and Information Science*, vol 2278. Springer, Cham. https://doi.org/10.1007/978-3-031-76335-9_7
- Pardos, Z. A., & Bhandari, S. (2023). Learning gain differences between ChatGPT and human tutor generated algebra hints. Retrieved from *arXiv:2302.06871v1*
- Sottolare, R., Ballinger, C., & McGroarty, C. (2023, May). Considerations in the Design of Realistic Agents for Serious Games. In *The International FLAIRS Conference Proceedings* (Vol. 36).
- Sottolare, R., Ballinger, C. B., Litvinas, M., Hu, S., & McGroarty, C. (2024, May). Using Genetic Algorithms to Automate Scenario Generation and Enhance the Training Value of Serious Games for Adaptive Instruction. In *The International FLAIRS Conference Proceedings* (Vol. 37).
- Sottolare, R. (2024, December). Max, A GPT Concept for Battalion Staff Training. Demonstrated at the Soar Technology booth (#138) during the Interservice/Industry Training Simulation & Education Conference, Orlando, Florida.
- U.S. Army Training and Doctrine Command. (2024). *The Army Learning Concept for 2030-2040*. [TRADOC Pamphlet 525-8-2].
- Wadhvani, P. (2023, January). *AI in education market statistics, trends & growth opportunity 2032*. Global Market Insights. <https://www.gminsights.com/industry-analysis/artificial-intelligence-ai-in-education-market>
- Wang, Y., Zhou, Q., & Ledo, D. (2024). StoryVerse: Towards Co-authoring Dynamic Plot with LLM-based Character Simulation via Narrative Planning. *Proceedings of the 19th International Conference on the Foundations of Digital Games*. Presented at the Worcester, MA, USA. doi:10.1145/3649921.3656987
- Zapata-Rivera, D., Forsyth, C. M., Graf, A., & Jiang, Y. (2024). Designing and Evaluating Evidence-Centered Design based Conversations for Assessment with LLMs. *Joint Proceedings of the Human-Centric eXplainable AI in Education and the Leveraging Large Language Models for Next Generation Educational Technologies Workshops (HEXED-L3MNGET 2024)*. CEUR-WS.org, online https://ceur-ws.org/Vol-3840/L3MNGET24_paper3.pdf
- Zapata-Rivera, D., Lehman, B., Lopez, A., Gooch, R., & Long, R. (2025, accepted). Designing and evaluating AI-enhanced interactive reports with teachers. Paper submitted as part of a symposium at the Annual Meeting of the American Educational Research Association. Denver, CO.

A large, thick red curved line that starts from the top right, curves downwards and to the left, and then curves back up towards the right, framing the top portion of the page.

**SECTION III– OPPORTUNITIES AND
IMPLICATIONS OF GENERATIVE AI FOR
TRAINING ENVIRONMENTS**

*Arthur C. Graesser¹, Paige M. Lawton², and Anne M. Sinatra²,
Eds.*

*The University of Memphis Institute for Intelligent Systems¹
US Army Combat Capabilities Development Command (DEVCOM)
– Soldier Center²*

CHAPTER 11 – GENERATIVE AI FOR INSTRUCTIONAL SYSTEMS DESIGN SCALABILITY, EFFICIENCY, AND QUALITY CONTROL

Ramona Shires¹, Robert McCormack¹, Shawn Weil¹, and Gregory Goodwin²

Aptima¹; US Army Combat Capabilities Development Command (DEVCOM) Soldier Center²

Introduction

The Generalized Intelligent Framework for Tutoring (GIFT) was originally created to economize the development of intelligent tutoring systems (ITSs) by leveraging a modular, open-standard architecture (MOSA) to maximize reuse of its components when creating new ITSs (Robson & Barr, 2013). Thus, the goals of maximizing scalability, efficiency, and quality control are part of the DNA of GIFT.

Despite making progress towards realizing these goals, GIFT has yet to fully achieve them. Authoring new courses in GIFT still requires a significant effort (McCarthy & Sinatra, 2023). Quality control challenges remain. For example, in scenario-based experiential training contexts, measuring the performance of individuals and teams objectively can be difficult, even for expert observers (Foltz & Gilbert, 2023). And finally, neither GIFT nor many other learning technologies have end-to-end self-improvement capabilities that would maximize the efficiency of developing and maintaining them (Nye et al., 2019).

Artificial Intelligence (AI), and specifically Generative AI models, represent a transformative advancement in the field of Instructional Systems Design (ISD), bringing unprecedented capabilities to scale, optimize, and ensure the quality of instructional materials. ISD, a structured approach to designing and delivering effective training, traditionally relies on manual and labor-intensive processes to analyze requirements, design content, and implement learning solutions. The integration of Generative AI is poised to revolutionize these workflows by introducing automation, adaptability, and precision.

This chapter explores the transformative impact of Generative AI in ISD, focusing on three key areas: scalability, efficiency, and quality control. The **Scalability** section delves into how Generative AI enables instructional designers to address the challenges of dynamic learning environments through the use of reusable templates and prompt strategies, large data set processing, and model-agnostic systems. In the **Efficiency** section, the chapter highlights the role of AI in automating repetitive cognitive tasks, facilitating human-AI collaboration, and maintaining adherence to structured ISD methodologies to optimize resource use and accelerate development processes. The **Quality Control** section discusses how Generative AI ensures reliability and consistency in instructional content, emphasizing its capabilities in standardizing data inputs, managing file conventions, and providing transparent and explainable outputs.

The chapter also presents real-world **Use Cases and Applications** of Generative AI in military and industry contexts, such as the Navy's Ready Relevant Learning (RRL) initiative and the Army's GIFT tool, illustrating its transformative potential in aligning training delivery with operational needs. Finally, the **Future Perspectives** section reflects on the ethical and technical challenges of Generative AI while exploring opportunities for integration with emerging technologies like augmented and virtual reality.

By examining these topics, the chapter provides a comprehensive overview of how Generative AI can revolutionize ISD, empowering organizations to deliver scalable, efficient, and high-quality training solutions for diverse learner populations.

The Role of Generative AI for ISD

At its core, Generative AI is a groundbreaking technology that uses advanced algorithms to analyze large amounts of data and create new, meaningful content automatically. A key innovation behind Generative AI is the use of *transformers* (Vaswani et al., 2017), a type of machine learning architecture designed to process sequential data, such as text or speech, while capturing long-term relationships and context. Transformers work by assigning importance, or "attention," to different parts of the data, enabling the model to understand the nuances of language and meaning. Building on this foundation, large language models (LLMs), such as GPT (Generative Pre-trained Transformer), are an application of transformers that are trained on vast datasets containing text from books, websites, and other sources. These models can generate coherent and contextually relevant text, answer questions, translate languages, and even create code. In simple terms, Generative AI acts like a highly skilled assistant, capable of understanding and producing human-like content by recognizing patterns in data and combining them in innovative ways. This technology is reshaping industries by automating creative and analytical tasks that were once thought to require human intelligence.

Within the ISD process, the potential applications of Generative AI are vast and transformative, reshaping how training is analyzed, designed, and delivered. One of its most powerful capabilities lies in automating the analysis of learner needs, allowing instructional designers to process large volumes of data—such as survey results, performance metrics, or skill assessments—with unprecedented speed and precision. This enables the identification of knowledge gaps and learning preferences, forming a foundation for more targeted and effective training solutions. Generative AI also excels in creating tailored training modules at scale, adapting content to accommodate diverse learner populations with varying needs, languages, contexts, or technical requirements. As industries and technologies evolve rapidly, training programs often struggle to keep pace with new demands, but AI offers a solution by dynamically generating content that evolves alongside these changes. This adaptability allows instructional designers to address challenges like delivering consistent, high-quality training to geographically dispersed learners or creating scenario-specific modules for highly specialized roles. By offloading time-intensive tasks and streamlining workflows, Generative AI empowers designers to focus on strategic, high-level objectives, ultimately leading to more engaging, accessible, and impactful training experiences that meet the needs of modern learners and organizations.

AI's role extends beyond simple automation. It acts as an enabler, augmenting human creativity and decision-making in the ISD process. Traditionally, ISD practices are constrained by several significant limitations. These include the labor-intensive nature of analyzing learner needs, designing content, and assessing outcomes, as well as the inherent challenges of scaling training solutions to meet the needs of diverse learner populations. Manual processes often struggle to keep pace with rapidly evolving organizational requirements, resulting in outdated or misaligned training content. Furthermore, ensuring consistency and quality across instructional materials—particularly when designing for large, geographically dispersed, or culturally diverse audiences—remains a persistent challenge. These limitations highlight the pressing need for innovative approaches to scalability, efficiency, and quality control in ISD.

Generative AI, particularly LLMs, address these challenges by introducing powerful capabilities that transform how ISD processes are executed. LLMs excel in analyzing and synthesizing vast amounts of data, enabling instructional designers to rapidly identify knowledge gaps, performance trends, and learner preferences. This data-driven insight supports targeted, evidence-based training strategies that were previously difficult or impossible to achieve at scale. Generative AI enhances scalability by autonomously generating tailored instructional materials that adapt to different languages, cultural contexts, and technical requirements, ensuring that training solutions are accessible and relevant for diverse audiences. In terms of efficiency, LLMs automate time-intensive tasks such as curriculum development, scenario creation, and learner assessments, significantly reducing the time and resources needed to implement training programs.

Finally, Generative AI improves quality control by leveraging structured frameworks, templates, and real-time data validation techniques to maintain consistency, accuracy, and alignment with organizational goals.

The integration of LLMs into ISD represents a paradigm shift, allowing organizations to overcome traditional limitations while delivering personalized, scalable, and high-quality learning experiences. By tackling these persistent challenges, LLMs enable instructional designers to focus on higher-order objectives, such as ensuring content is pedagogically sound, engaging, and aligned with strategic goals. These benefits reinforce the relevance of Generative AI in modern training environments, providing organizations with a powerful tool to meet the growing demand for effective, adaptive, and efficient training solutions. Let us explore how Generative AI directly addresses the current limitations in ISD and paves the way for a more innovative and impactful future.

Scalability

Scalability is one of the most significant advantages that Generative AI brings to ISD. By addressing the challenges of reaching diverse audiences and accommodating differing instructional needs, Generative AI enables instructional designers to create and deploy training solutions capable of reaching thousands—or even millions—of learners while adapting content to individual needs. The following delves into the key strategies and technologies that make scalability achievable within ISD frameworks.

Scalability Through Advanced Prompt Engineering Strategies

Generative AI achieves scalability in ISD by leveraging advanced prompt engineering techniques that maximize the model's ability to generate adaptable and context-specific analyses, designs, and instructional content. Prompts are the instructions or inputs provided to a Generative AI model to guide its response or output. They can range from simple questions or commands to more complex, structured guidelines designed to elicit specific types of content. Prompt engineering is the process of crafting and refining these prompts to optimize the model's performance, ensuring that it generates accurate, relevant, and high-quality outputs tailored to the desired use case. By integrating dynamic variables and modular prompts, instructional designers can tailor content to different analytic needs, learner profiles, cultural contexts, and technical requirements without rewriting extensive portions of the prompt.

One key strategy is the use of layered prompts that combine general instructional frameworks with scenario-specific details. This approach allows designers to define high-level training objectives while delegating context-specific customization to the AI. For instance, a base prompt might outline a general lesson structure, while additional layers specify the audience, job role, or technical complexity. This modularity ensures consistency in design while enabling rapid adaptation to new use cases.

Another critical technique is iterative refinement, where instructional designers experiment with and optimize prompts based on the model's outputs. By systematically testing variations and incorporating feedback, designers can fine-tune prompts to achieve consistent, high-quality results across projects. Furthermore, advanced prompt engineering allows for integration with external data sources, such as learner performance metrics or operational requirements, enabling the AI to generate data-informed and highly relevant training solutions at scale.

By focusing on advanced prompt engineering strategies, Generative AI enhances scalability in ISD, enabling instructional designers to efficiently analyze training requirements and create instructional content that is both adaptable and consistent. This approach reduces redundancy, streamlines workflows, and empowers designers to meet the growing demands of diverse and evolving training environments with precision and efficiency.

Leveraging AI to Process and Synthesize Large Data Sets

The capacity to analyze large datasets underpins the scalability of instructional design. AI and machine learning models excel in processing complex, unstructured data, providing actionable insights that inform targeted training solutions. By synthesizing information from assessments, surveys, and operational metrics, AI supports evidence-based practices. This capability enables instructional designers to identify gaps in knowledge, detect performance trends, and refine training materials, ensuring that content remains relevant and effective over time.

Generative AI has unlocked new possibilities for ISD by enabling its application to emerging industries and job roles. In sectors experiencing rapid technological advancements, such as cybersecurity, healthcare, and green energy, Generative AI addresses unique training needs. This scalability is made possible by Generative AI's ability to process and synthesize large data sets, which enables the system to identify and incorporate relevant variables, scenarios, and domain-specific knowledge seamlessly into training solutions. The capacity to handle such complexity ensures that content remains both comprehensive and adaptable to novel challenges.

Compatibility with Various AI Frameworks and Tools

The scalability of Generative AI is further enhanced by its model-agnostic nature. This approach ensures that AI solutions integrate seamlessly with frameworks and tools, maximizing their applicability across organizations. Model-agnostic systems are designed for interoperability with existing Learning Management Systems (LMS) and training platforms, allowing for scalable implementations without requiring significant infrastructure changes. Additionally, the flexibility to align AI capabilities with organizational requirements ensures that training programs remain adaptable and future-proof.

Real-world applications illustrate the potential of Generative AI to achieve scalability. For example, integration of AI within GIFT demonstrates how these models can create adaptive training scenarios tailored to individual learners within ITSs. This automation streamlines content generation and enables dynamic assessments, ensuring scalability while maintaining instructional quality. Similarly, the Navy RRL initiative employs AI to align training delivery with operational needs, optimizing schedules and resource allocation to meet the demands of a rapidly changing environment. These examples, discussed in more detail in a later section, highlight the practical benefits of Generative AI in delivering scalable solutions that address both organizational and learner-specific requirements.

Scalability in ISD is no longer a distant aspiration but a tangible reality enabled by Generative AI. Through variable approaches, reusable templates, data analysis, and adaptability to novel domains, instructional designers can effectively meet the growing demands of dynamic learning environments. The integration of model-agnostic systems and real-world implementations, such as GIFT and Navy RRL, further solidifies AI's role as a pivotal tool in the evolution of instructional design.

Efficiency

Efficiency is a cornerstone of ISD, and Generative AI introduces groundbreaking ways to streamline processes, reduce time demands, and optimize resources. By automating repetitive tasks, enhancing collaboration, and maintaining adherence to structured methodologies, Generative AI enables instructional designers to achieve more with less effort.

Multi-Prompt Batch Generation and Automation

Generative AI models address inefficiencies in instructional design processes through techniques like multi-prompt batch generation. This methodology automates the creation of multiple instructional variants, interrelated analytics tasks, and repetitive workflows with speed and precision. By leveraging this approach, instructional designers can dynamically produce content tailored to diverse needs, such as adapting lesson plans for various learner profiles, translating materials into different languages, or customizing training for specific cultural contexts and scenarios. For example, a single workplace safety training template can be quickly adapted to suit different job roles, industries, and regulatory requirements. This not only accelerates the content development process but also ensures that training materials remain highly relevant, accurate, and impactful.

Human-AI Teaming

Generative AI is designed to complement human expertise rather than replace it, creating a partnership where both humans and AI contribute to the efficiency of ISD. In this collaborative relationship, AI manages computational and repetitive tasks, freeing instructional designers to focus on higher-level creative and strategic decisions. This collaboration fosters an environment where creativity flourishes, as AI generates suggestions, examples, and drafts that spark innovative ideas. Additionally, the inclusion of AI in the workflow reduces the likelihood of errors by analyzing and validating data against predefined objectives. By optimizing the division of labor between humans and AI, the design process becomes more streamlined, allowing teams to achieve greater productivity and maintain pedagogical integrity.

Structured Process Control

In ISD, efficiency must go hand in hand with consistency and accuracy. Generative AI supports structured process control by embedding ISD methodologies into its workflows, ensuring that every step of the process adheres to best practices. This integration enables the automation of critical checkpoints where AI validates that content aligns with learning objectives and organizational goals. Furthermore, AI enhances version control by managing multiple iterations of instructional materials and maintaining a clear history of changes. The use of standardized templates reinforces uniformity across training programs, reducing variability and enhancing reliability. By integrating these process controls, Generative AI minimizes deviations from established protocols, ensuring that efficiency does not come at the cost of quality.

Technology in Practice

The real-world applications of Generative AI offer compelling examples of its efficiency-enhancing potential. Insights from a 2024 Interservice/Industry Training, Simulation and Education Conference (I/ITSEC) tutorial presentation (Shires & McCormack, 2024) highlight how AI-driven solutions address the needs of military training environments by enabling scalable content creation, facilitating data-driven adjustments, and optimizing resource allocation. For instance, Generative AI can rapidly develop training modules tailored to specific military operations and scenarios, ensuring that content remains aligned with operational demands. Moreover, AI's ability to analyze performance data enables real-time updates to instructional content, maintaining its relevance and effectiveness. The automation of administrative tasks, such as scheduling and resource allocation, further frees personnel to focus on mission-critical activities, demonstrating how AI enhances both operational and instructional efficiency in complex environments. Efficiency in ISD has been redefined by the capabilities of Generative AI, which enables organizations to achieve unprecedented levels of productivity and quality. Through batch generation, human-AI teaming, and structured process control, Generative AI streamlines the creation and delivery of training programs. The integration of these technologies, as evidenced by their success in military applications, highlights the transformative impact of AI on ISD. By embracing these innovations, organizations can deliver impactful

learning experiences with unparalleled efficiency, setting new benchmarks for instructional design in the digital age.

Quality Control

In the context of ISD, quality control is a critical element that ensures the reliability, consistency, and effectiveness of training programs. The integration of Generative AI into ISD workflows introduces new dimensions to quality management, enhancing the capacity to maintain high standards across all aspects of instructional design. Consistency is foundational to effective instructional design, as it minimizes variability and ensures that learners receive uniform training experiences.

Consistency of Data Inputs and Structure

Consistency is a cornerstone of quality control in instructional design, as it ensures that training materials are cohesive, uniform, and aligned with learning objectives. Generative AI addresses this need by standardizing data inputs and structures, which are essential for producing outputs that meet high-quality standards. This process begins with clearly defining input parameters, such as training objectives, learner profiles, and contextual requirements, and ensuring these inputs follow a structured format. By requiring standardized inputs, AI systems can minimize variability and enhance the uniformity of the generated outputs, regardless of the specific content or context.

Structured templates and predefined frameworks play a pivotal role in maintaining this consistency. These templates can serve as the foundation for a wide variety of training modules, ensuring that key elements—such as terminology, formatting, instructional flow, and learning objectives—are consistent across different iterations of content. For instance, a Generative AI system used to create technical training materials can enforce the use of specific industry-standard terms, maintain consistent formatting, and ensure that content is organized in a logical sequence that aligns with pedagogical best practices. This reduces the risk of discrepancies or omissions that could confuse learners or detract from the overall effectiveness of the training.

Furthermore, Generative AI can incorporate advanced tagging and metadata systems to ensure that data inputs are organized and easily retrievable. For example, instructional designers can tag content with relevant keywords, version numbers, and intended audiences, enabling the AI to automatically select and assemble the most appropriate components for specific use cases. This structured approach ensures that training materials remain consistent not only in individual modules but also across broader training programs, supporting organizational goals of scalability and repeatability.

Another advantage of consistency in data inputs and structures is the ability to streamline updates and revisions. Because content is built on a standardized foundation, changes made to one part of the framework—such as updating a safety regulation or company policy—can be automatically propagated across all affected modules. This reduces the time and effort required for manual updates while ensuring that all learners receive accurate and up-to-date information.

By implementing and enforcing consistent data structures, Generative AI enhances the reliability, coherence, and scalability of instructional materials. This approach allows instructional designers to focus on refining content quality and pedagogical soundness, knowing that the underlying system is maintaining a strong foundation of consistency across all outputs. As a result, learners benefit from a seamless and uniform experience, while organizations achieve greater efficiency and quality in their training programs.

Ensuring Content Reliability

The accuracy of instructional content is paramount, as errors can undermine the credibility and effectiveness of training programs. Generative AI supports accuracy by leveraging advanced data validation techniques and cross-referencing content against established objectives. For instance, AI can automatically flag inconsistencies or inaccuracies in generated materials by comparing them to a set of predefined criteria or standards. Furthermore, continuous feedback loops, enabled by AI, allow instructional designers to refine content based on real-time performance data and learner assessments. This iterative process not only improves the precision of instructional materials but also ensures alignment with the desired learning outcomes.

File Naming Conventions and Strategies

Effective file management is often overlooked but plays a vital role in maintaining quality control within ISD projects. Generative AI aids in this area by implementing standardized file naming conventions and organizational strategies. By automating these processes, AI ensures that files are named consistently and intuitively, making it easier for teams to locate, manage, and version instructional materials. For example, AI systems can append metadata such as creation dates, project names, and version numbers to file names, providing clear identifiers that facilitate seamless collaboration and version control. These strategies reduce the risk of errors stemming from misplaced or outdated files, enhancing the overall efficiency and quality of the ISD workflow.

Ensuring Transparency in AI-Driven Processes

As Generative AI becomes an integral part of ISD, the need for transparency in AI-driven processes has grown significantly. Explainability—the ability to understand and interpret how AI systems generate outputs—is essential for building trust and ensuring accountability. Generative AI achieves this by providing detailed rationales for its decisions and outputs. For instance, when generating instructional content, AI can document the data sources, methodologies, and algorithms used, allowing instructional designers to verify the validity and appropriateness of the generated materials. This level of transparency not only enhances trust in AI systems but also empowers designers to make informed decisions and adjustments.

Opportunities and Challenges for GIFT

GIFT was originally created to economize the development of ITSs by implementing architectural standards including learner, domain, and pedagogical model frameworks as well as a user interface and defined data/information flow among those components. By constructing GIFT as a MOSA, any of these components could be re-used and/or improved upon optimizing the return on investment of these components and simplifying the creation of a new ITS (Robson & Barr, 2013).

Thus, GIFT's MOSA design is fundamentally about improving efficiency by maximizing re-use of many components of an ITS. At a minimum, a new ITS could be created with GIFT by simply creating a domain module and using existing architectural elements including pedagogical and learner modules that already exist. However, in practice, even the creation of a new domain module can quickly become a complex, time-consuming task. At a minimum, authors must create the core training content, assessments to gauge learner comprehension, and remedial content for learners who need it.

Remedial content must be created and linked to a pedagogical approach. GIFT's Engine for Management of Adaptive Content (EMAP) can deliver remedial content using both the Component Display Theory

(CDT) which includes at least two types of expository and two types of inquisitory content for each concept, and the Interactive, Constructive, Active, Passive (ICAP) framework with at least four types of content per concept (Goldberg et al., 2015; Sottolare et al., 2019). Regarding scalability, as noted above, LLMs are well suited to generate training content. With the right prompt engineering, many commercial LLMs can deliver useful multimedia content on most any topic. GIFT does not currently have any tools to automate a workflow starting with prompt generation and ending with integration of the final training media into a GIFT course. But it is likely that someone who is knowledgeable on course authoring in GIFT could save significant time by leveraging an LLM to generate content, assessments, and remediation for a new GIFT course. Future work will be needed to refine and automate such a workflow in GIFT. Ideally this workflow would guide prompt generation for the core and remedial content and would also add needed metadata to the content so that it could be leveraged by the desired pedagogical framework.

In terms of efficiency, LLMs have been shown to provide a means to rapidly refine GIFT's pedagogical framework. GIFT supports the use of static rules and decision trees to guide the delivery of training and remediation, but creation of these rules is both time-consuming and assumes that one can know in advance all possible learner states and reactions to the learning environment and the optimal tutoring response to any given learner state. While this may be possible for some very constrained learning domains, it is not a viable approach if one hopes to build a tutor with a broad capability to train across nearly all domains of knowledge and skill (Nye et al., 2019).

It was always recognized that for an intelligent tutor like GIFT to be most effective, it would need to behave more like an agent, learning from prior interactions with learners to improve its own policies and optimize its pedagogical framework over time (Sottolare et al., 2019). The emergence of powerful LLMs and trainable neural networks therefore makes this vision of GIFT, as an intelligent agent, much more viable.

An example of the application of machine learning to optimize pedagogy based on the learner's profile involves learners taking a course on counterinsurgency using the ICAP pedagogical framework (Spain et al., 2019). By observing learner performance following various types of feedback, reinforcement learning could be applied to train policies across a range of different learner profiles. Although the generalizability of these policies for other training domains has yet to be established, using this self-optimizing approach makes it possible to create generalizable policies in real-time as more training content is created and delivered through GIFT. So, in this regard, machine learning offers a clear path to improve the scalability of pedagogical policies that drive GIFT to tailor training to the needs of the learner.

With regard to quality control, as noted, the GIFT architecture supports the import, collection, modification, and export of learner data. GIFT does not operate in isolation, but relies on a larger learning ecosystem, known as the Total Learning Architecture (TLA), to access learner records and competency data used to tailor training to the individual or unit and then following a training event, export new assessments to update those learner records and competency frameworks (Goodwin et al., 2017). The data standard adopted by the TLA is known as xAPI (Experience API), an open-standard, human-readable, statement of learner activity that is enriched with contextual information to enable accurate interpretation. Smooth operation both among the components of GIFT and between GIFT and the larger TLA requires the consistent application of standards to all xAPI statements and other messages. If measurement criteria are inconsistent across the training applications or their components, or if records describing performance are not standardized, then a considerable amount of noise or error is going to degrade the tracking and management of competencies in the TLA. As more training systems are integrated into the TLA, maintaining end-to-end standards for performance assessment and reporting will become a significant challenge. Although this challenge has yet to be formally addressed in any published research report, it is likely that LLMs will be a valuable tool to explore large learner record stores to find inconsistencies in performance assessment or reporting across related xAPI records.

Another way in which LLMs can improve quality control is in standardizing assessment of complex behaviors like team communication during collective training events. GIFT's architecture has been developed to support team/collective training. Team training entails not only instructing and monitoring individual team members but also interactions among team members and assessing such team constructs as leadership, cooperation, and shared mental models. Developing an AI-based agent capable of team training is another significant challenge (Sottolare et al., 2018).

One of the biggest challenges for team training is the analysis of verbal communication among team members. Communication in a team context is critical for coordination, decision-making, support, and leadership (Sottolare et al., 2018). Analyzing team communication involves the classification of speech acts. For example, determining whether a particular speech act is a request, a command, or a sharing of information makes it possible to diagnose patterns of communication within and across teams and to link those patterns to the quality of team performance. Manually classifying team communications is a time-consuming process because the communication has to be played back in real time for expert observers to do classification.

Research has shown that LLMs are capable of correctly classifying speech acts among members of an infantry squad during a training event approximately 78% of the time (Spain et al., 2025). This level of performance was achieved through iterative prompt engineering within the investigators' team communication analysis framework. Although there is still room for improvement in the accuracy of the classification, even a classification rate of 78% can still provide significant time savings to expert human raters. The incorporation of such tools into GIFT is an ongoing process and this application of LLMs represents a very promising approach to standardizing assessment of complex team processes through automation.

In summary, AI tools like LLMs have the potential to improve scalability, efficiency, and quality control for GIFT and indeed evidence of this already exists in the case of refining GIFT's pedagogical logic and for classifying speech acts in collective training (Spain et al., 2019; Spain et al., 2025). Opportunities still exist to explore the use of these tools for things like generating all the content needed to create a new domain knowledge file, to provide a means of enforcing community of learning standards for assessing performance, and to create tools that can assess complex behaviors against those standards.

Use Cases and Applications

Generative AI is revolutionizing the landscape of ISD by introducing advanced tools and methodologies that enhance training readiness, scalability, and quality. Its application in real-world scenarios demonstrates its transformative potential across various contexts, from military training to industry-specific solutions. This section highlights notable use cases and applications, focusing on the Navy's RRL initiative, the "NAUTICAL for GIFT" project, and additional examples from military and industry environments.

Navy RRL: Integrating AI to Support Readiness and Training

The Navy's RRL initiative exemplifies how Generative AI can be harnessed to support readiness and training. This program aims to align training delivery with operational needs, ensuring that Sailors are equipped with the right skills at the right time. By integrating AI-driven tools, the RRL initiative optimizes the design and delivery of instructional content, making it both timely and relevant.

Generative AI enables the Navy to analyze vast datasets, such as performance metrics and operational requirements, to analyze tasks, learning outcomes, and instructional media to provide ISDs with training recommendations to identify and mitigate skill gaps and tailor training programs accordingly. This data-

driven approach ensures that training materials are not only accurate but also aligned with the evolving demands of naval operations. Furthermore, AI streamlines administrative tasks, such as scheduling and resource allocation, allowing instructors and trainees to focus on mission-critical activities. By bridging the gap between training and operational readiness, the RRL initiative highlights the strategic value of Generative AI in enhancing military preparedness.

NAUTICAL for GIFT: Aligning Generative AI with Intelligent Tutoring Systems

The "NAUTICAL for GIFT" project represents a pioneering effort to integrate Generative AI with ITSs. GIFT could leverage AI to deliver personalized and adaptive training experiences. Generative AI expands the capabilities of GIFT, enabling it to create dynamic and context-specific instructional scenarios.

One of the key innovations of this project is the use of Generative AI to automate the development of training scenarios that are tailored to individual learner needs. For example, Generative AI can generate role-specific scenarios for learners in different job functions, ensuring that the training is both relevant and engaging. Additionally, the system employs AI to monitor learner performance in real time, providing adaptive feedback and adjusting the difficulty of training tasks as needed. This integration of Generative AI with ITS underscores the potential of AI to enhance the effectiveness and scalability of personalized learning solutions.

Additional Examples from Military and Industry Contexts

Beyond the Navy RRL and NAUTICAL for GIFT, Generative AI is making significant inroads into other military and industry applications. In the military domain, AI-powered training solutions are being used to prepare personnel for complex scenarios, such as cybersecurity threats and disaster response operations. These systems leverage AI to simulate realistic environments, allowing trainees to practice decision-making and problem-solving skills in a controlled setting.

In the private sector, industries such as healthcare, manufacturing, and energy are adopting Generative AI to address their unique training challenges. For instance, in healthcare, AI is being used to develop interactive training modules for medical procedures and patient care. These modules incorporate real-time feedback and adaptive learning paths, ensuring that healthcare professionals gain the skills they need to provide high-quality care. Similarly, in the manufacturing sector, Generative AI supports the creation of safety training programs that are customized for different equipment types and operational settings. These examples illustrate the versatility of Generative AI in addressing various training needs across sectors. By automating the design and delivery of instructional content, AI not only reduces the time and cost associated with training development but also ensures that the materials are highly relevant and impactful.

The application of Generative AI in ISD showcases its potential to transform training and education. Through initiatives like the Navy RRL and NAUTICAL for GIFT, as well as broader industry applications, AI-driven tools are enabling organizations to achieve new levels of efficiency, scalability, and quality. These use cases provide a glimpse into the future of instructional design, where AI serves as a critical enabler of innovative and effective training solutions.

Future Perspectives

As Generative AI continues to redefine the boundaries of ISD and training, it brings both opportunities and challenges that will shape its future trajectory. On the one hand, AI offers unparalleled capabilities to scale training programs, enhance efficiency, and ensure quality. On the other, it raises important ethical considerations and security concerns that must be addressed to fully realize its potential.

In most ways, GIFT is well-positioned to take advantage of the potential of the latest generation of AI tools. Although GIFT was designed with scalability in mind, significant work is still needed to build new courses in GIFT. This challenge has not lessened as GIFT has evolved to support complex collective training exercises and to operate within the TLA supporting live and virtual experiential learning.

The potential of LLMs to generate curated multimedia content at the behest of a well-constructed prompt offers enormous potential for GIFT to realize the dream of being a truly scalable intelligent tutor. Of course, GIFT does much more than simply deliver content. By understanding the learner as represented in records of their past performance, experiences, and formal training, GIFT strives to build tailored learning experiences to accelerate comprehension and performance of individuals and teams.

Assessing performance in these training contexts is challenging. Even expert human observers, have a limited ability to observe large numbers of trainees at once, whether in a collective training event or with many individual trainees in multiple lanes. Both from the standpoint of quality control and efficiency, there is a need to be able to automate as many assessments as possible. As noted above, LLMs and other forms of AI are already proving to be instrumental in automating and standardizing the assessment of complex behaviors of individuals and teams in the context of training.

The evolving capabilities of AI present immense opportunities. Advances in explainability, adaptive learning, and integration with emerging technologies such as augmented reality (AR) and virtual reality (VR) are likely to expand the horizons of what Generative AI can achieve in ISD. However, as important as automated assessment is, it should always support an expert observer/trainer. The human expert provides the validation of all assessments and feedback and ensures that the automated assessments can be trusted and are unbiased.

The contributions of Generative AI to ISD—scalability, efficiency, and quality control—are transformative. By enabling scalable content generation, AI addresses the dynamic needs of modern learners. Its efficiency-enhancing features streamline workflows, allowing instructional designers to focus on creativity and innovation. Meanwhile, robust quality control mechanisms ensure that training programs remain consistent, accurate, and transparent.

The future of Generative AI in ISD is bright, marked by its potential to revolutionize the way organizations approach training and education. By navigating the challenges and leveraging the opportunities, AI can serve as a cornerstone of innovative, effective, and ethical instructional design. As technology evolves, its role in shaping the learning experiences of tomorrow will only continue to grow, offering unprecedented possibilities for education and training at scale.

References

- Goldberg, B., Hoffman, M. & Tarr, R. (2015). Authoring Instructional Management Logic in GIFT Using the Engine for Management of Adaptive Pedagogy (EMAP). *Design Recommendations for Intelligent Tutoring Systems: Authoring Tools and Expert Modeling Techniques*, 319-334
- Goodwin, G., Folsom-Kovarik, J., Johnson, A., Schatz, S., & Sottolare, R. (2017). Exploring Assessment Mechanisms in the Total Learning Architecture (TLA), Chapter 3 in *Design Recommendations for Intelligent Tutoring Systems: Assessment Methods*, 29-39
- Foltz, P. & Gilbert, S. (2023). Team Tutoring in Intelligent Tutoring Systems SWOT Analysis, Chapter 8 in *Design Recommendations for Intelligent Tutoring Systems: Strengths, Weaknesses, Opportunities, and Threats (SWOT) Analysis of Intelligent Tutoring Systems*, 91-100.

- McCarthy, J. & Sinatra, A.M. (2023). Authoring Tools in Intelligent Tutoring Systems SWOT Analysis, chapter 5 in *Design Recommendations for Intelligent Tutoring Systems: Strengths, Weaknesses, Opportunities, and Threats (SWOT) Analysis of Intelligent Tutoring Systems*, 63-69.
- Nye, B., Thaker, R., Surana, N., Auerbach, D., & Brawner, K. (2019). Toward Self-Improving Multi-Agent Tutoring Architectures: Progress and Future Directions, Chapter 5 in *Design Recommendations for Intelligent Tutoring Systems: Self Improving Systems*, 41-54.
- Robson, R. & Barr, A. (2013). Lowering the Barrier to Adoption of Intelligent Tutoring Systems through Standardization, Chapter 2 in *Design Recommendations for Intelligent Tutoring Systems: Learner Modeling*, 7-12.
- Shires, R. & McCormack, R. (2024) *Unleashing the Potential: Harnessing Large Language Models and Generative AI in Military and Industry Applications*. In Proceedings of the 2024 Interservice/Industry Training Simulation and Education Conference (I/ITSEC). Orlando, FL.
- Sottolare, R. A., Burke, C.S., Salas, E., Sinatra, A. M., Johnston, J. H., & Gilbert, S. B. (2018). Designing adaptive instruction for teams: A meta-analysis. *International Journal of Artificial Intelligence in Education*, 28(2), 225-264.
- Sottolare, R., Graesser, A., Hu, X., & Sinatra, A. (2018). Design Recommendations for Intelligent Tutoring Systems, Vol. 6: Team Tutoring.
- Sottolare, R., Sinatra, A., & Defalco, J. (2019). Considerations for Modeling Adaptive Instruction as a Complex Self-Improving System, Chapter 4 in *Design Recommendations for Intelligent Tutoring Systems: Self Improving Systems*, 29-40
- Spain, R., Min, W., Kumaran, V., Pande, J., Saville, J. & Lester, J. (2025). Applying Large Language Models to Enhance Dialogue and Communication Analysis for Adaptive Team Training. *International Journal of Artificial Intelligence in Education*. 10.1007/s40593-025-00479-5.
- Spain, R., Rowe, J., Goldberg, B., Pokorny, B. & Lester, J. (2019). *Enhancing Learning Outcomes Through Adaptive Remediation with GIFT*. In Proceedings of the 2019 Interservice/Industry Training Simulation and Education Conference (I/ITSEC). Orlando, FL.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention Is All You Need*. *Advances in Neural Information Processing Systems*, 30, 5998-6008. [Available on arXiv:1706.03762](https://arxiv.org/abs/1706.03762)

CHAPTER 12 – ARIZONA STATE UNIVERSITY CREATEAI PLATFORM

Ishrat Ahmed^{1,2}, Paul Alvarado², Siddharth Jain², Tracy Arner¹, Elizabeth Reilley², and Danielle S. McNamara¹

Learning Engineering Institute, Arizona State University¹; Enterprise Technology, Arizona State University²

Introduction to CreateAI: Bridging AI Innovation and Accessibility at Arizona State University (ASU)

The rapid evolution of artificial intelligence (AI) has created an unprecedented opportunity for innovation across industries, but its complexity often presents a steep learning curve for many (Roberts & Candi, 2024; Gunner, 2025). This highlights the need for user-friendly tools that bridge the gap between AI's potential and its practical application. Developed by the AI Acceleration unit at Arizona State University (ASU), the CreateAI Platform is breaking down these barriers, revolutionizing how AI is built, explored, and deployed. Designed to empower the ASU community, CreateAI is more than a technology platform—it is a transformative ecosystem that integrates pre-built models, streamlined workflows, and scalable infrastructure. By minimizing costs and complexity, CreateAI lowers barriers to entry, and empowers users to engage with AI. Its emphasis on accessibility, security, and adaptability ensures that innovation is accelerated while maintaining a user-friendly and secure environment.

CreateAI's mission is simple yet ambitious: to make cutting-edge AI tools and resources accessible to all, regardless of technical expertise. For individuals and teams without engineering or machine learning backgrounds, the platform demystifies AI by offering intuitive, easy-to-use tools. This shift represents a critical milestone in bridging the gap between groundbreaking research and practical, real-world applications of AI. The development of this platform requires investment in industry-leading language models, engineering and data science resources, as well as costs for technology infrastructure and computing.

One of CreateAI's defining features is its vendor-agnostic (e.g., OpenAI, Amazon Web Services, Google, Meta, etc.) cloud-independent design, ensuring access to best-in-class technologies regardless of their source. This flexibility allows the platform to evolve rapidly in response to new advancements, supporting ASU's pursuit of revolutionary, evidence-based educational experiences. By personalizing and enhancing these experiences, CreateAI aims to make education more cost-effective and accessible for all.

The CreateAI Platform is also built with security and adaptability at its forefront. Operating entirely within ASU's "walled garden" ensures that no data is shared with third parties, providing a secure and private environment for AI exploration. An additional feature that enhances security and adaptability is the modular design of the platform. Modular design refers to a system architecture where components are designed independently but can work seamlessly together, which is compatible with the GIFT (Generalized Intelligent Framework for Tutoring; gifttutoring.org) mission. This modular design enables the integration of emerging technologies, from vector databases to guardrail mechanisms for generative AI models. In a landscape where change is constant, CreateAI is designed to be agile and secure. From its foundational principles to real-world applications, CreateAI represents the future of AI development—accessible, secure, and ever-evolving.

The Role of Generative AI

Generative AI (GenAI) is transforming how we interact with technology by enabling the creation of intelligent, dynamic applications across various academic and non-academic domains. The CreateAI platform leverages the power of Generative AI and allows one to explore, prototype, and build AI powered applications for teaching, learning, workflow management, and research. It simplifies the process by handling complex backend tasks like managing large language model (LLM) providers, vector databases, authentication/access, and data logging. In this section, we describe the technical features of the CreateAI platform.

Technical Description of CreateAI Platform

The CreateAI platform is an Application Programming Interface (API) layer that allows internal ASU staff and developers to securely and safely access LLMs for a multitude of use cases. It is a versatile system that integrates robust architecture, advanced data management, and enhanced security features with seamless user interactions to support development of a variety of applications. The core functionalities that make the platform adaptable, scalable, and secure for evolving technological needs are described below:

System Architecture and Flexibility. These features ensure that the platform can adapt and grow with evolving technologies and user needs.

Modular Design: The CreateAI platform supports modular architecture which allows flexibility and scalability. Modularized design allows components within the platform to be swapped in and out as needed. For example, as models are updated or deprecated, they can be easily added or removed without disrupting the platform's overall functionality. Similarly, there may be several vector databases or database types required to store data. A modularized design allows for seamless exchange between the various databases, depending on the specific use case. This approach is also a key feature of GIFT, ensuring flexibility and interoperability.

Access to multiple LLMs: The CreateAI platform provides a large range of cutting-edge LLMs tailored to meet the needs of modern workflows. Currently, the platform provides access to over 30 different language models from leading vendors such as OpenAI, AWS, Google, and Microsoft. These models support interactions across multiple modalities, including text, audio, and image, offering versatile and comprehensive AI capabilities. The list of the models and their modalities across the CreateAI platform and its applications are provided in Table 1. This list is expected to expand, of course, as GenAI evolves.

Table 1: List of all the available LLMs, their providers, supported modalities, and availability across the CreateAI platform and its applications: (a) Model Comparison is a tool to compare different AI models; (b) MyAI Builder is a tool to build customized AI solutions, and (c) ASU GPT is a conversational AI (more about these applications are described in the next section).

Large Language Models (LLMs)			Enabled in CreateAI Platform & Products			
Provider	LLM Name	Modality	API Access	Model Comparison	MyAI Builder	ASU GPT
Amazon Bedrock	Titan G1 Express	Text	✓	✓	✓	
	Titan G1 Light	Text	✓	✓	✓	
	Claude 3 Sonnet	Text, Vision	✓	✓	✓	✓
	Claude 3.5 Sonnet	Text, Vision	✓	✓	✓	✓
	Claude 3 Haiku	Text, Vision	✓	✓	✓	✓
	Claude 3 Opus	Text, Vision	✓	✓	✓	✓
	Claude 2.1	Text	✓	✓	✓	✓
	Claude 2	Text	✓	✓	✓	✓
	Claude Instant (Claude v1.2)	Text	✓	✓	✓	✓
	Jurassic 2 Ultra	Text	✓			
	Jurassic 2 Mid	Text	✓			
	Command	Text	✓			
	Command Light	Text	✓			
	Llama 3.2 - 1B	Text	✓	✓	✓	✓
	Llama 3.1 - 3B	Text	✓	✓	✓	✓

	Llama 3.1 - 11B	Text, Vision	✓	✓	✓	✓
	Llama 3.1 - 90B	Text, Vision	✓	✓	✓	✓
	Llama 3.1 - 8B	Text	✓	✓	✓	✓
	Llama 3.1 - 70B	Text	✓	✓	✓	✓
	Llama 3.1 - 405 B	Text	✓	✓	✓	✓
	Mistral - 7B	Text	✓	✓	✓	✓
	Mixtral - Mistral 8 x 7B	Text	✓	✓	✓	✓
	Mistral Large	Text	✓	✓	✓	✓
	Nova Pro	Text, Vision	✓	✓	✓	✓
	Nova Lite	Text, Vision	✓	✓	✓	✓
	Nova Micro	Text	✓	✓	✓	✓
Reranker Models (Amazon Bedrock)	Cohere Reranker	Reranking	✓			
	Amazon Reranker	Reranking	✓			
Google AI	Bison v1 and v2 - API Only	Text	✓			
	Gemini Pro	Text	✓	✓	✓	✓
	Gemini Pro 1.5	Text, Vision, Audio	✓	✓	✓	✓
	Gemini 1.5 Flash	Text, Vision, Audio	✓	✓	✓	✓
	Gemini Pro Vision	Vision	✓			
	Imagen3	Image Gen	✓			

OpenAI	GPT 4o	Text, Vision	✓	✓	✓	✓
	GPT 4o mini	Text, Vision	✓	✓	✓	✓
	GPT o1	Text	✓	✓	✓	✓
	GPT o1 Mini	Text	✓	✓	✓	✓
	GPT 3.5	Text	✓	✓	✓	✓
	GPT 4	Text	✓	✓	✓	✓
	GPT 4 Turbo	Text, Vision	✓	✓	✓	✓
	Dalle 3	Image Gen	✓			
	Whisper	Audio	✓			
	OpenAI TTS	Speech Gen	✓			
Microsoft Azure	GPT 3.5	Text	✓			
	GPT 4	Text	✓			
Arizona State University	Llama 3 1B Fine Tuned	Text	✓			
	Llama 3 1B	Text	✓			

The variety of models available allows the users (e.g., faculties, staffs, researchers) to choose the best fit for specific tasks, from drafting documents, summarizing information, to creating intelligent chatbots and beyond. Offering multiple models within the platform provides significant benefits, including the flexibility to switch between different LLMs, allowing users to choose the best model based on their use case and budget. Unlike vendor AI solutions that rely on expensive annual license subscriptions, the platform operates on a per token pricing model, making it more economical, especially for users with limited or occasional usage needs.

Small Language Models (SLMs): The CreateAI platform supports flexibility by allowing users to host and customize their own small language models (SLMs), empowering them to tailor AI solutions to specific needs. Supporting SLMs with fewer than 3 billion parameters, the platform enables users to browse available models on Hugging Face (Wolf, 2019), provide the corresponding model URL, and integrate the selected model seamlessly into the CreateAI platform with all core functionalities enabled by default. This

process ensures users can leverage models best suited to their unique requirements, fostering innovation and personalization in AI applications.

Furthermore, the platform is developing a robust pipeline that allows users to upload their documents, select a base SLM, and request fine-tuning and deployment of the model. Once trained, the personalized model can be used directly within the platform or shared with others, facilitating collaborative and adaptable use cases. To demonstrate the platform's potential, two SLMs have been trained and deployed using ASU-specific data, providing accessible, real-worked examples for users. By empowering users to choose, fine-tune, and deploy their own models, CreateAI enhances flexibility, encourages experimentation, and simplifies AI development for various applications.

Data and Knowledge Management. These features enhance the organization, retrieval, and processing of data for better performance.

Retrieval-Augmented Generation: While LLMs offer significant benefits, such as generating human-like text, they are constrained by a fixed context window, i.e., they have a limit on how much information they can process one at a time. This limitation restricts their ability to retain details from earlier parts of a long conversation or document, making it challenging to manage complex or extended discussions effectively. The CreateAI platform leverages Retrieval-Augmented Generation (RAG) to address this limitation. RAG is a technique that enhances LLM capabilities by allowing them to retrieve relevant information from external sources such as user uploaded documents, rather than relying solely on their fixed training data. By dynamically accessing external data sources and retrieving relevant information through vector search operations (Lewis et al., 2020), RAG bridges the gap between limited context windows and the need for large-scale data integration. This approach ensures that the model can generate more accurate and contextually informed responses, even when dealing with extensive datasets or complex queries. This is particularly beneficial for users working with large datasets, enabling seamless integration and contextual understanding without sacrificing accuracy or scalability.

Vector Database: The CreateAI platform leverages advanced AWS OpenSearch vector databases to streamline and enhance the efficiency of AI application development. A vector database is a type of database designed to store, manage, and search high-dimensional vector embeddings, which are numerical representations of data such as text, images, or audio. Unlike traditional databases that rely on exact keyword matching, vector databases use mathematical similarity measures (e.g., cosine similarity) to find and retrieve data points that are most relevant to a given query.

This integration addresses common bottlenecks such as data ingestion (the process of collecting, processing, and storing incoming data efficiently) and query performance (the speed and efficiency with which a system retrieves relevant information from a database), which are critical challenges in building robust AI systems. The platform offers two options to cater to varying operational needs: OpenSearch Serverless and OpenSearch Cluster. OpenSearch Serverless provides a lightweight, scalable solution ideal for applications requiring flexibility, while OpenSearch Cluster delivers enhanced ingestion speeds, hybrid search capabilities (combining keyword-based and vector-based search for more accurate and relevant results), and advanced re-ranking features (intelligently reordering search results based on relevance, user behavior, or contextual factors). This dual approach ensures that users can tailor their solutions to specific requirements, optimizing high-speed data handling for real-time AI workloads. By integrating these powerful tools, the CreateAI platform empowers developers to build AI applications that are not only efficient but also scalable and responsive to dynamic demands.

Embeddings as a Service: The CreateAI platform provides embeddings as a service, enabling users to independently perform RAG on their own infrastructure. Embeddings as a service refers to a system that generates dense vector representations of data (e.g., text, images, or other modalities), which can be used

for tasks like similarity search, clustering, or retrieval. These embeddings are essentially vector representations of strings, and every LLM functions as an embedding model with an additional layer that converts the embedding back into text. This service supports multimodal and multilingual embeddings, allowing users to work seamlessly with various datasets, such as images and text in multiple languages. The platform allows users to leverage various open-source embeddings directly, offering flexibility and ease of use while empowering them to perform RAG more effectively, ensuring improved accuracy and relevance in their applications.

Interaction and Communication. These features enhance the user experience through seamless and effective interaction models:

Multiple File formats: The platform supports 36 different file formats, enabling users to upload and interact with a wide variety of content types. These formats include documents, spreadsheets, presentations, PDFs, and more ensuring compatibility with commonly used file types in educational and professional settings.

Voice to Voice Service: The platform includes a voice-to-voice service, enabling users to interact through speech. Spoken input is processed and responded to with natural-sounding voice output, enhancing accessibility and making the application more user-friendly for a variety of interaction scenarios. Users can leverage any language model available within the platform for voice-to-voice interactions, without being limited to models that natively support audio input or output. Additionally, the platform allows users to customize the voice and integrate RAG functionality, providing dynamic, personalized, and contextually relevant responses.

Safety and Security. These features are designed to protect data, ensure ethical usage, and maintain platform integrity.

Enhanced Security: One of the critical challenges in deploying LLMs is ensuring the security and privacy of user data. The CreateAI platform addresses this concern by implementing strict security measures, including agreements with LLM providers and cloud infrastructure partners that prohibit data from being used for model training. This ensures that sensitive information remains confined within the platform's cloud infrastructure. To further reinforce these efforts, the platform ensures AI security and user privacy compliance through continuous auditing of its partners and the platform itself, maintaining a robust and trustworthy framework for data protection.

Safety, Security, and Access control: Security and cost efficiency are critical considerations in the design of the CreateAI platform. To ensure secure access and usage, the platform employs multiple layers of authentication, including Single-Sign-On (SSO) and developer-level authentication mechanisms. These measures prevent unauthorized access and ensure that requests are verified, thus maintaining the platform's integrity. By imposing usage limits on LLMs, the platform also mitigates the risk of excessive computational expenses, promoting cost efficiency.

Ethical AI Engine: Ensuring ethical AI outputs is critical to building trust and promoting responsible use of AI technologies. We have developed Ethical AI Engine, a comprehensive framework for detecting potential harmful content in AI outputs, ensuring alignment with ethical standards (Liu, et al., 2024). Building on Stanford's HELM (Holistic Evaluation of Language Models) framework (Bommasani et al., 2023), which evaluates language models across metrics such as accuracy, robustness, fairness, and efficiency, the Ethical AI Engine expands on these principles to create a more versatile evaluation system. This suite of automated evaluation algorithms scores LLM-powered chatbots across multiple dimensions, including domain-specific accuracy, fairness, robustness, and responsible use in higher education. The ethical AI engine

represents a significant step forward in fostering ethical AI practices, ensuring that AI technologies are both effective and socially responsible.

Data Logging: The CreateAI platform's robust data logging capabilities further enhance its reliability and safety. By logging every request and employing an LLM for summarization and safety checks, CreateAI ensures that malicious or harmful queries can be flagged for review. This proactive approach to safety not only enhances user experience but also sets a precedent for responsible AI deployment. Data generated from applications within the platform allow for evaluation of usage patterns and reasons, providing valuable insights to optimize performance, improve user experience, and tailor AI solutions to better meet user needs. While vendor solutions often rely on external data storage, which depends on the vendor's ability to manage and transmit reliable data, CreateAI stores data internally for analytical purposes. This internal storage allows for integration with other contextual data, providing deeper insights to enhance safety and support the continuous improvement of the platform.

The CreateAI platform consists of features and capabilities that are a transformative step toward significantly lowering the barriers for non-technical developers to design, build, and deploy LLM-powered applications. The platform addresses key challenges related to security, scalability, and usability so that users need only bring their creativity to build LLM-powered solutions. The platform not only simplifies access to advanced AI tools but also maintains a strong focus on safety and ethics. Support of multimodal applications and high-speed, secure data retrieval makes it a versatile solution for a variety of different use cases, ranging from educational tools to enterprise applications.

CreateAI Platform Applications

ASU has created three different applications using this CreateAI Platform to demonstrate its versatility and adaptability as indicated in Table 1. These applications can be used by the ASU community (e.g., faculties, staffs, researchers, etc.). These include **Model Comparison**, which allows users to compare different AI models; **ASU GPT**, a conversational AI tailored to meet the varying needs of the ASU community, such as helping marketers generate materials, assisting faculty in refining or creating assignments, and enabling staff to summarize documents for clarity; and **MyAI Builder**, which empowers users to build their own customized AI solutions.

Model Comparison: The Model Comparison tool empowers the users to select multiple LLMs to compare their strengths and find the best fit for their particular use case. This is important because LLMs vary in their responses due to differences in each provider's approach to training methods, training data, parameters, etc. The users can choose up to 6 different models (see Figure 1), allowing them to submit a prompt and review differences in quality and format of responses, estimated cost, and speed (Figure 2).

The value of this tool lies in its ability to help users identify the model that best suits their specific tasks and then leverage that model with other CreateAI products, such as ASU GPT and MyAI Builder. For instance, some models are better at following instructions. Users can write a test prompt to assess each model's ability to follow instructions and evaluate their performance based on the output. Alternatively, users may rely on pre-existing benchmarks (standardized tests or datasets used to evaluate the performance of AI models) to select models; however, these benchmarks are often based on datasets that may not align with the user's unique needs or use case. This tool bridges the gap by enabling tailored model evaluation for more precise and effective AI application.

Figure 1: Model Comparison Interface: User can choose up to 6 different models from 30 models.

Figure 2: Model Comparison Interface: Selected Model Output of a prompt provided by a user

View Model Outputs

Compare and analyze model capabilities, speed, and interpretation.

Tell me about Arizona State University

<p>GPT 4 OpenAI</p> <p>Thunderbird School of Global Management. Each campus is distinctive and academically focused on certain aspects of the overall university mission.</p> <p>ASU offers a wide range of undergraduate and graduate degree programs across various disciplines. The university has several colleges including Barrett, The Honors College; College of Health Solutions; Herberger Institute for Design and the Arts; Ira A. Fulton</p> <p>Best match Tokens 378 Cost \$ 2.2680 36.0 s</p>	<p>Gemini Flash 1.5 Google DeepMind Best Cost</p> <p>Arizona State University (ASU) is a public research university located in Tempe, Arizona. ASU is a comprehensive university with four campuses in the Phoenix Metropolitan Area: Tempe, West Valley, Downtown Phoenix, and Polytechnic. ASU also offers courses and degrees through ASU Online and at the ASU Colleges at Lake Havasu City in western Arizona. The university is known for its commitment to innovation, its diverse student body, and its focus on real-world applications of knowledge. ASU is governed by the Arizona</p> <p>Best match Tokens 618 Cost \$ 0.0185 28.4 s</p>	<p>GPT 4o OpenAI Fastest</p> <p>Arizona State University (ASU) is a prominent public research university located in the Phoenix Metropolitan Area, Arizona. It is known for its innovative approach to education and research, offering a wide range of undergraduate, graduate, and professional programs across multiple campuses and online platforms.</p> <p>**Campuses and Locations:** - **Tempe Campus:** The main campus, known for its research and</p> <p>Best match Tokens 409 Cost \$ 0.6135 16.7 s</p>
<p>GPT 4o Mini OpenAI</p> <p>Arizona State University (ASU) is a public research university located in the Phoenix metropolitan area of Arizona. It is known for its innovative approach to education and its commitment to accessibility, research, and community engagement. Here are some key aspects of ASU:</p> <p>### Campuses ASU has four main campuses:</p> <p>Best match Tokens 543 Cost \$ 0.0326 26.6 s</p>	<p>O1 Mini OpenAI</p> <p>Arizona State University (ASU) is a prominent public research institution located in the Phoenix Metropolitan Area, Arizona. Recognized for its innovation, inclusivity, and extensive academic offerings, ASU serves a diverse student body across multiple campuses and online platforms.</p> <p>### **Campuses and Locations** ASU operates four main campuses within the Phoenix area:</p> <p>Best match Tokens 966 Cost \$ 0.0580 20.3 s</p>	<p>Nova Pro Amazon</p> <p>Certainly! Arizona State University (ASU) is a prominent public research university located in Arizona, known for its innovative approach to education and research. Here are some key points about ASU:</p> <p>### Overview - **Founded:** 1885 - **Type:** Public research university</p> <p>Best match Tokens 691 Cost \$ 0.2211 20.5 s</p>

ASU GPT: The second application is ASU GPT, an AI companion tailored for the university that allows users to upload internal ASU data and interact with it in a conversational setting leveraging multiple language models. This ensures that the AI companion provides tailored support for students, faculty, and staff regarding university-related queries. Users can choose to engage with any of the 30 advanced LLMs (see Table 1), ask questions, and explore ideas both safely and securely within ASU’s ecosystem. Interactions within this application are more conversational in nature, allowing users to engage in text-based conversations while also uploading and interacting with their own documents to find relevant information and receive personalized response.

In this application, users can also upload an image for text transcription and have voice-to-voice service enabled, providing enhanced accessibility and convenience for interaction needs. The key benefit of this platform is its ability to offer users access to multiple vendor solutions instead of relying on a single LLM provider. This feature provides users with greater flexibility and choice enabling the ASU community access to cutting-edge AI technology in an ever-changing landscape.

MyAI Builder: The third application is MyAI Builder which allows users at ASU to develop their own AI experiences and leverage the power of Generative AI by building custom AI-powered applications tailored to specific goals. The user can select from 32 models, configure a system prompt, leverage RAG, and develop a Generative AI chatbot that includes a user interface (see Figure 3a and 3b).

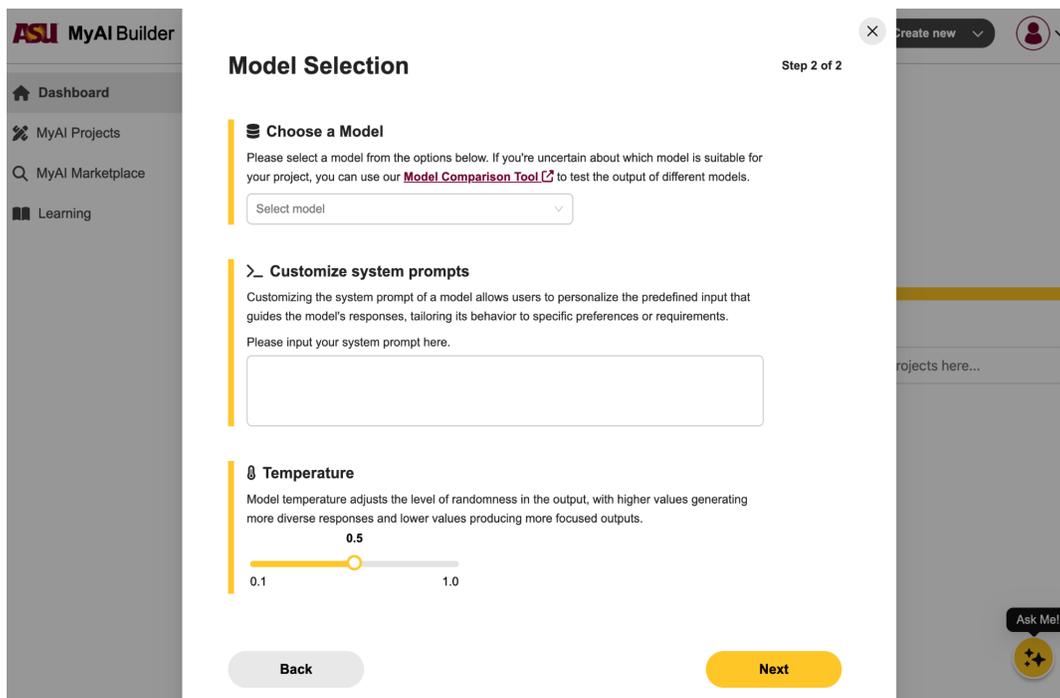
Figure 3a: MyAI Builder Interface. In this first interface, users can add a project name, and project description.

The screenshot displays the 'Create a project' interface within the MyAI Builder application. The interface is a modal window titled 'Create a project' with a close button (X) in the top right corner. It is labeled 'Step 1 of 2'. The form contains three main sections, each with a title and a brief instruction:

- Project name:** Includes a text input field and the instruction 'Create a name for your project'.
- Project description:** Includes a larger text area for input, the instruction 'Please write a brief description of your project', and a character count '0/240'.
- Use cases:** Includes a dropdown menu and the instruction 'Select your use case'.

A yellow 'Next' button is positioned at the bottom right of the form. The background shows a sidebar with navigation options: 'Dashboard', 'MyAI Projects', 'MyAI Marketplace', and 'Learning'. The top navigation bar includes a 'Create new' dropdown and a user profile icon. There is also an 'Ask Me!' button and a plus sign icon in the bottom right corner of the background.

Figure 3b. MyAI Builder Interface. In this second interface users can choose a model, and provide system prompts to build their own AI solution.



In addition to these, there are multiple settings a user can adjust. For instance, a user can choose whether to use RAG vs System Prompt for the application depending on their use case. In the case of RAG, users can leverage their source documents, in multiple file formats (e.g., docx, pptx, csv, etc.) and ASU-specific websites, to build a custom GPT. On the other hand, users can choose system prompts where they can use the general knowledge of LLMs but provide a set of rules such that the custom GPT behaves in a particular way. Furthermore, the integration of voice-to-voice functionality in MyAI Builder enhances accessibility, enables hands-free and natural interaction, supports faster and multilingual communication, and improves real-time user engagement across a variety of scenarios. All these features within MyAI Builder make it accessible to users without coding knowledge, empowering educators and professionals to create AI-driven solutions without relying on technical expertise.

Overall, the CreateAI platform offers a comprehensive suite of applications that cater to varying needs within the ASU community, enhancing the way users interact with AI technology. The Model Comparison app provides a valuable tool for selecting the most suitable LLM for specific tasks, ensuring optimal performance and efficiency. ASU GPT serves as a robust AI companion, facilitating secure and insightful interactions with advanced language models, while MyAI Builder empowers users to create AI solutions tailored to their unique objectives. Together, these applications not only simplify access to cutting-edge AI capabilities but also foster innovation and collaboration.

Notable Use Case: Building an ITS using CreateAI Platform

The CreateAI platform offers a unique opportunity to create custom GPT models tailored to specific applications. One innovative use case of this could be the development of an Intelligent Tutoring System (ITS). An ITS is a computer-based system designed to provide personalized feedback and adapt to a learner's performance, replicating the effectiveness of one-on-one tutoring (Graesser et al., 2012). The development of such systems has been both time-consuming and code-intensive, requiring significant expertise in programming and AI development (Aleven et al., 2009; Koedinger et al., 2004). To alleviate this, authoring tools for ITS development already exist (Aleven et al., 2009; Sottolare et al., 2015) to allow users to develop their own ITSs. However, authors who successfully use authoring tools typically require specialized skills in programming and AI, which limits these tools from being accessible to a broader audience. With the CreateAI platform, the process of building an ITS becomes significantly more streamlined and accessible. Using the MyAI Builder application, educators and developers can create custom ITSs by uploading relevant documents as the knowledge base. The platform supports prompt engineering to define how the system interacts with learners, enabling it to evaluate responses and adapt accordingly at every turn. This eliminates the need for complex coding, instead using plain language, thus making the development process more accessible and user-friendly.

Moreover, the no-code capabilities of MyAI Builder simplify the creation of advanced tutoring systems, allowing educators, trainers, and subject matter experts to focus on content and pedagogy rather than technical implementation. By leveraging CreateAI, users can design ITSs that deliver personalized, interactive, and adaptive learning experiences across various domains. This innovation not only reduces the barriers to developing ITSs but also accelerates the adoption of AI-driven education tools to enhance learning outcomes.

Example AI Solutions in Education Domain

In addition to enabling the development of ITSs, the CreateAI platform offers versatile solutions for a wide range of educational use cases. Its flexibility and adaptability make it a powerful tool for addressing needs in teaching, learning, and academic innovation. We highlight a few example use cases to showcase the capabilities of the CreateAI platform.

One prominent application of the CreateAI platform is the development of tools designed to enhance the ideation process by generating creative and varying ideas. By utilizing MyAIBuilder, bots can be developed to empower students in exploring novel concepts and perspectives, making the bots particularly valuable for tasks like brainstorming ideas for essays, projects, or research. Through effective prompt engineering, these tools can act as a thought partner, assisting students in generating outlines, suggesting unique perspectives, and refining arguments. This functionality can enable users to overcome mental blocks, unlock new possibilities, and efficiently iterate on their ideas. The resulting iterative interaction highlights the transformative potential of the CreateAI platform in driving creativity and productivity.

Additionally, The CreateAI platform may enable the creation of AI mentors that guide students through coursework challenges while fostering the development of practical and professional skills. These mentors may provide personalized support, addressing questions with contextually relevant advice, and real-time feedback to build students' confidence in articulating ideas and solutions. By serving as collaborative design partners, the AI mentors can introduce students to the practical use of AI tools, helping them enhance creativity and efficiency in tasks such as brainstorming, refining ideas, and solving complex problems. This framework may emphasize AI as a complement to human expertise, offering students hands-on experience with AI in ways that align with professional learning design practices.

Finally, the CreateAI platform enables the development of AI-driven tutors that deliver personalized and adaptive learning experiences. Using tools like MyAI Builder, educators can easily design tutors by uploading course-specific materials, such as lecture notes, multimedia content, and assessments, to create a comprehensive knowledge base. Through advanced prompt engineering, the tutors can be programmed to interact with students in a contextually relevant and engaging manner, guiding them through complex topics, answering questions, and providing targeted feedback to address misunderstandings. The platform's ability to analyze student performance allows the tutors to dynamically adjust their strategies, ensuring individualized learning experiences. By supporting multimodal inputs such as text or image, the platform enables the creation of tutors tailored to various fields, from STEM to the arts. This adaptability empowers educators to design AI tutors that enhance learning and foster student success. Additionally, data logging and adherence to established data privacy standards within ASU's system streamline the process of analyzing chatbot effectiveness in the classroom. Since data is generated and stored internally, there is no need for crosswalks, external data integration, or coordination with multiple teams, reducing complexity and privacy concerns.

These are just a few examples of what the CreateAI platform can achieve, showcasing its versatility and transformative potential. By enabling users to create personalized tutors, collaborative mentors, or innovative tools, the platform empowers educators, students, and professionals to unlock new possibilities in learning and problem-solving. Its flexibility ensures that it can adapt to a wide range of needs, from education to creative industries, highlighting its value as a dynamic and accessible AI-driven solution.

Recommendations for GIFT Framework

The Generalized Intelligent Framework for Tutoring (GIFT) is a modular framework designed to facilitate the development of adaptive and personalized ITSs, revolutionizing the way educational content is delivered and tailored to individual learners (Sottolare et al., 2013). Our CreateAI platform holds immense potential to advance the design of future ITSs in many ways. For example, potential approaches that could be implemented within GIFT are below.

1. *Leverage No-Code Development Tools:* GIFT could incorporate no-code development capabilities similar to the CreateAI platform's MyAI Builder. By enabling educators and subject matter experts to create ITSs without extensive programming knowledge, GIFT can significantly lower the barriers to adoption. Providing a user-friendly interface for uploading knowledge bases and configuring tutoring logic would empower non-technical users to design customized ITS solutions.
2. *Enable Prompt Engineering for Adaptive Learning:* Integrating prompt engineering tools into GIFT would allow users to define how the system interacts with learners and adapts to their performance. This feature could enhance the flexibility of tutoring systems, enabling tailored feedback and dynamic evaluation of learner responses, thereby replicating the effectiveness of one-on-one tutoring.
3. *Streamline Knowledge Base Integration:* GIFT should support easy integration of course-specific materials, such as lecture notes, textbooks, and multimedia content. Automated processing of these materials into a structured knowledge base would reduce setup time and enable ITSs to deliver more contextually accurate responses.
4. *Facilitate Rapid Prototyping and Deployment:* Incorporating features that allow rapid prototyping and deployment of ITSs can help educators and developers test and iterate their systems quickly. For example, a pipeline for fine-tuning and deploying small language models on user-provided datasets could enhance the scalability and versatility of GIFT.

5. *Promote Collaboration Among Educators and Developers:* GIFT could incorporate features that allow educators and developers to collaborate on ITS creation, sharing templates, resources, and best practices. This collaborative environment would foster innovation and accelerate the adoption of AI-driven tutoring systems.

By implementing these recommendations, GIFT can establish itself as a leading framework for developing flexible, accessible, and effective ITSs.

Discussion

CreateAI is a versatile platform designed to empower all users including faculties, staff, and researchers, regardless of their technical expertise, by offering tools to design, deploy, and customize AI-driven solutions. For technical users, the platform provides robust API connections, enabling seamless integration with advanced language models and cutting-edge technology available on the platform. For non-technical users, products such as MyAI Builder and ASU GPT provide intuitive interfaces that allow them to achieve their goals using natural language, without needing to worry about complexities such as security, data logging, and user interface design: these features are seamlessly integrated into the product. By catering to varying user needs, CreateAI fosters an environment where anyone can engage with and benefit from AI technology. This platform is being beta tested by faculty, staff, and researchers based on their unique interests and need. The platform allows for rapid prototyping of ideas leveraging leading Generative AI language models. While initial adoption has begun, a comprehensive analysis of its current usage and impact on learning or research productivity is still underway.

The effectiveness of the CreateAI platform in educational settings is presumably related to rooted in three critical factors: access to reliable course-specific data, collaboration with subject matter experts, and iterative improvements driven by user feedback. These elements form the foundation for creating AI systems that meaningfully support students in their learning journey while aligning with educational goals and standards.

Reliable Course-Specific Data

The CreateAI platform leverages reliable and up-to-date course materials, including lecture transcripts, text documents, images, and videos, to create tailored AI solutions. By integrating these resources, the platform ensures that AI responses are contextually accurate and relevant to the specific course content. For instance, in technical courses such as biology or computer science, where specialized tools and terminologies are prevalent, CreateAI can leverage faculty approved information to guide students. This platform capability enables the creation of virtual assistants that complement classroom instruction and aid students in mastering complex concepts. Documents do not have to be centralized into one technology or repository - they can be stored dynamically in many different places and the platform enables the ability to access the information within an AI system.

Collaboration with Faculty and Subject Matter Experts

Faculty involvement is essential in shaping how AI supports students in the classroom. CreateAI emphasizes the importance of involving faculty and subject matter experts in the development process by offering intuitive tools (e.g., MyAI Builder) that make AI customization accessible, even for those without technical expertise. The platform allows instructors to engage in prompt engineering through user-friendly interfaces, enabling them to shape how AI interacts with students without requiring programming skills.

This facility empowers educators to tailor the AI's responses to align with their teaching styles, whether by providing direct answers or guiding students toward solutions that promote critical thinking. Through MyAI Builder, CreateAI significantly lowers the barrier to entry for educators, encouraging wider adoption of AI tools in classrooms. Instructors can focus on pedagogical objectives rather than technical implementation, ensuring that the AI system delivers responses that are both contextually accurate and academically aligned. This collaborative approach bridges AI capabilities with instructional expertise, enhancing the platform's flexibility and usability while enriching the educational experience for learners.

User Feedback for Iterative Improvement

The MyAI Builder application developed within the CreateAI platform is designed to incorporate user feedback as a cornerstone of its development process. Student input helps identify common learning challenges and areas where AI can provide meaningful support. For example, in courses requiring proficiency in tools like Excel, the platform enables the creation of AI chatbots that guide students through tool-specific tasks, tailored to their assignments or field of study. Advanced features, such as analyzing screenshots of student work to provide contextualized instructions, demonstrate the platform's ability to address real-world student needs effectively. Iterative improvements driven by feedback ensure that CreateAI solutions remain intuitive, practical, and aligned with user expectations.

The CreateAI platform excels in integrating reliable data, faculty collaboration, and student feedback into its AI development process. These factors are expected to not only improve the quality of AI interactions but also foster trust and adoption among students and educators. As AI continues to evolve, maintaining a collaborative and feedback-driven approach will be key to creating tools that enrich the learning experience and support academic needs. Additionally, the forthcoming capability to host custom embedding models opens new research avenues into the integration of domain-specific knowledge with advanced AI systems.

Conclusions

The CreateAI platform developed by ASU represents a significant advancement in AI infrastructure, enabling robust and secure development of AI-powered applications. Its suite of features addresses key challenges in deploying LLMs while adhering to high standards of security, ethics, and performance. The CreateAI platform exemplifies the potential of AI infrastructure to balance performance, security, and ethical considerations. Its innovative features not only address existing challenges in AI application development but also pave the way for more accessible and responsible use of Generative AI technologies. As the platform evolves, its contributions to AI research and application development are likely to expand, further solidifying its role as a leader in the field. Future work could explore the scalability of these features in larger, more diverse environments and assess their impact on broader AI adoption.

Acknowledgements

We would like to express our gratitude to the individuals whose support and contributions made this work possible. We acknowledge the insights and guidance from our colleagues and thought leaders, whose vision and guidance were pivotal in shaping this work. Special thanks to Arizona State University for providing the resources and infrastructure for the development and implementation of the CreateAI platform. We are grateful to the faculty, staff, and students who participated in the developing, testing and refinement of the platform, offering valuable feedback that helped shape its design and features. The work reported here was supported the ASU Learning Engineering Institute and the Institute of Education Sciences, U.S. Department of Education, through Grants R305N210041 and R305T240035 to Arizona State University. The opinions

expressed are those of the authors and do not represent views of ASU, the Institute of Education Sciences, or the U.S. Department of Education.

References

- Aleven, V., McLaren, B. M., Sewall, J., & Koedinger, K. R. (2009). A new paradigm for intelligent tutoring systems: Example-tracing tutors. *International Journal of Artificial Intelligence in Education*, *19*(2), 105-154.
- Bommasani, R., Zhang, D., Lee, T., & Liang, P. (2023). Improving transparency in AI language models: A holistic evaluation. *HAI Policy & Society*.
- Graesser, A. C., Conley, M. W., & Olney, A. (2012). Intelligent tutoring systems. In K. R. Harris, S. Graham, T. Urdan, A. G. Bus, S. Major, & H. L. Swanson (Eds.), *APA Handbook of Educational Psychology, Vol. 3. Application to learning and teaching* (pp. 451–473). American Psychological Association. <https://doi.org/10.1037/13275-018>
- Gunner, R. (January 8, 2025) Survey: Generative AI makes tasks simple, but developing that AI is anything but. *IBM Newsroom*. Retrieved Jan 15, 2025 from <https://newsroom.ibm.com/blog-survey-generative-ai-makes-tasks-simple.-but-developing-that-ai-is-anything-but/#:~:text=This%20speaks%20to%20the%20skills,same%20traits%20are%20the%20rarest>
- Koedinger, K. R., Aleven, V., Heffernan, N., McLaren, B., & Hockenberry, M. (2004). Opening the door to non-programmers: Authoring intelligent tutor behavior by demonstration. In *Intelligent Tutoring Systems: 7th International Conference, ITS 2004, Proceedings* (pp. 162-174). Springer Berlin Heidelberg.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474.
- Liu, S., Shourie, V., & Ahmed, I. (June, 2024). The Higher Education Language Model Multidimensional Multimodal Evaluation Framework: Evaluating LLM-powered chatbots in higher education. [White paper]. Retrieved January 13, 2025, from https://issuu.com/asu_uto/docs/highered_language_model_evaluation_framework
- Roberts, D. L., & Candi, M. (2024). Artificial intelligence and innovation management: Charting the evolving landscape. *Technovation*, *136*, 103081.
- Sottolare, R. A., Brawner, K. W., Goldberg, B. S., & Holden, H. K. (2013). The generalized intelligent framework for tutoring (GIFT). In *Fundamental issues in defense training and simulation* (pp. 223-233). CRC Press.
- Sottolare, R., Graesser, A., Hu, X., & Brawner, K. (Eds.). (2015). Design recommendations for intelligent tutoring systems: Authoring tools and expert modeling techniques.
- Wolf, T. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*. <https://doi.org/10.48550/arXiv.1910.03771>

CHAPTER 13 – USING GENERATIVE AI TO GUIDE EXPERIENTIAL LEARNING

Lisa N. Townsend¹, Paige M. Lawton¹, Robert A. Sottolare², Randall Spain¹, Richard N. Landers³, Benjamin Goldberg¹, and Andy Smith⁴

US Army Combat Capabilities Development Command (DEVCOM) Soldier Center¹;
SoarTech, LLC, an Accelint Company²; University of Minnesota³; North Carolina State University⁴

Introduction

When it comes to the traditional education and training model, it is often focused on the learning of declarative and procedural knowledge, and often this is not an optimal method for increasing learning transfer and long-term human competence, or tacit knowledge. Tacit knowledge enables performing a task with hyper-cognitive reasoning, decision-making, and virtually automated behaviors (Huskey, 2022). It can only be constructed through experiences and learning through various forms of feedback, coaching, and reflection. One challenge in modern institutional training and education environments is the low engagement and percentage of retention and transfer of taught material, creating low return on training and education investment. While traditional instructional methods can achieve satisfactory levels, they simply are not adept enough or aligned with how humans naturally learn or enable maximum competence in tasks and traits they need to excel. Yet in both the civilian university setting, and in the Department of Defense (DOD), most institutional training courses still employ a traditional didactic instructional model that has been used for centuries. Meanwhile, modern learning science and practical research show that more active learning processes supported with technology can make institutional instruction more engaging, create longer retention of learning, and potentially increase the transfer of instructed knowledge and skills.

This requires a new instructional model termed competency-based experiential learning (CBEL) (Owens, 2023). This model has been tested in applied research for the US Navy and now, for the last several years, has been heavily researched for the US Army (Goldberg, et al., 2021) in support of the US Army Learning Concept 2020-2040 (US Army, 2017) and the Synthetic Training Environment (STE) program.

The CBEL model also became a learning engineering project that is experimenting with this application into a military institutional learning environment that currently incorporates a traditional instructional method. As learning engineering is a process that employs advancements in science, technologies, and best practices related to human learning for advancing the fundamental ability to survive and thrive in the modern world, applying CBEL attributes within this environment presents a chance to understand the benefits and address the challenges.

An opportunity to conduct this learning engineering effort was provided by the US Army's United States Military Academy (USMA) Department of Military Instruction (DMI). The DMI teaches military science (MS) across three-courses over a four-year program. They already integrate the curriculum program with a well-equipped, state-of-the-art, Simulation Center, providing an opportunity for data collection and a potential testbed. The focus area for CBEL was determined to be Land Navigation (LandNav), as it is a topic that Cadets are introduced to during summer training and then it is practiced and applied throughout the remaining four years of school in a crawl, walk, run approach, lending itself to both numerous opportunities for injecting and enhancing experiential learning (EL) and for longitudinal assessment. LandNav is also an area where current metrics show Cadets struggling on the complex tasks, and Instructors spending time ensuring progress and understanding of the associated tasks – tasks required for endurance and success.

Incorporating Generative Artificial Intelligence (AI) aspects could be beneficial and influential in this application. Current AI considerations for this effort include the use of Adaptive Instructional Systems (AISs) technology to inform modifications of the synthetic environment to match the level of competence as a trainee progresses towards mastery and also collecting low-level data that is both saved and used by AI built into the AIS for the tasks being trained. Understanding how colleagues across academia and industry are incorporating Generative AI (GenAI) can provide additional valuable insight and lessons learned on how it could be manipulated to direct and influence experiential learning.

The Role of Generative AI

Using GenAI to Create Novel Narratives for Learning

A promising area of research demonstrating the potential for GenAI to direct and influence experiential learning is through narrative centered learning. Narrative-centered learning describes a family of research on how to utilize interactive stories, often implemented in 3D game environments, to actively engage students in situated problem solving. Though relatively new, GenAI has already shown its ability to have a tremendous impact in several aspects of this research, including scenario generation, adaptive scaffolding, and natural language dialogue analysis.

A key challenge for both narrative-centered learning and experiential learning environments is the design and creation of novel learning activities that incorporate the desired learning objectives while not being repetitive to users. GenAI, especially large language models (LLMs), have been shown to excel at generating compelling narratives, including plot events, character trajectories, and end goals, based on short, natural language input story arcs. Early systems like SceneCraft (Kumaran et al., 2023) utilized this technology to enable scenario designers to generate, and refine, short narratives which could then be dynamically embodied in a 3D game environment. Building on this work, NarrativeGenie (Kumaran et al., 2024) was developed utilizing advances in OpenAI's GPT4 model. NarrativeGenie is an interactive narrative generation framework that is able to generate sequences of narrative events, that can be customized and aligned with the author's intent both at a granular event level, as well as at a higher-level narrative event graph representation. After refinement, the generated narrative can then be embodied in a 3D game environment using pre-defined character and environment assets, dynamically generating all non-player character dialogues, emotes, and gestures. While currently narrative focused, it is not hard to imagine leveraging these same technologies to generate rich, customizable scenarios for experiential learning (EL) applications such as the LandNav course discussed earlier.

GenAI has also grown in importance as a tool for analyzing data collected during EL activities. LLM-based frameworks have shown particular promise for automated understanding of natural language data, particularly conversation data. Dialogue act recognition is a key task involving categorizing utterances for their intent or function. GenAI frameworks have shown great promise on this task for both middle school student text conversations (Acosta et al., 2024; Kim et al., 2024), as well as for analyzing team dialogue (Min et al., 2021; Pande et al., 2023) across multiple training exercises, including Squad Overmatch (Johnston et al., 2019) and Critical Combat Care Air Transport Team (Robinson et al., 2023) as part of the Team Communication Analysis Toolkit (TCAT) project, sponsored by the US Army DEVCOM Soldier Center. Such capabilities, combined with improved automated speech recognition techniques, could provide the foundation for more robust and effective AI companion agents or coaches. These analytics can provide valuable insight to instructors and curricular designers, allowing them to better understand how trainees are performing in EL activities.

GenAI Throughout the Training Cycle

Using CBEL in training means objectives are clearly aligned, training is executed with linked interventions, and the subsequent After Action Review (AAR) is informed by data points within the training. Adaptive feedback generation can play a part in making this process effective, efficient, and more objective. Training that is adaptive will better allow for focused interventions as needed, with AAR content that highlights areas to work on, as well as areas to thrive on. Approaching training in this way provides a solid conduit from training planning to training evaluation.

Adaptive training is an instructional approach that dynamically adjusts the content, pace, difficulty, or delivery method of training to meet the specific needs, skills, and progress of individual learners or teams (Spain et al., 2012). The primary goal of adaptive training is to maximize learning efficiency and effectiveness by tailoring the training experience to optimize outcomes for diverse sets of participants. In pursuing this goal, adaptive training, while more effective, may also be more difficult to create, monitor, and review since it requires additional content and selection of interventions under more complex conditions.

The four phases of training—planning, preparation, execution, and review—provide a structured framework for effective military training. By integrating GenAI methods into the four phases of training, military organizations can create a more responsive, efficient, and effective training ecosystem that evolves to meet emerging challenges and operational demands. GenAI methods have been implemented to create, automate, and optimize training processes thereby reducing the workload, skill and time needed to support adaptive training. Examples of how GenAI may be applied in each phase are discussed below.

In the planning phase, GenAI plays a pivotal role by streamlining the development of training plans and scenarios. By processing vast amounts of historical data, doctrine, and mission requirements, GenAI can create comprehensive, contextually appropriate operational orders (OPORDs) and mission scenarios. For example, AI models can analyze prior training exercises and real-world operations to generate scenarios that align with specific mission-essential competencies (MECs) and anticipated challenges. GenAI also enhances collaboration during the Military Decision-Making Process (MDMP). It can propose multiple courses of action (COAs), simulate potential outcomes, and assist commanders in selecting optimal strategies. This capability accelerates planning, reduces cognitive load, and ensures that training plans are tailored to meet operational needs, leaving planners with more time to focus on strategic decision-making.

During the preparation phase, GenAI supports adaptive training by automating the creation and customization of scenarios based on defined training objectives. Instead of relying on static, pre-built scenarios, GenAI can dynamically generate diverse and realistic training environments. For instance, if the goal is to train a naval strike group in anti-air warfare, the AI could create scenarios featuring a variety of adversarial tactics, weather conditions, and resource constraints, ensuring trainees are exposed to a wide range of challenges. Generative AI also enables continuous scenario refinement. Based on real-time input from instructors or past trainee performance data, the AI can introduce variations to better align with skill levels or address gaps in competency. This adaptability ensures that the scenarios remain relevant and challenging, enhancing trainee engagement and preparedness.

GenAI significantly enhances adaptive training during the execution phase by enabling real-time scenario adaptation. As trainees engage in exercises, the AI monitors their actions and performance against established objectives and metrics. It then adjusts the scenario dynamically, introducing new elements or modifying existing ones to ensure that trainees are continually challenged in meaningful ways. For example, if a team demonstrates proficiency in a particular mission task, such as securing a perimeter, the AI could escalate the scenario by introducing unexpected adversarial actions or environmental changes, such as a communications blackout. Conversely, if trainees struggle with a task, the AI might simplify the scenario

or provide additional resources, such as virtual coaching or step-by-step guidance. GenAI also facilitates individualized learning by tailoring interventions to each trainee. For instance, an AI-powered virtual instructor could generate personalized feedback or practice drills, ensuring that every participant receives the support needed to progress effectively with their learning.

In the review phase, GenAI revolutionizes the AAR process by automating data analysis and generating insightful feedback. By analyzing performance data captured during the training exercise, the AI can identify patterns, pinpoint areas of improvement, and highlight successful strategies. The AI can also generate detailed AAR reports, linking trainee actions to mission objectives and MECs. These reports might include visualizations of performance trends, summaries of key decision points, and recommendations for future training. For example, if a scenario involved a complex evacuation under fire, the AI could analyze communication logs, resource allocation, and response times to provide actionable insights on teamwork and decision-making. GenAI ensures that AARs are not only comprehensive but also timely. Immediate feedback helps trainees retain lessons learned and apply them to subsequent training iterations. Furthermore, the insights from AARs can inform the planning and preparation phases of future exercises, creating a continuous improvement loop that enhances training effectiveness.

GenAI introduces a new era of efficiency, realism, and adaptability in training. During the planning phase, it accelerates scenario design and ensures alignment with mission objectives. In preparation, it dynamically generates and refines scenarios to meet training needs. During execution, GenAI enables real-time adaptation to trainee performance, ensuring that exercises remain engaging and relevant. Finally, in the review phase, it automates and enriches AARs, linking feedback directly to mission goals and fostering continuous learning. By integrating GenAI into adaptive training methods, organizations can create a transformative training ecosystem that prepares individuals and teams for the complexities of modern operations.

GenAI Enhanced Feedback and Coaching to Improve Transfer and Retention

During the four phases of training described above, the execution and review phases provide opportunities for coaching. This is coaching that is adaptive and will create a more valuable experience for the trainee with guidance specific to individual needs. Coaching can more readily be effective and efficient when tailored in this way with the support of GenAI. In applying CBEL at USMA, coaching will be incorporated both in AI form (adaptive feedback) and from the instructor (based on individual and group performance). This interdisciplinary engineering process approach involves comprehensive input from technical experts, learning experts, and subject matter experts (SMEs).

In some learning domains, successful transfer can be more challenging to achieve due to an increased distance between what can be learned in a formal instructional environment and the behaviors needed for successful performance. For example, leadership training is by its nature more ambiguous than technical skills training due to a lack of “correct” context-independent answers. In technical skills training, the ability of a trainee to perform at a validated standard can be defined clearly, whereas in leadership training, there is a much broader and less concrete set of competencies that may be relevant to any particular performance context. To address this, the traditional goal of leadership training is to develop these base competencies, with the assumption that trainees will gradually learn to identify when and where to use those skills as they progress in their careers. One effective way to bridge this gap sooner is through leadership coaching, a formal process by which a leadership coach guides trainees in the application of what they have learned to their day-to-day work context, setting difficult but achievable goals, and creating a sense of accountability. Unfortunately, leadership coaching is very resource-intensive, requiring 1.5 to 2 hours of a coach’s time for every 1 hour of coaching received by each trainee. Coaching is also provided outside of a traditional training context in that it is generally applied after the conclusion of relevant training at the time there is

opportunity for transfer to occur, creating significant logistical challenges. Thus, it is not practical to deploy such a system to all who would benefit.

A GenAI based leadership coach can potentially address this challenge directly by converting what was previously a very human-resource-intensive and location-bound system into an automatable, scalable digital solution that could theoretically be applied across the entirety of the DOD at a relatively low per-person cost. Developing an AI coach that successfully mimics human coaching is very challenging, as doing so requires simulating both the coaching skills and context-related subject matter expertise of trained coaches into an automated solution with no human oversight at the point of use. In one such project intended for use at the United States Naval Academy (USNA), a coach like this was built by blending coaching experts with doctoral training in industrial-organizational psychology, GenAI experts with doctoral training in computer science, DevOps experts in the deployment of complex web-based systems, as well as traditional front-end and back-end developers working toward a positive user experience. This combination was necessary to create a compelling coaching experience, one in which 1) users are invited automatically to the system, 2) the coach engages in an open-ended conversation in which the coach probes and draws conclusions about the trainee's challenges and goals, 3) the coach engages in collaborative goal-setting to be executed over a two-week interval, 4) the system checks in via text message or email halfway through the interval, and then 5) the coach follows up in a subsequent meeting, a pattern repeated indefinitely. Thus, all coaching interactions must be not only completely unique to the trainee's personal concerns and conversational approach but also useful to their growth (Office of Naval Research, 2024).

Such a complex system also necessitates a complex evaluation method at multiple levels of analysis, starting from initial prototyping but continuing indefinitely into use. For a project with USNA, in the initial prototyping phase, this was accomplished by asking SMEs internal to the core development team to engage with the system role-playing as trainees and provide structured feedback. After numerous prototyping rounds, this transitioned to focus groups with trainees providing group feedback on a facilitated session with the AI coach. After more prototyping rounds, this transitioned to one-on-one think-aloud sessions. Across these rounds and those that followed, the development goal was always to ensure that the GenAI system was robust to all possible conversational directions and content in that conversational content needed to remain aligned to system goals while also meeting the expectations of users. For example, each user test also contributed to building a "red teaming" list of scenarios that the development team re-tested with each subsequent prototype. Once the system was deemed robust by SMEs based upon user feedback, it was ready for deployment with authentic users, which necessitated another evaluation method still. Specifically, once embedded within the course schedules of USNA midshipmen, the system's success can now be evaluated by tracking changes over time within user groups in terms of various self-report and student performance metrics, as well as expert review of transcripts from the system itself.

GenAI Supported Feedback and Coaching Through Adaptive Learning Experiences

GenAI can also significantly shape how coaching and feedback are delivered to learners during experiential learning activities. Feedback and coaching are critical to learning and skill development (Shute, 2008). Feedback can reduce uncertainty by showing individuals how to correctly perform tasks (Billings, 2012) and can help enhance mental models and develop schemas to avoid future mistakes (Moreno, 2004). Specific forms of feedback such as corrective feedback can function to highlight errors, replace incorrect actions with appropriate responses, and reinforce correct behaviors (Badami et al., 2011). Determining how feedback should be delivered, the content of the feedback message, and the level of cognitive processing required is critical for optimizing learning and promoting engagement among trainees (Fahid et al., 2021).

There is growing evidence that GenAI can provide effective coaching and feedback to support adaptive learning experiences. LLMs can be used to generate and deliver real-time explanatory feedback and to automatically craft feedback messages that are tailored to trainees' errors and level of performance (Nguyen

et al., 2023). GenAI can also be used to craft robust interactive learning experiences, allowing learners to engage in conversations with agents that can provide highly effective active learning experiences. Recent research also shows that while LLMs can automatically generate feedback to support engaging learning experience, feedback and coaching message and statements need to be valid, aligning to the concept being taught, thus feedback evaluation is an important step in ensuring effective GenAI-driven learning experiences. One potential way to address the need to evaluate GenAI driven feedback is to use Reinforcement Learning (RL) (Scarlatos et al., 2024). RL refers to a family of machine learning techniques focused on creating agents that perform actions in an environment to optimize a numerical reward. Using RL, the AIS can learn the best type of feedback (e.g., positive reinforcement, constructive criticism, detailed explanations) for different users based on their responses, progress, and preferences. The RL model would receive feedback from the user (e.g., how well they understand the feedback or how they improve after receiving it) and adjust its future responses to coach accordingly. RL can also be used to determine the optimal time and amount of feedback to provide to maximize learning and engagement (Jeuring et al., 2022; Smith et al., 2022).

Discussion and Recommendations for Future Research Using GenAI

Gen AI is emerging as a powerful tool in narrative-centered experiential learning, where interactive stories in 3D environments engage learners in problem-solving scenarios. By generating rich narratives, providing adaptive feedback, and analyzing dialogue, AI supports the creation of dynamic, personalized learning experiences while easing the burden of content creation. Tools like NarrativeGenie (Kumaran et al., 2024) exemplify how LLMs can drive both the development and analysis of immersive training environments.

GenAI enhances the entire training cycle by automating and personalizing key phases—planning, preparation, execution, and review—that contribute to an adaptive, experiential learning environment where progression to competence mastery is a goal. It streamlines scenario generation, supports real-time adjustments based on trainee performance to inform future exercises, and contributes to rich, data-driven AARs. This integration creates a responsive and efficient training ecosystem tailored to evolving operational demands.

GenAI offers a scalable solution to leadership coaching by simulating personalized, goal-oriented support traditionally delivered by human coaches—an especially valuable approach in domains where skill transfer is complex and context-specific, making skill retention more challenging. The project at USNA demonstrated how an AI coach can sustain tailored coaching cycles, track trainee progress, and evolve through rigorous prototyping and feedback, making leadership development more accessible and adaptive across large organizations.

GenAI is transforming how feedback and coaching are delivered by generating real-time, personalized responses that adapt to learner performance, adding to the overall learning experience. To ensure this feedback is both effective and aligned with learning goals, RL can be used to evaluate and optimize the type, timing, and delivery of feedback based on individual user needs. This integration enhances engagement, supports skill development, and helps close the gap between training and real-world performance.

The most promising aspect of GenAI and LLMs lies in their ability to personalize and adapt learning experiences at scale, transforming static instruction into dynamic, learner-centered journeys. By generating interactive narratives, adaptive feedback, and context-aware coaching, they enable training environments that evolve with individual needs—something previously limited by time, cost, and human resources.

Broader opportunities include creating immersive, AI-driven training ecosystems that span technical, cognitive, and leadership development. Tools like NarrativeGenie illustrate how AI can revolutionize scenario development, while AI coaching platforms open pathways to foster mentorship and guidance across large organizations. Integrating RL further strengthens these systems by ensuring continuous improvement based on user feedback and outcomes.

Key challenges involve ensuring feedback quality, maintaining alignment with instructional goals, and safeguarding user trust. To address these, training must incorporate rigorous human-in-the-loop evaluation, domain-specific expertise, and transparent design practices during the four phases of training—planning, preparation, execution, and review. Doing so takes time and resources, but provides the structure to create more responsive, efficient, and effective training that can readily evolve to meet emerging challenges and operational demands. Investing in robust prototyping, testing, and ongoing performance monitoring will be critical to harnessing GenAI's full potential responsibly.

Recommendations for GIFT and STEEL-R Overall

When considering the trends and advancements of GenAI in the adaptive learning space, it is important to consider their implementation explicitly aligned to existing architectures and standards informed data strategies. The Generalized Intelligent Framework for Tutoring (GIFT) was designed to accelerate the authoring, delivery, and evaluation of adaptive learning services and solutions across military and professional education contexts. GIFT was developed to address the challenges of implementing intelligent tutoring at scale and leverages a modular approach to enable re-use and generalizable workflows where possible (Sottolare et al., 2013). The STE Experiential Learning for Readiness (STEEL-R) data strategy was instantiated to extend GIFT's compliance with the Total Learning Architecture (TLA) and to make assessments managed within GIFT interoperable with competency frameworks and ecosystem type learning management systems. These joint architectures provide a platform for the integration of GenAI to support CBEL. The CBEL learning cycle employs a systemic process of immersive simulation capabilities and modern adaptive instructional technology that stems from the STEEL-R science and technology framework (Goldberg et al., 2021; Owens et al., 2022). CBEL utilizes the STEEL-R framework to set up the simulation for students to interact with, as well as set up real-time data collection, stimulus strategies, and performance assessment algorithms that an instructor can monitor and/or control remotely in real-time. While significant advancements have been made to support rapid creation of intelligent tutoring functions across distributed, virtual and live learning environments, many challenges remain, preventing widespread adoption. The following recommendations outline how GIFT and STEEL-R can evolve to leverage GenAI for CBEL through scenario creation, adaptive instruction, multimodal assessment, and personalized coaching, ultimately enabling more scalable, data-driven EL environments.

One of the core opportunities for expanding GIFT's authoring capabilities in the preparation phase is the addition of a dedicated Narrative Generation Module within GIFT that would provide instructors with an AI-assisted authoring environment capable of generating complete, instructionally aligned scenarios from high-level inputs such as learning objectives, competency targets, and contextual constraints. Research in narrative-centered learning and automated scenario generation demonstrates the feasibility of generating coherent narrative arcs, event sequences, and dialogue structures that can be ingested by simulation engines (Kumaran et al., 2023). The module would output scenario structures, narrative arcs, event sequences, and dialogue in formats that can be processed by GIFT's scenario engine. This approach supports a scalable, modular authoring process that reduces the workload of manual scripting without compromising instructor control or domain alignment – setting up the simulation for students to interact with.

During training execution, GIFT can integrate a Real-Time Adaptation Engine that processes STEEL-R's structured event data. As learner actions, decisions, communications, and environmental interactions are

encoded through STEEL-R's standardized schema, the Adaptation Engine can evaluate performance against predefined metrics as well as detect deviations from expected behavioral patterns and subsequently initiate targeted instructional adjustments. For instance, if, during a LandNav task, a trainee repeatedly misidentifies certain terrain features, the Adaptation Engine can trigger a corrective prompt or adjust the scenario to provide additional practice under controlled conditions. In a team-based training exercise, the same mechanism could be used to detect breakdowns in communication, such as delayed acknowledgement of commands, and inject an adaptive intervention to reinforce the appropriate communication/coordination protocols. The execution of these adaptive responses may involve modifying the parameters of the scenario, delivering context-specific guidance, or dynamically adjusting task difficulty to maintain an optimal level of difficulty. Integrating this capability establishes a closed-loop content and evaluation cycle in which scenario logic, learner performance data, and instructional adaptations are tightly synchronized. This allows GIFT to respond to learner needs precisely and consistently across training iterations - setting up real-time data collection, stimulus strategies.

For post-training assessment, GenAI can provide advancement of GIFT and STEEL-R in the form of an Automated AAR module that transforms STEEL-R's event logs into structured, competency aligned feedback. As STEEL-R captures the series of learner actions, decisions, communications, and system states, the module could then analyze the data streams in order to identify relevant patterns of performance such as delayed decisions, ineffective communications, or repeated procedural errors. Returning to our LandNav example, the module may detect when a trainee consistently deviates from optimal route planning or fails to identify terrain features before committing to a movement decision. The module would use analyses of these errors to automatically generate performance summaries that highlight key points in the exercise, link observed behaviors to mission essential competencies and provide evidence-based recommendations for targeted remediation. The automation of this process results in reduced instructor workload for AAR, increases consistency of AAR across training events, and accelerates the timeline of feedback delivery – through the use of performance assessment algorithms that an instructor can monitor and/or control remotely in real-time. In addition, STEEL-R's standardized data structures ensure that resulting AAR outputs can be fully traced to raw event logs.

In order to execute these capabilities, GIFT and STEEL-R require a set of architectural design elements that support clear oversight, flexibility, and control as AI capabilities are added. GIFT should integrate a pluggable service model that allows new AI components like narrative generators and adaptation engines to be integrated without disrupting existing workflows. Additionally, STEEL-R should include an AI audit trail that records when AI systems create content, make recommendations, and employ adaptations to ensure that everything is traceable and accountable. Control from the instructor's standpoint should be through built-in review and override options in GIFT's authoring tools. Finally, long-term scalability depends on system interoperability and data integration with all AI generated output (scenario metadata, adaptation decisions, AAR summaries) following the TLA-aligned data standards and STEEL-R schemas. This consistency supports cross-system analytics, long-term learner modeling, and easy integration of future AI services to allow GIFT and STEEL-R to function as a coordinated training environment for adaptive experiential learning at scale.

Conclusions

The integration of CBEL into a traditional military instructional environment marks a significant step in advancing learning engineering within defense education. Leveraging the infrastructure and curriculum of USMA DMI— focus on LandNav—provides longitudinal opportunities for testing innovative approaches that align training with immersive learning objectives, resulting in enduring competencies. As GenAI technologies are incorporated to support AISs and enhance scenario realism and personalization, CBEL emerges not only as a way to deepen learning but also as a catalyst for cultural and methodological

transformation. Additionally, observing and learning from implementation challenges and successes in this context will help chart a roadmap for broader applications and improvements in EL across military and civilian sectors.

References

- Acosta, H., Lee, S., Bae, H., Feng, C., Rowe, J., Glazewski, K., ... & C. Lester, J. (2024). Recognizing Multi-Party Epistemic Dialogue Acts During Collaborative Game-Based Learning Using Large Language Models. *International Journal of Artificial Intelligence in Education*, 1-25.
- Badami, R., VaezMousavi, M., Wulf, G., & Namazizadeh, M. (2011). Feedback after good versus poor trials affects intrinsic motivation. *Research quarterly for exercise and sport*, 82(2), 360-364.
- Billings, D. R. (2012). Efficacy of adaptive feedback strategies in simulation-based training. *Military Psychology*, 24(2), 114-133.
- Fahid, F. M., Rowe, J. P., Spain, R. D., Goldberg, B. S., Pokorny, R., & Lester, J. (2021, June). Adaptively scaffolding cognitive engagement with batch constrained deep Q-networks. In *International conference on artificial intelligence in education* (pp. 113-124). Cham: Springer International Publishing.
- Goldberg, B., Owens, K., Hellman, K., Robson, R., Blake-Plock, S., Hoffman, M. & Gupton, K. (2021). Forging Proficiency and Readiness through an Experiential Learning for Readiness Strategy. In Proceedings of the 2021 Interservice/Industry Training Simulation and Education Conference (IITSEC). Orlando, FL.
- Huskey, R. (2022, January 5). The science of 'Flow states', explained by a cognitive science researcher : ScienceAlert. ScienceAlert. <https://www.sciencealert.com/the-science-of-why-flow-states-feel-so-good-according-to-a-cognitive-scientist>
- Jeuring, J., Keuning, H., Marwan, S., Bouvier, D., Izu, C., Kiesler, N., ... & Sarsa, S. (2022). Towards giving timely formative feedback and hints to novice programmers. In Proceedings of the 2022 Working Group Reports on Innovation and Technology in Computer Science Education (pp. 95-115).
- Johnston, J. H., Phillips, H. L., Milham, L. M., Riddle, D. L., Townsend, L. N., DeCostanza, A. H., & Fitzhugh, S. M. (2019). A team training field research study: Extending a theory of team development. *Frontiers in Psychology*, 10, 1480.
- Kim, Y. J., Acosta, H., Min, W., Rowe, J., Mott, B., Chaturvedi, S., & Lester, J. (2024, November). Dual Process Masking for Dialogue Act Recognition. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 15270-15283).
- Kumaran, V., Rowe, J., & Lester, J. (2024, November). NARRATIVEGENIE: generating narrative beats and dynamic storytelling with large language models. In Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (Vol. 20, No. 1, pp. 76-86).
- Kumaran, V., Rowe, J., Mott, B., & Lester, J. (2023, October). Scenecraft: Automating interactive narrative scene generation in digital games with large language models. In Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (Vol. 19, No. 1, pp. 86-96).
- Min, W., Spain, R., Saville, J. D., Mott, B., Brawner, K., Johnston, J., & Lester, J. (2021). Multidimensional team communication modeling for adaptive team training: A hybrid deep learning and graphical modeling framework. In Proceedings of International Conference on Artificial Intelligence in Education (pp. 293–305). Springer, Cham.
- Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional science*, 32(1), 99-113.
- Nguyen, H. A., Stec, H., Hou, X., Di, S., & McLaren, B. M. (2023, August). Evaluating chatgpt's decimal skills and feedback generation in a digital learning game. In *European conference on technology enhanced learning* (pp. 278-293). Cham: Springer Nature Switzerland.
- Office of Naval Research (2024). Automated and adaptive coaching using large language models. (Award No. N00014-24-1-2336).
- Owens, K. P. (2023). *Learning engineering for competency-based experiential learning design*. In Proceedings of the Human Factors and Ergonomics Society 67th International Annual Meeting (pp. 1545–1550). Sage.
- Owens, K., Goldberg, B., Blake-Plock, S., Robson, R., Mishra, S. (2022). Introduction to Competency-Based Experiential Learning. IITSEC Orlando FL 2022.

- Pande, J., Min, W., Spain, R. D., Saville, J. D., & Lester, J. (2023, June). Robust team communication analytics with transformer-based dialogue modeling. In *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 639-650).
- Robinson, F. E., Huffman, L. C. S., Bevington, L. C. D., French, D., Rothwell, C., Stucky, L. C. & Hughies, A. (2023). Team coordination style is an adaptive, emergent property of interactions between critical care air transport team personnel. *Air Medical Journal*, 42(3), 174-183.
- Scarlato, A., Smith, D., Woodhead, S., & Lan, A. (2024, July). Improving the validity of automatically generated feedback via reinforcement learning. In *International Conference on Artificial Intelligence in Education* (pp. 280-294). Cham: Springer Nature Switzerland.
- Shute, V. J. (2008). Focus on formative feedback. *Review of educational research*, 78(1), 153-189.
- Smith, A., Spain, R. D., Rowe, J., Goldberg, B., & Lester, J. (2022, May). Formalizing Adaptive Team Feedback in Synthetic Training Environments with Reinforcement Learning. In *Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym10)* (p. 117).
- Sottolare, R. A., Brawner, K. W., Goldberg, B. S., & Holden, H. K. (2013). The Generalized Intelligent Framework for Tutoring (GIFT). 10.13140/2.1.1629.6003.
- Spain, R. D., Priest, H. A., & Murphy, J. S. (2012). Current Trends in Adaptive Training With Military Applications: An Introduction. *Military Psychology*, 24(2), 87–95.
<https://doi.org/10.1080/08995605.2012.676984>
- U.S. Army (2017). The U.S. Army Learning Concept for Training And Education 2020-2040. TRADOC Pamphlet 525-8-2.

CHAPTER 14 - OPPORTUNITIES FOR LEVERAGING GENERATIVE AI FOR CREATING REMEDIATION CONTENT IN GIFT

Anne M. Sinatra, Paige M. Lawton, and Randall Spain

US Army Combat Capabilities Development Command (DEVCOM) – Soldier Center

Introduction

The capabilities of Generative Artificial Intelligence (GenAI) are rapidly advancing, creating new and exciting opportunities to facilitate the development of adaptive instructional systems. Its capabilities range from content creation (e.g., generating instructional materials, questions, remediation content) to sophisticated functions like student modeling, domain modeling, and tutorial planning. Integrating GenAI into the Generalized Intelligent Framework for Tutoring (GIFT) could significantly reduce the authoring burden of creating intelligent tutoring systems (ITSs) and create unique opportunities for investigating how GenAI can support personalized learning experiences. GIFT is an open-source ITS framework and highly flexible as it is intended to support tutoring in an author-defined domain. One of the challenges of creating a framework such as GIFT is that it requires course authors to provide or create a large amount of content to support different levels of adaptivity within the system. The necessary amount and variations of this content grows relative to the desired level of system adaptivity.

GIFT supports adaptivity with respect to content through two approaches: (1) the Domain Knowledge File (DKF) which facilitates adaptation with external simulation-based training environments, and (2) the Adaptive Courseflow Object, which relies on author-provided content (e.g., PowerPoint, multiple choice questions) to facilitate adaptive hypermedia-based training experiences. In this chapter, we discuss the Adaptive Courseflow Object and opportunities for using GenAI to automatically create remediation content that supports personalized training experiences.

GIFT's Adaptive Courseflow Object

As a modular ITS architecture, GIFT provides GIFT course creators with a suite of tools for rapidly creating intelligent tutors and deploying adaptive instructional experiences. These tools, many of which are available through GIFT's Course Creator, support the development of intelligent tutors across a range of instructional domains. Central to this authoring process is GIFT's Adaptive Courseflow Object which serves as the primary mechanism for structuring instructional activities that dynamically respond to learner needs.

In contrast to a fixed, linear sequence of instruction, the Adaptive Courseflow Object enables branching instructional paths and personalized learning experiences. Instructional decisions are managed by GIFT's Engine for Management of Adaptive Pedagogy (EMAP) which evaluates inputs from the Learner Model (including measures of mastery, confidence, and past performance) to determine the most appropriate next step in the instructional sequence. The adaptive logic embedded in the Adaptive Courseflow Object is grounded in established learning science theories, most notably Component Display Theory (Merrill et al., 1992) and the ICAP framework for cognitive engagement in learning (Chi & Wylie, 2014). The authoring scheme for an Adaptive Courseflow Object operationalizes these theories by organizing instruction into five primary quadrants. These include the: (1) Rules quadrant, in which learners are introduced to the core principles and concepts of a domain; (2) Examples quadrant, where learners see how the rules are applied in content; (3) Recall quadrant, where learners demonstrate declarative and procedural knowledge associated

with the rules; (4) Practice quadrant, where learners demonstrate knowledge and skills in novel or simulated contexts, and; (5) Remediation quadrant, where learners receive targeted instructional support if they have not reached mastery.

GIFT courses built with Adaptive Courseflow Objects guide learners through these distinct quadrants. Students begin by reviewing multimedia training materials aligned with the Rules quadrant. Following this foundational instruction, they transition to the Examples quadrant to view additional content illustrating examples of the target concepts. Subsequently, students transition to the Recall quadrant, where their comprehension is assessed through a series of quiz questions. If sufficient mastery has been demonstrated, the student may proceed to the Practice quadrant to apply their knowledge within a training simulation designed to support hands on learning or skill transfer.

Complementing these core instructional activities, is the Remediation quadrant that addresses cases in which learners fail to demonstrate mastery of a course concept. Remediation is designed to provide targeted, ICAP-inspired remedial content tailored to the learner's needs. Three forms of remediation are currently supported: Constructive response activities that require learners to generate explanations or solutions; Active response activities that engage learners through guided interaction; and Passive response materials delivered through media such as videos, text resources, or web-based content. Constructive and Active remediation activities are authored using GIFT's built-in templates, while Passive remediation can be authored using various applications. Applications that support Interactive remediation using pedagogical agents are currently being developed.

As part of the authoring processes, course creators curate instructional content, tag materials with metadata to support adaptive selection, construct assessments using GIFT's survey authoring tools, configure practice scenarios, and develop remediation resources aligned with instructional goals. At runtime, this metadata-driven architecture enables GIFT to continuously monitor learner performance, update the Learner Model, and dynamically select instructional content aligned to a learner's instructional pathway. The result is a responsive learning environment in which instruction is personalized to optimize learner engagement, progression, and mastery over time.

While Adaptive Courseflow Objects provides a framework for authoring adaptive course experiences, curating and creating course content is resource intensive, requiring a considerable amount of time and expertise. Course authors must populate the Rules and Example quadrants with relevant instructional material; develop and design a large pool of questions that accurately measure comprehension and vary in difficulty for the Recall quadrant; develop effective and realistic practice scenarios; and assemble remediation content that targets different levels of understanding and comprehension. In addition to these tasks, authors must manually tag each piece of content with appropriate metadata so GIFT's EMAP can select and sequence the appropriate instructional materials. These course authoring tasks can be a significant bottleneck in developing and deploying new adaptive courses in GIFT. This is precisely where GenAI tools can transform the authoring process.

In particular, GenAI offers the potential to automatically produce remediation content that responds to a learner's performance. This capability can be integrated with learner modeling modules that detect misunderstandings, disengagement, or domain-specific errors during instructional activities. Based on these observations, authors can leverage multimodal large language models (LLMs) to generate tailored instructional materials like simplified explanations, targeted examples, or practice scenarios. These materials can then be integrated within the Adaptive Courseflow logic to deliver remediation content that aligns with the learner's needs as well as the intended pedagogical strategy. By integrating GenAI into the course authoring process, developers can significantly reduce the time required to produce and revise multiple versions of content (McNeill, 2024). Rather than manually creating content for different

remediation strategies, authors can specify remediation triggers based on learner states and prompt a generative system to automatically produce relevant content for remediation (Yu & Guo, 2023).

However, the integration of GenAI into ITS design also raises critical considerations regarding trust, transparency, and instructional accountability. Learners and instructors must be able to understand, evaluate, and, if needed, override system-generated content and assessments. In the case of GIFT, there would be advantages to providing interfaces during the Adaptive Courseflow Object authoring process to facilitate the course authors' examination of the accuracy of the materials and remediation content that has been generated by AI before it is shown to a learner. There are additional challenges when AI is used in real-time during a tutoring session, as it cannot be reviewed by the course author in real-time. Establishing transparency around how generative content is produced and aligned with instructional goals is essential for building trust in ITS environments. Design choices that support metadata clarity, traceable reasoning, and authoring controls help ensure that GenAI is not only a more efficient option when producing remediation content, but also a dependable collaborator that reliably aligns content with instructional goals. Ensuring this alignment requires responsible and transparent handling of factors such as data management, policy disclosure, and instructional design decisions (Bozkurt & Sharma, 2024).

The Role of Generative AI in Developing Remedial Content

AI for Content Creation

Developing engaging and effective content for online and blended learning requires substantial resources and can be labor intensive. The high standards for learning content, which can include videos, images, interactive elements, and adaptive assessments, create substantial workloads and challenges for instructional designers. In addition to content creation, instructional designers must also source or produce multimedia elements, write explanatory text, and develop assessments that identify the varying needs and prior knowledge of learners. These requirements can complicate efforts to meet the increasing demand for timely and relevant learning resources, while also ensuring that learning modules achieve their desired outcomes (Kerres & Buntins, 2020).

As instructional designers continue to explore ways to scale adaptive learning systems, GenAI presents a practical tool for supporting multimodal content generation. Recent studies highlight the potential that GenAI has in reducing manual burden while maintaining the quality of instructional content (Xiaoyu et al., 2025). In recent years, LLMs, image generators, and voice synthesis tools have been used to facilitate drafting instructional content, generating assessments, and producing multimodal content; however, the intent is not for GenAI to replace expert instruction but to augment it and to lessen workload. For instance, GenAI may be used to quickly provide initial content/outputs that can be approved or refined by authors to ensure that content aligns with pedagogical goals (Chng, 2023). This focus on GenAI capabilities has inspired research which examines how generative tools can be used to automate tasks, personalize content based on learner data, and stimulate the creation of remediation materials.

Recent studies have documented how GenAI tools are being integrated into instructional design workflows to support content creation across formats. Bolick and da Silva (2024) examined how generative platforms such as ChatGPT, Midjourney, and Descript are being used to automate content generation such as instructional text, visual displays, animations, and narrated media. They found that these tools reduce the time required to produce materials and allow authors to focus on higher-order instructional decisions. Additional work highlights how LLMs are being used to generate personalized feedback and instructional prompts within ITSs (Maity & Deroy, 2024). These systems can analyze learner performance, such as incorrect answers, hesitations, or skipped steps, that signal a need for remediation and generate remediation

content that is contextually relevant and pedagogically aligned to learner needs. For example, if a learner repeatedly has trouble with a concept or task within a simulation, the system can generate remediation in the form of a simplified explanation, a worked example, or a visual aid that has been adapted to the learner's trend of errors.

Illustrative Example: Adaptive Master Gunner Course in GIFT

In this section, we describe an adaptive computer-based training course developed for the US Army's Maneuver Center of Excellence (MCoE) Master Gunner Course. The course covered basic concepts associated with the Detect, Identify, Decide, Engage, and Assess (DIDEA) process while providing trainees with remedial content. We provide a general overview of the course and the manual GIFT course authoring process that we used. Further, we discuss how GenAI could have been used to expedite the GIFT course authoring process.

The Master Gunner Course and the DIDEA Process

The US Army's MCoE is responsible for training Soldiers and leaders for the maneuver force. Among its training is the Master Gunner Course, which develops non-commissioned officers into experts on their respective weapon platforms. The course is traditionally delivered through instructor-led classroom sessions that include heavily detailed PowerPoint presentations. These presentations provide the foundational knowledge required for advanced gunnery, covering complex technical procedures and tactical information.

Central to the gunnery curriculum is the DIDEA process, a foundational cognitive model for direct-fire engagement. DIDEA provides a cyclical framework that enables crews to rapidly detect potential targets, positively identify them as threats, decide on an appropriate course of action, engage the target with the optimal weapon system, and finally, assess the effects of that engagement to determine if the threat has been neutralized. The procedural and doctrinal nature of this topic makes it an ideal candidate for leveraging GenAI to efficiently author an adaptive computer-based course that includes engaging instructional content and associated remediation activities that can enhance and accelerate a trainee's learning experience.

GIFT Course Authoring

Without the use of AI, we developed a series of courses in GIFT to address selected components of the DIDEA process (Detect, Engage, and Assess) to create an adaptive learning experience. A primary challenge was developing sufficient variations of the original course materials, which consisted of PowerPoint slides provided by course instructors. These PowerPoint slides served as a starting point to identify course concepts and learning objectives and to create an extensive multiple-choice question-bank used in the Recall phase of Adaptive Courseflow Object. All newly created questions were shared with the Subject Matter Experts (SMEs) for review/approval to confirm that they were accurate.

For remediation content, we initially used an alternate version of the content slides that were provided by the SMEs. While we identified opportunities to use unique features in GIFT to support more interactive remediation, such as using audio files of Fire Commands and with supplemental questions, these approaches were not utilized due to time and resource constraints. Instead, to add variation, we manually created Active and Constructive remediation tasks that asked learners to review and summarize course concepts in their own words before being shown an expert summary. Manually developing the remedial activities for each course concept was labor intensive and required validating that newly created content

accurately aligned to the original course material. Incorporating GenAI into this process could have significantly reduced the authoring workload while increasing the volume of available materials. Indeed, recent research has shown LLMs can ingest existing documents or slide decks to produce enriched learning materials, including detailed explanations, summaries, and varied assessment questions (Attard & Dingli, 2024; Morales-Chan et al., 2024; Moore et al., 2023). In our pilot study, applying such a process could have dramatically accelerated the creation of question banks and remedial content from the initial SME-provided PowerPoint files. This would have transformed the authoring phase from a laborious content creation task into a more efficient process of content evaluation, allowing the course author to focus their efforts on refining and validating the AI-generated materials with the SMEs, rather than starting from scratch.

Question Bank Generation

One of the main tasks for an author within the Adaptive Courseflow Object is to populate the question bank for each course concept. For the Master Gunner GIFT courses, a significant amount of effort went into reviewing the PowerPoint presentations and creating questions of different difficulty levels that aligned to the course concepts. To facilitate this process, LLMs could be provided with the original content slide for each of the concepts and asked to generate multiple choice questions based on it. The use of GenAI for this purpose would significantly decrease the time and effort required to create a larger number of questions of varying levels of difficulty for each concept. These questions could then be reviewed by the course instructor to confirm their accuracy. For example, Gopi et al., (2024) show how LLMs can be used to generate adaptive questions. These researchers used a retrieval-augmented generation (RAG)-based approach along with a dual-prompt approach to dynamically develop quiz questions. The first prompts, the Contextual Question Prompt, formulated questions based on user topics and chat history, while the Answer Question Prompt managed question responses with explanations, ensuring relevant and contextually accurate interactions. This approach achieved an 88.5% topic relevancy score and even allowed for quizzes to be generated on-demand from user-uploaded documents, showcasing a scalable and effective model for generating topic-relevant multiple-choice quiz questions.

A similar approach could be used in GIFT to generate quiz questions of varying difficulty (easy, medium, difficult) aligned to course content. In addition to generating questions, GenAI could also automate metadata tagging as part of the process. Each question could be tagged with details used by the GIFT system such as course concept, difficulty level, and/or expected performance thresholds. The goal is that this would make it easier to align questions with GIFT's adaptive logic and support real-time learner modeling.

Remediation Content Generation: PowerPoints

A key component of remediation within the GIFT architecture involves presenting targeted instructional materials, often in the form of PowerPoint slides, as they are often pre-prepared, and readily available to the course authors. However, often a course author will only have one version of slide materials created. In the Adaptive Courseflow Object, GIFT can present different initial content or remediation content to the learner based on attributes (e.g., motivation level), and their performance. GIFT's EMAP has pre-programmed preferences on the types of content that should be presented based on individual learner attributes, and it selects the content using metadata. GenAI could be used by providing an initial version of a slide, and asking for different versions that correspond with different characteristics of learning material, such as more visual (image focused) or more text based. Hashiyada et al. (2025) presents a framework that uses LLMs and RAG to automatically create detailed and personalized learning materials from lecture slides. The system works by extracting the structure from PowerPoint slides and then using RAG to add detailed content from a knowledge base, which can be customized to a student's learning style.

Remediation Content Generation: Examples, Constructive Prompts, and Interactive Agents

In addition to creating slides, GenAI can also be leveraged to create different forms of multimedia content ranging from videos (Chen et al., 2025), podcasts (Menon et al., 2025), and visual novels (Gursesli et al., 2025) to engaging games (Ordoumpozanis & Apostolidis, 2025) and LLM-empowered pedagogical agents to support remediation (Liu et al., 2025). One of the challenges encountered with creating the Master Gunner GIFT courses was that materials such as audio examples of FIRE Commands were not readily available, and it would take coordination, time, and resources to create them. GenAI could be utilized to help reduce these issues by generating example verbal FIRE Commands and potentially creating Audio files and variations of the files. They then could be shared with SMEs to ensure that they are accurate and then incorporated into the Adaptive Courseflow remediation. GenAI could have also been utilized to create additional materials such as videos which could have been used in a similar manner as part of remediation. GenAI could also have contributed to generating a much larger amount of constructive and summary prompts based on the initial learning materials which would have created a much larger variety of remediation content that learners could encounter.

Challenges

While GenAI offers promising efficiencies for producing remediation content, its integration into ITSs introduces challenges that impact instructional quality, authoring workflows, and system transparency.

Pedagogical Precision

With the use of GenAI, consistent alignment of produced content with instructional goals is not automatically guaranteed. One of the inherent risks with GenAI is that generative outputs can vary in depth of information, writing tone, and content validity (Talaver & Vakaliuk, 2025). The dynamic nature of AI-generated content means that responses can vary, sometimes unpredictably, depending on input context and model training. This variability can decrease trust in learning systems because it raises concerns about consistency and reliability. Additionally, while GenAI excels in pattern recognition and data-driven predictions, it lacks an intrinsic understanding of educational goals, ethical considerations, or contexts. This limitation necessitates careful system design to ensure that AI-generated content aligns with pedagogical principles and student expectations. To maintain pedagogical integrity, authors must validate that the generated content reinforces the intended learning outcomes and does not over-simplify or misrepresent domain knowledge. Maity and Deroy (2024) emphasize the need for human-in-the-loop review processes to address this challenge.

Authoring Control and Oversight

One of the advantages of GenAI is that it can produce large volumes of content quickly; however, this presents opportunities for inconsistencies in the quality of generated content. As a result, authors may spend more time adjusting and revising outputs to better align them with logic and learner needs. For instance, McNeill (2024) found that process efficiency gains are limited by the difficulty of engineering effective prompts, lack of learner personalization and context, and the need to verify the accuracy of outputs. Choi et al. (2024) further emphasizes the need for human oversight to ensure instructional quality, even when AI is used to generate adaptive materials. Implementation of deliberate review and contextual checks can be used to mitigate risk and ensure that AI generated content supports the intended learning outcomes (Stefaniak & Moore, 2024).

Transparency and Traceability

Learners and instructors must also be able to understand how and why remediation content was generated. A lack of transparency within the system can undermine trust and introduce complications within efforts to refine adaptive pathways. Recent literature reinforces the importance of maintaining traceability when integrating GenAI within ITSs. This is essential not only for allowing authors to verify the source and rationale of AI generated content to ensure instructional integrity (Choi et al., 2024), but also for preserving learner trust in adaptive systems (Garcia-Lopez & Trujillo-Linan, 2025). GIFT's modular architecture provides opportunities to embed traceability through instructional messages and content logs; however, these must be intentionally designed and maintained. GenAI presents an opportunity to automate this process and reduce authoring workload when it comes to metadata annotations and traceable logic.

Instructional Coherence Across Modalities

One of the advantages of ITS frameworks like GIFT is the ability to present material in different modalities based on learner attributes, such as motivation. Generated remediation content may take the form of text, images, examples, or other formats. A student with lower motivation regarding the content may be more receptive to images or examples, whereas a highly motivated student may benefit more from in-depth text-based content. Using GenAI to produce these materials in different formats presents a challenge when it comes to maintaining instructional coherence. Heilala et al., (2025) discuss the present limitation of most GenAI platforms, being that they are typically designed to handle one format at a time or tend to specialize in one format over others, such as better-quality text output vs. image generation. While text-based AI tools are most commonly used in educational applications, multimodal tools tend to be less developed and lack the integration needed to support instructional design. The lack of consistency in the quality of instructional content between formats can contribute to breakdowns in pedagogical alignment. For instance, a language model may emphasize clarity of a concept in text, but a visual produced by an image generator may oversimplify the same concept. This potential mismatch highlights the need for additional authoring oversight to ensure instructional consistency across modalities.

Discussion and Recommendations for GIFT

The series of Master Gunner GIFT courses described in this chapter provides an exemplar for creating and populating content in Adaptive Courseflow Objects. Leveraging GenAI to reduce the authoring burden, particularly for creating remedial content by creating variations based on existing provided content, could significantly streamline the ITS development timeline. While GenAI can rapidly create remedial content, course instructors and SMEs still have a vital role in the course development process. Their expertise is critical for reviewing and validating course material and engineering and revising prompts within the GenAI instructional design cycle. From the SME's perspective, the validation process is similar, but the content generation phase becomes significantly less effortful for the course author. This workload reduction is even greater if the author is also the SME. However, since the content is machine-generated, experts must review it with increased mindfulness. They need to be especially vigilant for potential errors or "hallucinations" that can be introduced by GenAI.

An additional area for consideration is integrating GenAI models and services directly into the GIFT authoring workflow. This feature could be transformative for rapidly creating adaptive training experiences. The core idea is that GenAI could act as an assistant at every step of the course creation process, helping instructors and course creators with domain modeling, content generation, adaptive branching and remediation, assessment creation, and feedback authoring.

Furthermore, GenAI can be used to create a dynamic and intuitive interface for managing complex adaptive courses within GIFT –i.e., it would allow course authors to see the “big picture” of a course, its content, and how all the components connect. By using natural language prompts, course creators could instantly search for content, view metadata, and generate custom on-the-fly visualizations of the course structure. Using GenAI in this manner would allow authors to easily see the relationships between all course objects, identify potential student pathways, and pinpoint content needing review, dramatically simplifying the process of authoring, testing, and refining courses to promote quality assurance.

While GenAI presents a promising tool for producing adaptive remediation within ITSs, its integration into operational training environments requires further empirical validation. GenAI tools within ITS frameworks such as GIFT should be tested to evaluate the tradeoffs between manually producing instructional materials and working with a GenAI agent to produce content and courses. As outlined earlier, challenges related to instructional alignment, authoring control, and multimodal coherence must be addressed before GenAI tools can be reliably scaled for authoring of remediation content. These challenges underscore the need for structured evaluation of both the generated content itself and of the authoring processes and instructional outcomes that result from AI-assisted design. A logical next step is to conduct a controlled evaluation comparing manually authored remediation content with content generated by LLMs. The primary goal would be to evaluate differences in authoring workload, instructional coherence, learner performance, and how well the content supports learners. A study of this nature would provide an opportunity to test how generative content can be integrated into GIFT’s Adaptive Courseflow, specifically how it fits with branching logic, remediation triggers, and metadata tagging for traceability. It could also identify the types of instructional tasks that are best suited for automation and those tasks where human oversight remains necessary.

Conclusions

The development of adaptive instructional experiences is largely limited by the workload needed to author high-quality, varied content to support learning. This chapter outlined how GenAI can be used within GIFT to mitigate this challenge, particularly by automating the creation of remediation materials that align with course concepts and instructional adaptations. Although the content was created manually for it, the Master Gunner GIFT Course example illustrates the potential of Generative AI for streamlining the production of instructional elements like question banks, alternative explanations of concepts, and other multimodal remediation materials that typically require substantial time and effort to generate. However, this opportunity for streamlining such processes introduces a new set of considerations for instructional quality and system transparency. It becomes necessary to maintain deliberate oversight to ensure that generated materials remain pedagogically aligned, coherent across modalities, and traceable within the adaptive framework. With this shift in workload, the role of the author becomes that of a validator of AI generated materials rather than the creator of the content. The adoption of a human-in-the-loop model is an essential part of the shift in authoring responsibilities in order to verify accuracy, maintain the integrity of instructional materials, and build trust with both learners and educators. By combining educator expertise with AI-driven content generation, personalized and adaptive instruction could be authored more efficiently, making it more accessible and more impactful for learners.

Acknowledgements

The research described herein has been sponsored by the US Army Combat Capabilities Development Command – Soldier Center. The statements and opinions expressed in this chapter do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred. We would like to acknowledge and thank Dr. Rebecca Robinson (Army University), Dr.

Benjamin Goldberg (Soldier Center), and the MCoE for their work on planning and coordinating the Master Gunner Course GIFT Pilot project which was an illustrative example in this chapter.

References

- Attard, A., & Dingli, A. (2024). Empowering educators: Leveraging large language models to streamline content creation in education. In *ICERI2024 Proceedings* (pp. 1312-1321). IATED.
- Bolick, A. D., & Da Silva, R. L. (2024). Exploring artificial intelligence tools and their potential impact to instructional design workflows and organizational systems. *TechTrends*, 68(1), 91-100.
- Bozkurt, A., & Sharma, R. C. (2024). Trust, credibility and transparency in human-AI interaction: Why we need explainable and trustworthy AI and why we need it now. *Asian Journal of Distance Education*, 19(2).
- Chen, Y., Lin, K. Q., & Shou, M. Z. (2025). Code2Video: A Code-centric Paradigm for Educational Video Generation. arXiv preprint arXiv:2510.01174.
- Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, 49(4), 219-243.
- Chng, L. K. (2023). How AI makes its mark on instructional design. *Asian Journal of Distance Education*, 18(2), 32-41.
- Choi, G. W., Kim, S. H., Lee, D., & Moon, J. (2024). Utilizing GenAI for instructional design: Exploring strengths, weaknesses, opportunities, and threats. *TechTrends*, 68(4), 832-844.
- García-López, I. M., & Trujillo-Liñán, L. (2025). Generative Artificial Intelligence in Education: Ethical Challenges, Regulatory Frameworks and Educational Quality in a Systematic Review of the Literature. In *Frontiers in Education* (Vol. 10, p. 1565938). Frontiers.
- Gopi, S., Sreekanth, D., & Dehbozorgi, N. (2024). Enhancing Engineering Education Through LLM-Driven Adaptive Quiz Generation: A RAG-Based Approach. In *2024 IEEE Frontiers in Education Conference (FIE)* (pp. 1-8). IEEE.
- Gursesli, M. C., Chen, S., Dewantoro, M. F., You, X., Anbar, E., Taveekitworachai, P., ... & Thawonmas, R. (2025). The Role of Large Language Model-Generated Stories in the Narrative Experience of Serious Visual Novel Games. *International Journal of Human-Computer Interaction*, 1-20.
- Hashiyada, K., Shi, W., & Yin, C. (2025). A Framework for Using LLMs and RAG to Realize the Automatic Generation of Learning Materials from Lecture Slides. In *International Conference on Learning Evidence and Analytics*.
- Heilala, V., Araya, R., & Hämäläinen, R. (2025). Beyond text-to-text: An overview of multimodal and generative artificial intelligence for education using topic modeling. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing* (pp. 54-63).
- Kerres, M., & Buntins, K. (2020). Recommender in AI-enhanced learning: An assessment from the perspective of instructional design. *Open Education Studies*, 2(1), 101-111.
- Liu, J., Chen, T., Li, S., Xia, Y., Zhu, H., Wu, R., ... & Wu, L. (2025). LLM-based pedagogical agent for ICU simulation instructor training: A quasi-experimental study. *Nurse Education Today*, 106901.
- Maity, S., & Deroy, A. (2024). GenAI and its impact on personalized intelligent tutoring systems. *arXiv preprint arXiv:2410.10650*.
- McNeill, L. (2024). Automation or innovation? A GenAI and instructional design snapshot. In *The IAFOR International Conference on Education-Hawaii 2024 Official Conference Proceedings* (pp. 187-194).
- Menon, V., Cherney, A., Cloude, E. B., Zhang, L., & Do, T. D. (2025, October). Evaluating the Impact of LLM-guided Reflection on Learning Outcomes with Interactive AI-Generated Educational Podcasts. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers* (pp. 99-106).
- Merrill, D. C., Reiser, B. J., Ranney, M., & Trafton, J. G. (1992). Effective tutoring techniques: A comparison of human tutors and intelligent tutoring systems. *The Journal of the learning sciences*, 2(3), 277-305.
- Moore, S., Tong, R., Singh, A., Liu, Z., Hu, X., Lu, Y., ... & Stamper, J. (2023, June). Empowering education with llms-the next-gen interface and content generation. In *International Conference on Artificial Intelligence in Education* (pp. 32-37). Cham: Springer Nature Switzerland.

- Morales-Chan, M., Amado-Salvatierra, H. R., & Hernandez-Rizzardini, R. (2024, July). AI-driven content creation: Revolutionizing educational materials. In *Proceedings of the Eleventh ACM Conference on Learning@Scale* (pp. 556-558).
- Ordoumpozanis, K., & Apostolidis, H. (2025). A Second-Generation Agentic Framework for Generative Ai-Driven Augmented Reality Educational Games. In *2025 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1-10). IEEE.
- Stefaniak, J. E., & Moore, S. L. (2024). The Use of Generative AI to Support Inclusivity and Design Deliberation for Online Instruction. *Online Learning*, 28(3), 181-206.
- Talaver, O. V., & Vakaliuk, T. A. (2025). A model for improving the accuracy of educational content created by generative AI. In *CEUR Workshop Proceedings* (pp. 149-158).
- Yu, H., & Guo, Y. (2023, June). Generative artificial intelligence empowers educational reform: current status, issues, and prospects. In *Frontiers in Education* (Vol. 8, p. 1183162). Frontiers Media SA.
- Xiaoyu, W., Zainuddin, Z., & Hai Leng, C. (2025). Generative artificial intelligence in pedagogical practices: a systematic review of empirical studies (2022–2024). *Cogent Education*, 12(1), 2485499.

CHAPTER 15 – SPECIAL CONSIDERATIONS AND NEXT STEPS FOR AI IN TRAINING ACROSS THE SERVICES, INDUSTRY, AND ACADEMIA

Lisa N. Townsend¹, Danielle McNamara², Robert A. Sottolare³, Benjamin Goldberg¹, and Wendi Van Buskirk⁴

¹US Army Combat Capabilities Development Command - Soldier Center¹; Arizona State University²; Soar Technology, LLC, an Accelint Company³; Naval Air Warfare Center-Training Systems Division⁴

Introduction

Artificial Intelligence (AI) has become a tool and a resource that is being more widely used and accepted in the training community. At its basis, AI refers to the simulation of human intelligence in machines, allocated to perform tasks typically requiring basic human cognitive processes such as learning, problem-solving, understanding natural language, and even creative thinking. From the training perspective this can also include applying the key components of Machine Learning (ML), Natural Language Processing (NLP), Computer Vision, and/or Robotics, to the development of intelligent systems (e.g., an Intelligent Tutoring System (ITS)). An ITS is a computer-based system designed to provide personalized feedback and adapt to a learner's performance, replicating the effectiveness of one-on-one tutoring (Graesser et al., 2012). ITSs are widely used in training across the Services, Industry, and Academia and they can provide opportunities to tailor learning experiences in real-time to an individual learner and adapt content based on learner performance (Mishra et al., 2026).

In Industry, AI is being used for fraud detection, risk management, customer service, personalizing recommendations, enhancing cyber security, and streamlining operations. For example, machine learning models can analyze transaction data to detect fraudulent activities, while AI-powered chatbots assist customers with inquiries and transactions. AI also plays a role in predicting market trends, helping investors make more informed decisions (Wu, 2025).

In Academia, AI is being used to create personalized learning experiences for students. AI-driven tools can assess individual learning styles and progress, offering tailored lessons and feedback. Virtual tutors powered by AI are also helping students with subjects ranging from math to language learning.

AI uses in Military training contribute to increasing subject retention, decreasing class time, adapting training to individual needs, developing an optimized curriculum, generating varying scenario content, and creating comprehensive student assessment materials. Department of Defense (DoD) leaders and Warfighters use AI to prepare for and make rapid, well-informed decisions by leveraging high-quality data and advanced analytics as part of a continuous, outcome-driven, and user-focused development, deployment, and feedback cycle. Quality data is the critical foundation, because all analytic and AI capabilities require trusted, high-quality data (Stensrud, 2024). The rapid development and integration of reliable intelligent agents capable of replicating human-like decision-making and behavior can lead to Warfighters who are immersed in more complex and challenging training exercises with a fraction of the manpower previously required for support and control. By leveraging AI-driven agents, dynamic and unpredictable scenarios can be developed, pushing Warfighters to adapt and refine their tactics in a high-fidelity training environment. AI can provide DoD leaders with insightful analytics, foundational models, and visualizations. This can result in DoD leaders having the capability to understand their domains better and elevate interpretation of the key variables impacting decisions for and outcomes in those domains.

Leveraging AI, can create more robust and impactful training experiences that better prepare Warfighters for the complexities of the modern battlefield (Stensrud, 2024; Wheeler Atkinson et al., 2023;).

This chapter brings together unique perspectives from Government, Industry, and Academia authors to illustrate how each community approaches and applies AI within the training domain. As these sectors use slightly different terminology, priorities, and examples, the sections intentionally reflect their distinct yet also interrelating viewpoints.

The use and acceptance of AI occurring at a somewhat disjointed and rapid pace causes a need for considerations to be defined and examined and for next steps to be insightfully curated and adhered to on a consistent basis as this tool and resource evolves and its applications increase. For training, whether in the Military, in Industry, or in Academia, these considerations and next steps have similar challenges and create similar concerns and thus there is much overlap in implementation, risk, and lessons learned. Grewal et al. (2024) argue for appropriate managerial and policy efforts because such efforts can minimize the perils (hallucinations, ethics challenges, etc.) and also help in achieving the benefits of utilizing AI. Considering the clear tension between policy responses and freedom to innovate, they advocate for planning in order to amplify the need to minimize the perils of AI. Just as importantly, some universal prohibited AI practices must include manipulating people's decisions, exploiting vulnerabilities, classifying people based on personal traits or protected status, and predicting future crimes.

The European Union (EU) recently issued an extensive policy regulating the use of AI in deployed systems (The EU Artificial Intelligence Act, 2025). This is the first significant policy on the broad use of AI in history and it most certainly will not be the last. The EU is a supranational political and economic union of 27 member states that are located primarily in Europe. The focus of the policy is on promoting the use of AI in a way that protects the health, safety, and fundamental rights of EU citizens – it protects by addressing:

- Health: restrict AI systems that have the capacity to cause harm
- Safety: restrict AI systems that manipulate through false or misleading information
- Fundamental Rights: restrict AI systems that classify individuals based on protected status, make predictions about future crimes

The DoD is addressing AI in numerous ways as well as the Military and US national legislation regarding AI usage is likely coming – it is only a matter of time. EU policy on AI will likely be used as a template, as it is currently influencing corporate policies. The guidelines in this policy are informative considerations as we increase our own awareness, understanding, compliance, and ethics in using AI in the Military, Industry, and Academia environments.

The Role of Generative AI: Special Considerations and Next Steps for the Military, Industry, and Academia

Generative AI pertains to a category of AI, focusing on producing novel content, instead of merely examining or manipulating pre-existing data. This section's discussion focuses on special considerations for Generative AI across the military (Army/Navy), Industry, and Academia and includes next steps for future research.

Overcoming Military/DoD Challenges in Applying Generative AI

In 2020, the US DoD became the first military in the world to adopt ethical principles for the use of AI in military operations (CDAO, Responsible AI, 2025). These principles were based on recommendations from the Defense Innovation Board. The five DoD AI ethical principles are Responsible, Equitable, Traceable, Reliable, and Governable. The Chief Digital and Artificial Intelligence Office (CDAO) leads these efforts to operationalize the commitment to Responsible AI (RAI) by creating policy, tools, and training for DoD components. The CDAO also participates in interagency and international RAI efforts to further learning and share best practices across multiple organizations to ensure Government personnel have the resources and tools to assess whether AI systems are safe, reliable, and trustworthy. The five principles for the ethical development of AI capabilities apply to both combat and non-combat functions and assist the US military in upholding legal, ethical and policy commitments in the field of AI. Further, the focus on RAI led to the development of six foundational RAI tenets: RAI Governance, Warfighter Trust, AI Product and Acquisition Lifecycle, Requirements Validation, Responsible AI Ecosystem, and AI Workforce (U.S. Department Responsible Artificial Intelligence Strategy and Implementation Pathway, 2022). Finally, the initial principles and subsequent tenets led to the AI Adoption Strategy in 2023. This strategy outlines the approach to strengthening the organizational environment in order to continuously deploy data analytics and AI capabilities for decision advantage.

Following this strategy leads to decision advantage outcomes such as superior battlespace awareness and understanding; adaptive force planning and application; fast, precise and resilient kill chains; resilient sustainment support; and efficient enterprise business operations ([Data, Analytics and Artificial Intelligence Adoption Strategy, 2023](#)). Furthermore, in order to reach these outcomes, overarching goals should focus on investing in interoperable, federated infrastructure; advancing the data, analytics and AI ecosystem; expanding digital talent management; improving foundational data management; delivering capabilities for the enterprise business and joint warfighting impact; and strengthening governance and removing policy barriers ([Data, Analytics and Artificial Intelligence Adoption Strategy, 2023](#)). These DoD investments in data, analytics, and AI address key operational problems identified, fill validated gaps to enhance warfighting capabilities of the Joint Force, and strengthen the enterprise foundation required to sustain critical enduring advantages. Strengthening decision advantage for the Department's warfighting and business operations is one key to maintaining a resilient future force that can address a far reaching array of operational problems, dynamically promote and deter, and overmatch in conflict, when necessary.

These principles, tenets, and outcomes appear throughout the special considerations and next steps discussed in the following Army, Navy, Industry, and Academia sections.

Overcoming Army Challenges in Applying Generative AI to Defense Training Systems

The US Army faces a set of interdependent challenges when considering enterprise-wide adoption of Generative AI into training, education and force development strategies. These challenges span technical, organizational, and operational dimensions, requiring coordinated mitigation strategies that align with DoD's Responsible AI principles while enabling the strategic use of Generative AI for Soldier readiness.

Challenge 1: Reliability and Trust in AI-Generated Content

Generative AI systems can produce hallucinations (i.e., plausible sounding but factually incorrect outputs) that pose significant risks in military training contexts where accuracy is critical. Warfighters must be able to trust that AI-generated scenarios, feedback, and instructional content reflect doctrinally aligned information and operationally relevant conditions. The Army's emphasis on reliable AI, as articulated in

the DoD AI ethical principles, demands that these services demonstrate consistent, verifiable performance before deployment in training environments.

Addressing this challenge requires building validation pipelines that combine automated verification with subject matter expert review. Domain-specific fine-tuning using authoritative Army publications such as Field Manuals and Army Training Publications can anchor AI outputs in established doctrine. Equally important is creating feedback mechanisms that allow instructors to flag and correct errors, enabling continuous improvement while building the evidence base needed to establish institutional and operational trust over time.

Challenge 2: Knowledge Acquisition for Scalable Content Generation

Effective AI-driven tutoring requires accurate representations across military performance and competence (i.e., knowledge, skills, behavior) definitions, including tactical decision-making frameworks, equipment specifications, and operational procedures that general-purpose language models lack in their training data. Compounding this is the need to scale any approach across all echelons. The Army's training enterprise encompasses thousands of Military Occupational Specialties (MOSs), each with unique competency requirements. Creating AI systems capable of generating accurate, contextually appropriate content, at the correct level of granularity, requires careful schematic structuring of domain knowledge that can extend across training audiences.

A path forward involves developing retrieval-augmented generation architectures that ground AI outputs in authoritative doctrine and validated training curricula. Knowledge graphs and ontologies can structure military domain expertise in formats accessible to Generative AI, but there will need to be careful consideration in maintaining security classifications. Automated scenario generation tools driven by specific training requirements can then enable rapid creation of varied, doctrinally accurate training materials, with instructional designers guiding and refining outputs through purpose-built authoring interfaces.

Challenge 3: Integration with Training Environments

The Army's Synthetic Training Environment (STE) and Live, Virtual, and Constructive (LVC) training capabilities present complex integration challenges for rapidly maturing Generative AI. Existing platforms may lack the interfaces or real-time capabilities needed for AI-powered instruction.

Meeting this challenge requires developing standardized APIs (Application Programming Interfaces) and middleware that enable Generative AI services to interface with various training platforms while adopting open standards such as xAPI for learning technology interoperability. Additionally, agent architectures should combine Generative AI for natural language interaction with validated tactical behaviors derived from doctrine and Subject Matter Expert (SME) input. Configurable difficulty levels can allow agents to adapt to learner proficiency, while rigorous human-subjects research evaluates both agent realism and training effectiveness.

Challenge 4: Security and Data Protection

Army training applications frequently involve sensitive information, including operational tactics, equipment vulnerabilities, and learner performance data. Generative AI systems trained on or exposed to such data present novel security concerns (e.g., risks of data leakage through model outputs, adversarial attacks designed to extract training data, and challenges in deploying AI capabilities on classified

networks). Compliance with DoD cybersecurity frameworks while leveraging cloud-based AI services adds additional complexity.

To address this, secure deployment architectures that enable on-premise or air-gapped AI inference are essential for sensitive applications. Privacy-preserving techniques including differential privacy and federated learning can protect training data, while regular adversarial testing and red-teaming exercises identify vulnerabilities before exploitation. Clear data governance policies must control what information is used for AI training and how learner data is protected throughout the AI lifecycle, ensuring compliance with NIST 800-171 and CMMC (Cybersecurity Maturity Model Certification) requirements.

Challenge 5: Personalized and Adaptive Training at Scale

The Army's vision for adaptive training requires AI systems capable of assessing individual Soldier performance, identifying knowledge gaps, and dynamically adjusting training content and difficulty. Generative AI can enable personalization through tailored feedback, explanations, and novel practice scenarios. However, effective personalization requires accurate learner modeling, which depends on integrating performance data from multiple training systems and exercises, a significant data infrastructure challenge that spans organizational and technical boundaries. Realizing this vision depends on integrated performance analysis frameworks that aggregate data across training events to build comprehensive learner profiles. Adaptive instructional algorithms can then use Generative AI to produce personalized learning pathways responsive to individual or squad needs, implementing competency-based progression models that target specific skill deficiencies with appropriate remediation. Throughout, adaptive systems must remain aligned with Army training doctrine and unit readiness requirements to ensure personalization serves operational objectives.

Addressing these challenges requires sustained investment in research and development, close collaboration between the Army research community, industry partners, and academic institutions, and iterative deployment approaches that allow learning from early implementations. By proactively mitigating these risks while pursuing the transformative potential of Generative AI, the Army can enhance training effectiveness, reduce development costs, and better prepare Soldiers for the complexities of multi-domain operations.

Overcoming Navy Challenges in Applying Generative AI to Defense Training Systems

The Navy is investing in AI in several domains ranging from the well-publicized Amelia, used for IT help desk support (Demarest, 2023), to decision support and display optimization for Anti-Submarine Warfare Operators (Naval Air Warfare Center Training Systems Division, 2024). In the training domain, there are efforts to utilize computer vision techniques to augment tactical system displays for training purposes (Schroeder et al., 2024). The Surface Micro-Adaptive Rules of the Road Training technical candidate being sponsored by the Office of Naval Research is developing an AIS (adaptive instructional system) to improve a navigation team's understanding and application of Nautical Rules of the Road (RoR) (Steinhauser, 2024). This program of research is utilizing ML to develop several capabilities including scenario generation, ship agent driving behaviors, and adaptive instruction. ML approaches are typically ideal solutions because they can model aspects of a task beyond what an instructor might reasonably be able to detect or articulate, such as identifying subtle patterns across very large sets of learner data (Schroeder et al., in press). However, typically military training applications are fraught with the small data set issue which makes utilizing ML particularly challenging.

Indeed, Karimi et al.'s (2024) review found that AISs using small data sets tended to focus on classroom education versus complex decision-making tasks typical of military training systems. Having a small amount of data to train models can lead to developing poor training systems with the potential of negative training outcomes which is why these complex domains are traditionally avoided. AI hallucinations highlight this issue. AI hallucinations are erroneous, false, or misleading responses generated from AI and eliminating them is virtually impossible (Jones, 2025). The giraffe problem (also known as giraffing) is an example of AI hallucinations. This well-known phenomenon occurs when AI has a misleading view on how often giraffes appear in real life (due to high number of giraffe pictures that are uploaded to the internet) and often confuses long vertical lines in fences, trees, and even lighthouse images as giraffes. This highlights the importance of the quality and quantity of the training data used to generate AI models.

When AI hallucinations occur, it can cause a significant negative training and/or operational impact to military personnel. Therefore, it is important to not introduce Generative AI into systems too early without rigorous validation efforts. Early adoption of these models without proper validation can lead to abandonment of the technology, learning inaccurate information, and trust issues. One area that requires further investigation is determining if our scientific knowledge and theory in trust in automation applies to trust in Generative AI.

The other challenge unique to military training is when we should invest in Generative AI solutions and when we should not. Not all learning tasks require or benefit from Generative AI models. Many times, simple rule or heuristic based algorithms and/or basic AI are sufficient. Therefore, system designers must determine whether the return on investment of these sophisticated models outweighs the learning outcomes achieved via techniques such as training effectiveness evaluations and value-added comparisons studies (Schroeder et al., in press).

Overcoming Industry Challenges in Applying Generative AI to Defense Training Systems

From the perspective of a training technology company that develops AI-powered solutions for the DoD, addressing the unique challenges of implementing Generative AI is essential. Below are a list of key challenges along with strategies for mitigating risks associated with the integration of Generative AI solutions into training systems.

Challenge 1: Data Security and Compliance with Regulations

Generative AI systems must operate within a secure environment to protect sensitive and classified data. To achieve this, a robust data pipeline should be established, featuring encryption, role-based access control, and detailed audit logs. These measures should safeguard data at every stage, from ingestion to deployment. Furthermore, all AI systems should undergo thorough security audits and penetration tests to identify and address vulnerabilities. Compliance with DoD cybersecurity frameworks, such as NIST 800-171 and CMMC, should be strictly enforced to maintain operational security.

DoD contractors face strict requirements for handling sensitive and classified information, and data breaches can result in severe penalties, loss of contracts, or reputational damage. By implementing secure data pipelines, encryption, role-based access controls, and adhering to frameworks like NIST 800-171 and CMMC, contractors reduce the likelihood of unauthorized access and data leakage. Regular audits and compliance ensure that systems meet DoD security standards, protecting the contractor from legal liabilities and reputational harm.

Challenge 2: Building Trust

For AI systems to be trusted by operators, their decision-making processes must be transparent. This involves employing interpretable AI techniques to provide insights into how outputs are generated. Documentation and visual aids should be created to help users understand these systems. Additionally, a feedback mechanism should be implemented, allowing users to report any erroneous or unexpected outputs. Human-in-the-loop reviews may be used to validate AI-generated content, especially in mission-critical scenarios, to ensure its reliability and accuracy.

If military personnel and decision-makers lack confidence in AI outputs, systems may face limited adoption, rendering investments in technology futile. Explainable AI techniques, transparent documentation, and human-in-the-loop reviews ensure that end-users understand and trust the AI's decision-making process. Establishing feedback mechanisms builds confidence in the system, reducing the risk of poor adoption rates and potential project cancellations.

Challenge 3: Addressing Bias and Ethical Use

AI systems are only as unbiased as the data they are trained on. To reduce bias, datasets should undergo rigorous audits, and use techniques like *adversarial debiasing*, a ML approach used to reduce or eliminate bias in models while maintaining high predictive accuracy (Zhang et al., 2018). This approach leverages the concept of adversarial training, where two models—a primary model and an adversary model—are trained simultaneously in a competitive framework. The goal is to produce predictions that are not only accurate but also independent of sensitive attributes (e.g., age). An ethics review board should oversee the development and deployment of Generative AI capabilities, ensuring their use aligns with DoD values. Ethical training should be provided to developers and users, highlighting potential unintended consequences and the importance of maintaining ethical standards in military applications.

However, Generative AI bias may or may not include ethical violations. For example, the over-representation of certain threat profiles produced by an automated scenario generation process could disproportionately include historical missions involving adversaries from specific regions, such as the Middle East or Southeast Asia. As a result, the Generative AI could predominantly generate scenarios featuring these threat profiles, environments, and cultural contexts. The impact of this type of bias could be narrower operational training and reduced preparedness for emerging threats. Trainees might develop proficiency in handling threats specific to certain regions but lack exposure to other potential adversaries. The system might fail to prepare Warfighters for nontraditional or evolving threats, such as cyber warfare, space-based operations, or humanitarian missions.

Unintended biases or unethical AI use could lead to operational failures or public scrutiny, damaging both the contractor's reputation and its relationship with the DoD. To mitigate biases or unethical use of AI, training system designers should diversify training data by including data from a wide range of military operations across different regions, terrains, and adversary profiles. Designers should also incorporate emerging threat scenarios. Another mitigation strategy involves using SMEs to validate AI-generated scenarios and ensure they reflect a broad spectrum of operational contexts. Of course, a methodology to detect bias in real-time could also be helpful in flagging potential biases during scenario generation.

Conducting bias audits, using debiasing techniques, and implementing oversight through ethics review boards help prevent ethical violations. Providing training on responsible AI use demonstrates a proactive stance, mitigating the risk of reputational damage and maintaining public trust.

Challenge 4: Ensuring Realism and Operational Relevance

Generative AI capabilities must be able to produce realistic and operationally sound outputs. If Generative AI systems produce unrealistic or operationally irrelevant outputs, they could undermine mission readiness, leading to costly rework or project delays. To guarantee realistic and relevant outcomes, SMEs should validate AI-generated scenarios and training materials. The models should be regularly updated and calibrated to reflect evolving tactics, techniques, and procedures (TTPs). AI scenario generators should also include safeguards to prevent the creation of unrealistic or tactically inappropriate situations. Validating AI-generated content with SMEs and continuously updating models to reflect evolving TTPs ensures operational relevance. This minimizes the risk of delivering ineffective or obsolete solutions that could jeopardize contract performance and future opportunities.

Challenge 5: Optimizing Performance and Scalability

AI systems deployed in military training environments must deliver consistent performance, even under demanding conditions. To achieve this, modular system architectures should allow for seamless updates and model improvements. Performance benchmarking and stress testing should ensure systems meet DoD training requirements. Leveraging distributed computing and cloud resources should enable scalability, supporting large-scale training exercises without compromising performance.

Generative AI systems that cannot scale or perform reliably under operational stress may fail to meet contractual requirements, leading to penalties or loss of trust with the DoD. Performance benchmarking, stress testing, and the use of scalable architectures reduce the risk of system failure during critical missions. Ensuring reliable performance protects the contractor from non-compliance with service-level agreements (SLAs) and potential contract disputes.

Challenge 6: Supporting Continuous Learning and Adaptability

Continuous learning should be integral to maintaining the relevance and effectiveness of Generative AI systems. These systems should incorporate mechanisms for reinforcement learning, allowing them to adapt based on real-world feedback. Adaptive training capabilities should tailor content to individual learners while maintaining consistency with operational goals. Transfer learning techniques should further enable rapid adaptation of AI to new scenarios or training domains.

Challenge 7: Facilitating User Adoption and Training

The success of any technology depends on its ease of use and the proficiency of its operators. User-friendly interfaces should simplify interaction with Generative AI systems, while tailored training programs should equip users with the skills needed to maximize the technology's potential. Additionally, a dedicated support system should address technical and operational concerns, ensuring smooth adoption and effective use.

A poorly designed or overly complex system may lead to user frustration, limiting the system's value and jeopardizing contract renewals. User-friendly interfaces, tailored training, and robust support services encourage user adoption and maximize the technology's impact. By addressing usability, contractors mitigate the risk of user dissatisfaction.

Challenge 8: Navigating Legal and Policy Violations

Compliance with legal and policy frameworks is critical in defense applications of AI. Collaboration with legal experts should ensure adherence to US and international laws governing AI use in military contexts.

Participation in inter-agency working groups should help shape policies and standards for Generative AI in defense. Moreover, clear documentation should outline the appropriate use cases and limitations of these systems, fostering responsible deployment.

Failure to comply with legal and policy requirements could result in fines, contract termination, or being barred from future DoD contracts. Collaborating with legal experts and participating in inter-agency policy discussions ensures that AI systems comply with all relevant laws and regulations. Clear documentation of system limitations and proper use cases reduces the risk of accidental misuse or non-compliance.

Challenge 9: Defending Against Adversarial Threats

Generative AI systems must be prepared to counter adversarial threats. Adversarial testing should be conducted to identify potential vulnerabilities. Defensive mechanisms should be implemented to detect and counter malicious inputs designed to deceive AI models. A rapid response team should be established to address emerging threats, ensuring the resilience of AI systems in contested environments.

AI systems that are vulnerable to adversarial attacks or exploitation could compromise national security and lead to severe consequences for the contractor. Adversarial testing and the development of defensive mechanisms safeguard against malicious attacks. A rapid response team ensures vulnerabilities are addressed swiftly, minimizing the risk of security breaches.

Challenge 10: Promoting Interoperability and Standardization

Finally, interoperability with existing and future DoD systems is essential. Generative AI systems should be designed to adhere to open standards, facilitating seamless integration across platforms. Collaboration with allied nations and defense partners should promote interoperable solutions, enhancing joint training capabilities. APIs and middleware should be developed to ensure smooth communication between Generative AI and other training technologies.

Non-interoperable solutions create integration challenges, delay deployments, and increase the risk of project failure. Designing AI systems to adhere to open standards and collaborating with allied nations ensures interoperability. Developing APIs and middleware for seamless integration minimizes the risk of costly rework or integration delays.

Demonstrating a commitment to security, trust, ethics, and performance builds strong relationships with the DoD. Furthermore, mitigating these risks protects the company from legal, financial, and reputational consequences, ensuring long-term growth and sustained partnerships.

Overcoming Academia Challenges in Applying Generative AI

The integration of Generative AI into postsecondary education presents several pedagogical concerns, primarily revolving around academic integrity, deep learning, and alignment with learning objectives. AI tools can enable students to generate assignments without engaging in critical thinking, leading to concerns about plagiarism and overreliance on automated responses. Additionally, AI-generated content may not always align with course objectives, making it difficult for instructors to ensure students are genuinely developing key competencies. To address these issues, institutions can implement AI-assisted assignments that encourage critical engagement rather than rote responses, integrate AI-detection tools to uphold academic integrity, and train faculty to guide students in using AI as a learning partner rather than a substitute. By thoughtfully incorporating AI in pedagogical practices, institutions can harness its potential while maintaining rigorous academic standards Mishra et al. (2025).

Beyond pedagogical concerns, instructor readiness remains a major barrier, as many faculty members lack AI literacy and are resistant to adopting AI-powered tools. This resistance is often driven by fears that AI will replace traditional teaching methods or require significant time investments to learn and integrate effectively. Additionally, students face their own challenges, such as unequal access to AI tools, the risk of overreliance on AI-generated content, and difficulties in critically evaluating AI outputs. These issues can be mitigated through professional development programs that enhance AI literacy among instructors, faculty workshops that promote AI experimentation, and the provision of institutionally supported AI tools to ensure equal access for all students. Furthermore, guidelines for responsible AI use, along with AI-integrated assessments that require human reflection, can help students develop critical evaluation skills while using AI ethically and effectively.

Technical and ethical challenges further complicate AI adoption in higher education. Issues such as AI accuracy, integration with Learning Management Systems (LMS), and lack of customization can hinder widespread use. Additionally, concerns about AI bias, data privacy, and intellectual property rights raise ethical and legal questions that institutions must navigate. To address these challenges, educational institutions must invest in AI tools that prioritize accuracy, transparency, and easy integration with existing systems. They should also establish fairness audits to mitigate AI biases, enforce strict data privacy policies, and educate faculty and students about intellectual property considerations when using AI-generated content. Finally, at an institutional level, the absence of clear AI policies, outdated assessment models, and limited funding for AI adoption remain significant barriers. Universities can respond by developing clear policies on AI use, revising accreditation and assessment models to accommodate AI-assisted learning, and allocating resources to support faculty and students in leveraging AI effectively.

Arizona State University (ASU) has proactively addressed many of these challenges through a series of comprehensive initiatives (see ai.asu.edu). Recognizing the importance of faculty preparedness, ASU has developed professional development programs aimed at enhancing AI literacy among educators. The "**Teaching and Learning with Generative AI**" self-paced course, designed by university experts, supports faculty and staff in effectively incorporating AI into their teaching practices. Additionally, ASU has established the **Generative Artificial Intelligence Community of Practice**, which fosters collaboration among faculty to explore and implement creative, ethical approaches to AI integration in education. These initiatives help instructors gain confidence in using AI as a teaching tool while ensuring alignment with best pedagogical practices.

To address technical and implementation barriers, ASU has invested in developing user-friendly AI tools that integrate seamlessly with existing educational platforms (Ahmed et al.,2026). The **MyAI Builder** platform empowers faculty and staff to create custom AI experiences, such as Generative AI-powered chatbots, tailored to specific course needs. This tool simplifies the development process, enabling users without extensive technical backgrounds to design and implement AI solutions that enhance the learning experience. Furthermore, MyAI Builder incorporates an **Ethical AI Engine**, which provides safeguards and guidelines to ensure AI-generated content aligns with ethical standards, institutional policies, and responsible AI principles. By embedding ethical considerations into AI development, ASU is fostering a culture of responsible AI use that prioritizes fairness, transparency, and academic integrity.

ASU has also taken significant steps to establish clear policies and guidelines for the ethical use of AI in education. The university's **collaboration with OpenAI** reflects its commitment to responsible AI integration, focusing on leveraging AI to enhance teaching, learning, and research while upholding academic integrity. ASU's AI strategy emphasizes **principled innovation**, prioritizing equity, privacy, and human-centered design to ensure that AI applications deliver social benefits and respect individual rights. By setting clear ethical standards and ensuring faculty, staff, and students have the necessary resources to use AI responsibly, ASU is paving the way for a more effective and equitable approach to Generative AI

in higher education. Through these initiatives, ASU demonstrates a comprehensive approach to overcoming the barriers associated with Generative AI in postsecondary education, fostering an environment where AI enhances learning without compromising ethical standards or educational quality.

Alignment Across Military, Industry, and Academia

Several shared themes and challenges emerge when applying Generative AI across military, industry, and academia. There are clear parallels in the challenges and vulnerabilities faced when applying Generative AI in military, industry, and higher education settings—despite the vastly different domains—suggesting a set of shared structural, ethical, and operational concerns:

1. **Trust and Transparency:** Building Warfighter confidence hinges on systems being explainable, ethical, and traceable. This aligns closely with the DoD's focus on *Warfighter trust* and ethical AI principles like responsibility, traceability, and governability. Building confidence in AI systems requires explainability, traceability, and responsible design. For the military, this ensures Warfighter trust in mission-critical tools; in education, it ensures academic integrity and ethical use.
2. **Security, Legal, and Policy Compliance:** Challenges around data security, legal constraints, and ethical AI use are fundamentally about protecting mission integrity and national security. These concerns are echoed in the DoD's Responsible AI governance structures. Whether protecting sensitive military data, personnel hiring records, or safeguarding student privacy and intellectual property, data security and ethical stewardship of AI-generated content remain universal challenges. Legal compliance and clear policy guidance are foundational in each environment.
3. **Robustness and Operational Relevance:** Ensuring realism, performance, and resistance to adversarial threats reflects the DoD's push toward operational readiness. The emphasis on validated, mission-relevant AI solutions mirrors efforts in Navy and Army (and DoD) training systems. Ensuring AI outputs are realistic and contextually meaningful is vital for both warfighting scenarios, industry practices, and curriculum alignment. Mismatched or hallucinated content can lead to poor training or academic disengagement.
4. **Adaptability, Continuous Learning and System Maturity:** Both the Army, Navy and broader DoD strategy emphasize sustaining AI effectiveness over time. Ongoing AI learning loops ensure systems evolve alongside shifting operational needs. Both sectors stress the need for continuous system improvement and stakeholder feedback. Whether it is updating training models for evolving mission or tasks, business goals, or revising AI tools to support course evolution, adaptability is key.
5. **Human-System Integration:** Challenges in user adoption and standardization are paralleled in the strategic investment in infrastructure and talent. Ease of use, training, and interoperability are crucial to scale AI responsibly across services. Human-system integration is essential to ensure usability, adoption, and scale.

6. **Bias and Ethical Use:** The risk of biased outputs affects these domains. Military applications face operational risks, while in industry and education, biased AI content can undermine fairness and learning outcomes. Both call for rigorous evaluation and fairness audits.
7. **User Readiness and Adoption:** Just as instructors may lack AI literacy in postsecondary education, service members or senior staff may struggle with unfamiliar interfaces or mistrust.
8. **Institutional Support and Infrastructure:** Military, industry, and academia often underscore the importance of investing in AI infrastructure, user-friendly tools, and professional development to ensure responsible and effective AI deployment.

Together, these connections reveal that the technical hurdles of deploying Generative AI are not isolated—they are deeply tied to cultural, ethical, and strategic priorities shared across these ecosystems. In essence, while the stakes and applications differ, the blueprint for successful Generative AI implementation that can be leveraged across each—trustworthiness, usability, ethics, and support—remains strikingly consistent.

Recommendations for GIFT and STEEL-R Overall

The Generalized Intelligent Framework for Tutoring (GIFT) provides a uniquely positioned software that can help address the challenges discussed throughout this chapter. As an open-source, modular AIS, GIFT offers service-oriented architectural capabilities that can enable accelerated integration of Generative AI into Army training systems. The following recommendations outline how GIFT and the Synthetic Training Environment Experimentation and Learning for Readiness (STEEL-R) (Goldberg et al., 2021) initiative can be leveraged to advance Generative AI adoption across military, industry, and academic contexts.

Recommendations

Recommendation 1: Establish GIFT as a Generative AI Integration and Interoperability Platform

GIFT's service-oriented architecture and well-defined interfaces position it as an ideal integration platform. The framework should be extended to include standardized APIs for connecting to LLMs, enabling GIFT-based tutors to leverage Generative AI for natural language dialogue, content generation, and intelligent feedback while maintaining the pedagogical control structures that ensure effective instruction. This integration should support both cloud-based and localized deployments to accommodate varying security requirements across classification levels.

Equally important is leveraging GIFT's Gateway module and demonstrated interoperability with a wide range of training applications (e.g., Virtual Battlespace (VBS), Unity-based simulations, mobile interaction, PowerPoint-based instruction). STEEL-R should build on existing GIFT-STE integration efforts to prototype AI-powered intelligent agents, dynamic scenario generation, and real-time adaptive instruction, incorporating Generative AI capabilities for enhanced roleplayer dialogue, procedural scenario variation, and AAR support.

Recommendation 2: Leverage GIFT's Domain Module for Knowledge-Grounded AI and Content Validation

To address challenges of knowledge acquisition and AI reliability, GIFT's Domain Module should be enhanced to support retrieval-augmented generation architectures. By structuring authoritative domain knowledge that includes Army doctrine, training curricula, and validated procedures in formats to guide Generative AI, GIFT can ground AI outputs in verified sources, reducing hallucinations and ensuring doctrinal accuracy. The concept library and misconception libraries already supported in GIFT provide a foundation for identifying and correcting AI-generated content that contradicts established knowledge.

Building on this foundation, GIFT should incorporate validation pipelines that assess AI-generated content before presentation to learners, including automated checks against doctrinal constraints, confidence scoring mechanisms, and SME review workflows. GIFT's authoring tools can be extended to support human-AI collaborative content development, enabling instructional designers to efficiently review, refine, and approve AI-generated materials while maintaining the quality standards essential for building and sustaining end user trust.

Recommendation 3: Extend GIFT's Learner Module for AI-Enhanced Personalization

GIFT's existing learner modeling capabilities should be enhanced to support Generative AI-driven personalization at scale. The Learner Module's competency tracking and performance assessment features can inform Generative AI systems about individual learner states, enabling more targeted content generation and adaptive feedback that responds to specific knowledge gaps and skill deficiencies. Integration with xAPI and other learning data standards will allow GIFT to aggregate performance information from varied training systems, creating comprehensive learner profiles that drive sophisticated personalization across the training enterprise while ensuring adaptive systems remain aligned with Army training doctrine and unit readiness requirements.

Recommendation 4: Implement Responsible AI Governance Through GIFT's Architecture

GIFT should incorporate explicit mechanisms for implementing DoD's Responsible AI principles throughout the AI lifecycle. This includes traceability features that log AI decision-making processes for audit and review, bias detection tools that monitor AI outputs for fairness across learner populations, and governance controls that enable administrators to configure AI capabilities according to organizational policies and security requirements. The framework's inherent extensibility allows these governance mechanisms to evolve as DoD's Responsible AI requirements mature and as new risks emerge from advancing AI capabilities, ensuring GIFT remains aligned with the DoD's commitment to ethical, reliable, and governable AI systems.

Recommendation 5: Enable Research, Evaluation, and Community Development

GIFT's embedded data collection and experiment management capabilities position it as a great platform for empirical research on the impact and return on investment of Generative AI in adaptive learning contexts. STEEL-R should leverage these capabilities to conduct rigorous training effectiveness evaluations comparing AI-enhanced instruction to traditional approaches, with research protocols that assess not only learning outcomes but also user trust, workload impacts, and potential unintended consequences of AI-assisted instruction. The framework should support A/B testing of different Generative AI configurations, enabling evidence-based decisions about AI deployment.

Complementing these research efforts, GIFT’s open-source model and established community of military, academic, and industry contributors provide a vehicle for collaborative advancement of Generative AI in training. The GIFT community should develop shared resources including prompt libraries, domain knowledge repositories, and validated AI integration patterns that accelerate adoption while promoting best practices. Regular workshops and publications can disseminate lessons learned, enabling the broader AIS community to benefit from experimentation conducted within GIFT and STEEL-R.

By implementing these recommendations, GIFT and STEEL-R can serve as catalysts for the responsible adoption of Generative AI in defense training, providing the technical infrastructure, validation frameworks, and research capabilities needed to realize the potential of AI while mitigating the risks identified throughout this chapter. The alignment between GIFT’s architectural principles (i.e., modularity, interoperability, and extensibility) and the requirements for trustworthy AI integration positions the framework as an enabler for the next generation of intelligent tutoring systems operating within industry, academia, and the military.

Conclusions

AI is rapidly transforming the landscape of training and operational readiness across the military, industry, and academia. Its integration into these domains—through tools like intelligent tutoring systems, predictive analytics, and generative models—offers unprecedented opportunities to personalize learning, optimize decision-making, and enhance performance. Yet, as AI becomes more embedded, the urgency to address its ethical, legal, and operational implications grows.

Across sectors, shared challenges such as trust, transparency, security, adaptability, and human-system integration underscore the need for a more thoughtful, and ultimately, more unified approach to AI governance and deployment. The emergence of comprehensive policies like the EU’s Artificial Intelligence Act signals a global shift toward responsible AI use—one that prioritizes health, safety, and fundamental rights. These principles are increasingly reflected in DoD strategies and are influencing corporate and academic practices alike.

To harness AI’s full potential while mitigating its risks, stakeholders must commit to continuous learning, robust infrastructure investment, and ethical oversight. The convergence of concerns across domains reveals that successful AI implementation is not just a technical endeavor—it is a cultural and strategic imperative. By aligning innovation with responsibility, we can ensure that AI remains a powerful ally in shaping the future of training, education, and national security.

Chapter Appendix

The tables on the following pages are summaries of the Navy, Army, Industry, and Academia Considerations and Next Steps.

Navy				
Special Considerations/Affordances for Training Applications	Potential Barriers	Next Steps	Risks	Benefits
<i>Models must have high fidelity to provide training value.</i>	Military training applications can be fraught with a small data set.	Utilizing computer vision techniques to augment tactical system displays for training purposes. Increase fidelity of data. Utilize machine learning. Understand the importance of the quality and quantity of the training data used to generate AI models.	AISs using small data sets tend to focus on classroom education vs complex decision-making tasks typical of military training systems.	Machine learning approaches are ideal solutions as they can model aspects of a task beyond what an instructor might be able to detect or articulate.
<i>Introducing Generative AI into systems at the right time – knowing when to invest for impact.</i>	Introducing AI too soon. Introducing AI without rigorous validation efforts.	System designers must determine whether the Return on Investment (ROI) outweighs the learning outcomes achieved using validation effort techniques such as training effectiveness evaluations and value-added comparisons studies.	Small amount of data to train models can lead to developing poor training systems with the potential of negative training outcomes. Lack of valid data can lead to AI hallucinations, causing a significant negative impact to military personnel.	Ability to model aspects of a task beyond what an instructor might reasonably be able to detect or articulate.
<i>Determining if our scientific knowledge and theory in trust and confidence in automation applies to trust in Generative AI.</i>	Must be realistic, have explainability, and understandability. Trust must be built.	Understand what performance guarantees there are when you change underlying assumptions. Ensure understandability and brevity of generated content.	Abandonment of the technology, learning inaccurate information, and trust issues.	Identifying subtle patterns across very large sets of learner data.

Army			
Special Considerations/Affordances for Training Applications	Potential Barriers	Next Steps	Benefits
<i>Integration with Training Environments</i>	<p>Platforms lacking the interfaces or real-time capabilities needed for AI-powered instruction.</p>	<p>Develop standardized APIs.</p> <p>Agent architectures should combine Generative AI with validated tactical behaviors derived from doctrine and SME input.</p> <p>Configurable difficulty levels.</p> <p>Rigorous human-subjects research.</p>	<p>Enables Generative AI services to interface with training platforms while adopting open standards for learning technology interoperability.</p> <p>Natural language interaction.</p> <p>Allows agents to adapt to learner proficiency.</p> <p>Evaluates both agent realism and training effectiveness.</p>
<i>Personalized and Adaptive Training at Scale</i>	<p>Requires accurate learner modeling, depending on integrating performance data from multiple training systems and exercises.</p> <p>Requires sustained investment in research and development, close collaboration between the Army research community, industry partners, and academic institutions, and iterative deployment approaches.</p>	<p>Integrate performance analysis frameworks that aggregate data across training events.</p> <p>Adaptive instructional algorithms using Generative AI.</p> <p>Align with Army training doctrine and unit readiness requirements.</p>	<p>Comprehensive learner profiles.</p> <p>Enhances training effectiveness, reduces development costs, and better prepares Soldiers for the complexities of multi-domain operations.</p> <p>Personalized learning pathways responsive to individual or squad needs implementing competency-based progression models that target specific skill deficiencies with appropriate remediation.</p> <p>Personalization serves operational objectives.</p>

Army (Continued)

Special Considerations/Affordances for Training Applications	Potential Barriers	Next Steps	Benefits
<p><i>Security and Data Protection</i></p>	<p>Training applications frequently involve sensitive information, including operational tactics, equipment vulnerabilities, and learner performance data.</p> <p>Compliance with DoD cybersecurity frameworks while leveraging cloud-based AI services.</p>	<p>Secure deployment architectures that enables on-premise or air-gapped AI inference.</p> <p>Engage in Privacy-preserving techniques.</p> <p>Regular adversarial testing and red-teaming exercises</p> <p>Clear data governance policies to control information used for AI training and how learner data is protected throughout the AI lifecycle, ensure compliance with NIST 800-171 and CMMC.</p>	<p>Safeguard sensitive applications.</p> <p>Protect training data.</p> <p>Identify vulnerabilities before exploitation.</p>
<p><i>Knowledge Acquisition for Scalable Content Generation</i></p>	<p>General-purpose language models lack accurate representations across military performance and competence definitions in their training data.</p> <p>Need to scale any approach across all echelons.</p> <p>Army’s training enterprise encompasses thousands of Military Occupational Specialties (MOSs), each with unique competency requirements.</p>	<p>Careful schematic structuring of domain knowledge that can extend across training audiences.</p> <p>Developing retrieval-augmented generation architectures.</p> <p>Knowledge graphs and ontologies.</p> <p>Automated scenario generation tools driven by specific training requirements.</p>	<p>AI systems capable of generating accurate, contextually appropriate content, at the correct level of granularity.</p> <p>Grounded AI outputs in authoritative doctrine and validated training curricula.</p> <p>Structured military domain expertise in formats accessible to Generative AI.</p> <p>Rapid creation of varied, doctrinally accurate training materials. Instructional designers guiding/refining outputs through purpose-built authoring interfaces.</p>

Army (Continued)

Special Considerations/Affordances for Training Applications	Potential Barriers	Next Steps	Benefits
<p><i>Reliability and Trust in AI-Generated Content</i></p>	<p>Hallucinations.</p> <p>Trust that AI-generated scenarios, feedback, and instructional content reflect doctrinally aligned information and operationally relevant conditions.</p>	<p>Demonstrate consistent, verifiable performance before deployment in training environments.</p> <p>Build validation pipelines that combine automated verification with SME review.</p> <p>Use authoritative Army publications such as Field Manuals and Army Training Publications for domain-specific fine-tuning.</p> <p>Create feedback mechanisms that allow instructors to flag and correct errors.</p>	<p>Anchor AI outputs in established doctrine.</p> <p>Continuous improvement while building the evidence base needed to establish institutional and operational trust over time.</p>

Industry				
Special Considerations/Affordances for Training Applications	Potential Barriers	Next Steps	Risks	Benefits
<i>Facilitating User Adoption and Training</i>	<p>Poorly designed or overly complex system.</p> <p>Usability in general.</p>	<p>Design user friendly interfaces.</p> <p>Tailor training programs.</p> <p>Dedicate support system to address technical and operational concerns.</p>	<p>User dissatisfaction and frustration.</p> <p>Limited system value and jeopardization of contract renewals.</p>	<p>Ease of use and proficiency of operators.</p> <p>Simplified interaction with system.</p> <p>Users with skills needed to maximize technology's potential.</p> <p>Smooth adoption and effective use.</p> <p>Increase the likelihood of long-term partnerships.</p>
<i>Navigating Legal and Policy Violations</i>	<p>Compliance with legal and policy frameworks.</p>	<p>Collaborate with legal experts.</p> <p>Participate in inter-agency working groups.</p> <p>Ensure clear documentation, outlining the appropriate use cases and limitations of these systems.</p>	<p>Fines, contract termination, or being barred from future DoD contracts.</p> <p>Accidental misuse or non-compliance.</p>	<p>Adherence to US and international laws governing AI use in military contexts.</p> <p>Contribute to shaping policies and standards for Generative AI in defense.</p> <p>Foster responsible deployment.</p>

Industry (Continued)				
Special Considerations/Affordances for Training Applications	Potential Barriers	Next Steps	Risks	Benefits
<i>Defending Against Adversarial Threats</i>	Adversarial threats Vulnerability to adversarial attacks or exploitation.	Conduct adversarial testing. Implement defensive mechanisms. Establish a rapid response team.	Compromise national security and lead to severe consequences for contractors. Security breaches	Identification of potential vulnerabilities. Detection and countering of malicious inputs designed to deceive AI models. Addressing emerging threats, ensuring the resilience of AI systems in contested environments. Safeguard against malicious attacks. Vulnerabilities addressed swiftly. Maintaining DoD's confidence in contractor solutions.
<i>Promoting Interoperability and Standardization</i>	Interoperability with existing and future DoD systems.	Designed to adhere to open standards. Promote interoperable solutions when collaborating with allied nations and defense partners. Develop APIs and middleware.	Integration challenges, delay deployments, project failure or loss of future work. Costly rework or integration delays.	Facilitation of seamless integration across platforms. Enhancing joint training capabilities. Smooth communication between Generative AI and other training technologies. Interoperability

Industry (Continued)				
Special Considerations/Affordances for Training Applications	Potential Barriers	Next Steps	Risks	Benefits
<i>Ensuring Realism and Operational Relevance</i>	<p>Time consuming.</p> <p>Shifting operational relevance.</p>	<p>Increase fidelity of data validation by SMEs of AI-generated scenarios and training materials.</p> <p>Regularly update and calibrate models to reflect evolving tactics, techniques, and procedures (TTPs).</p> <p>Include safeguards in AI scenario generators to prevent the creation of unrealistic or tactically inappropriate situations.</p>	<p>Undermining mission readiness, leading to costly rework or project delays.</p> <p>Delivering ineffective or obsolete solutions that could jeopardize contract performance and future opportunities.</p>	<p>Realistic and operationally sound outputs.</p> <p>Ensure operational relevance.</p>
<i>Data Security and Compliance with Regulations</i>	<p>Identification of vulnerabilities.</p> <p>Buy in</p> <p>Adherence maintaining.</p>	<p>Establish robust data pipeline (encryption, role-based access control, and detailed audit logs).</p> <p>Security audits and testing on AI systems to identify and address vulnerabilities.</p> <p>Comply with DoD cybersecurity frameworks.</p>		<p>Safeguarded data from ingestion to deployment.</p> <p>Increased operational security.</p> <p>Industry protections from legal liabilities and reputational harm.</p>

Industry (Continued)				
Special Considerations/Affordances for Training Applications	Potential Barriers	Next Steps	Risks	Benefits
<i>Addressing Bias and Ethical Use</i>	<p>Time for audit.</p> <p>Acceptance of novel techniques.</p> <p>Resources for incorporating ethical standards and diversification.</p>	<p>Rigorous dataset audits.</p> <p>Use techniques like <i>adversarial debiasing</i>.</p> <p>Leverage adversarial training.</p> <p>Ethics review board, aligned with DoD values.</p> <p>Ethical training provided to developers and users.</p> <p>Diversify training data to include data from wide range of military operations.</p> <p>Incorporate emerging threat scenarios.</p> <p>Use SMEs to validate AI-generated scenarios.</p> <p>Develop a methodology to detect bias in real-time.</p>	<p>Only as unbiased as the data they are trained on.</p> <p>Generative AI bias may or may not include ethical violations leading to operational failures, public scrutiny, damaging contractor reputation and relationship with DoD.</p> <p>Developed proficiency in handling specific threats but lacking exposure to other potential adversaries.</p>	<p>Reduced or eliminated bias in models.</p> <p>Maintaining high predictive accuracy.</p> <p>Predictions are accurate and independent of sensitive attributes.</p> <p>Insight on the development and deployment of capabilities.</p> <p>Proactively decrease unintended consequences.</p> <p>Understanding importance of maintaining ethical standards in military applications.</p> <p>Mitigating the risk of reputational damage and maintaining public trust.</p>

Industry (Continued)				
Special Considerations/Affordances for Training Applications	Potential Barriers	Next Steps	Risks	Benefits
<i>Supporting Continuous Learning and Adaptability</i>	<p>Shifting operational goals.</p> <p>Time consuming</p>	<p>Incorporate mechanisms for reinforcement learning.</p> <p>Tailor content of adaptive training capabilities to individual learners while maintaining consistency with operational goals.</p> <p>Enable rapid adaptation of AI to new scenarios/training domains.</p>	<p>Decreased relevance and effectiveness of the system.</p>	<p>Maintaining the relevance and effectiveness of Generative AI systems.</p> <p>Adaptation based on real-world feedback.</p>
<i>Optimizing Performance and Scalability</i>	<p>Delivering consistent performance, even under demanding conditions.</p> <p>Integrated and compatible modular system architectures.</p> <p>Systems that cannot scale or perform reliably under operational stress.</p>	<p>Allow for seamless updates and model improvements in modular system architectures.</p> <p>Conduct performance benchmarking and stress testing to ensure DoD requirements met.</p> <p>Enable scalability with distributed computing and cloud resources.</p>	<p>Failing to meet contractual requirements, leading to penalties or loss of trust with the DoD.</p> <p>System failure during critical missions.</p>	<p>Supporting large-scale training exercises without compromising performance.</p> <p>Compliance with SLAs and less contract disputes.</p> <p>Ability to build trust and confidence in performance between partners.</p>

Industry (Continued)				
Special Considerations/Affordances for Training Applications	Potential Barriers	Next Steps	Risks	Benefits
<i>Building Trust</i>	<p>Establishing transparency.</p> <p>Realistic</p>	<p>Employing interpretable AI techniques (documentation and visual aids).</p> <p>Including feedback mechanism.</p> <p>Human-in-the-loop reviews</p>	<p>Limited adoption</p> <p>Futile technology investments.</p> <p>Lack of confidence</p>	<p>Insights into how outputs are generated.</p> <p>Reporting of any erroneous or unexpected outputs.</p> <p>Validated content and reliability/accuracy</p> <p>End-users understand and trust AI's decision-making process.</p> <p>Confidence built in the system.</p> <p>Reduced risk of poor adoption rates and potential project cancellations.</p>

Academia				
Special Considerations/Affordances for Training Applications	Potential Barriers	Next Steps	Risks	Benefits
<i>Pedagogical Concerns</i>	The risk of AI use undermining academic integrity, reducing deep learning, and misaligning with learning objectives.	<p>Develop AI-assisted assignments that encourage critical thinking rather than rote responses.</p> <p>Use AI as a co-creator rather than a content generator to maintain deep learning engagement.</p> <p>Align AI tools with specific learning objectives through instructor oversight.</p>	<p>AI-assisted assignments may still be vulnerable to misuse or over-reliance on AI-generated content.</p> <p>Balancing AI's role as a co-creator while ensuring student effort can be challenging.</p> <p>Instructor oversight requires additional time and effort, which may not be feasible for all educators.</p>	<p>Encourages students to think critically and engage deeply with learning materials.</p> <p>Maintains academic integrity while leveraging AI to enhance learning.</p> <p>Ensures AI tools are used in alignment with pedagogical goals, improving learning outcomes.</p>
<i>Instructor Readiness and Resistance</i>	Lack of AI literacy, resistance to change, and the time investment required for effective AI integration.	<p>Provide professional development on AI literacy for educators and students.</p> <p>Encourage AI experimentation through faculty workshops and communities of practice.</p> <p>Develop easy-to-use AI guidelines and integration strategies for instructors.</p> <p>Offer institutional support for AI-based teaching innovations.</p>	<p>Resistance to AI adoption may persist despite training efforts.</p> <p>Faculty may feel overwhelmed by the additional workload required for AI integration.</p> <p>Training costs and time investments may strain institutional resources.</p>	<p>Increased faculty confidence and competence in AI-enhanced teaching.</p> <p>Encourages innovation and collaboration in AI-based pedagogy.</p> <p>Facilitates smoother AI adoption and enhances teaching effectiveness.</p>

Academia (Continued)				
Special Considerations/Affordances for Training Applications	Potential Barriers	Next Steps	Risks	Benefits
<i>Student Challenges</i>	Unequal access to AI tools, overreliance on AI for learning, and difficulty in evaluating AI-generated outputs.	<p>Provide equal access to AI tools through institutional licensing and funding support.</p> <p>Teach students critical AI literacy skills, including how to evaluate AI-generated content.</p> <p>Establish guidelines on responsible AI use to prevent overreliance.</p> <p>Design AI-integrated assessments that require human input and reflection.</p>	<p>Limited funding may result in inequitable access to AI tools.</p> <p>Students may still misuse AI despite guidelines and literacy efforts.</p> <p>Evaluating AI-generated outputs may require additional instructor oversight.</p>	<p>Promotes responsible AI use and critical thinking among students.</p> <p>Reduces disparities in AI access and usage across diverse student populations.</p> <p>Enhances students' ability to assess AI-generated information critically.</p>
<i>Technical and Implementation Barriers</i>	Issues with AI accuracy, lack of customization, and challenges integrating AI tools with Learning Management Systems (LMSs).	<p>Invest in AI tools that are accurate, transparent, and align with institutional values.</p> <p>Develop AI systems that integrate smoothly with LMS platforms (Canvas, Blackboard, etc.).</p> <p>Offer customization options for AI tools to fit course-specific needs.</p> <p>Regularly review AI-generated outputs for quality control.</p>	<p>AI tools may still produce errors or require constant updates.</p> <p>Compatibility issues may persist despite integration efforts.</p> <p>Customization demands may exceed institutional capabilities.</p>	<p>Increases reliability and trust in AI-driven educational tools.</p> <p>Enhances efficiency by ensuring seamless integration with LMS.</p> <p>Customization allows for more tailored and effective AI implementations.</p>

Academia (Continued)				
Special Considerations/Affordances for Training Applications	Potential Barriers	Next Steps	Risks	Benefits
<i>Ethical and Legal Concerns</i>	Concerns about AI bias, data privacy, and intellectual property rights in educational settings	<p>Implement fairness audits to reduce AI bias and ensure equitable outcomes.</p> <p>Enforce data privacy measures to comply with FERPA, GDPR, and other regulations.</p> <p>Educate faculty and students on intellectual property considerations in AI-generated work.</p> <p>Promote ethical AI use through institutional policies.</p>	<p>Bias reduction in AI is an ongoing challenge that requires constant monitoring.</p> <p>Compliance with privacy laws can be complex and resource-intensive.</p> <p>Intellectual property concerns may deter educators from using AI-generated content.</p>	<p>Improves fairness and ethical AI practices in education.</p> <p>Protects student and faculty data while ensuring regulatory compliance.</p> <p>Provides clearer guidance on responsible AI use in academic settings.</p>
<i>Institutional and Policy Barriers</i>	Absence of clear policies, outdated assessment models, and limited institutional funding for AI adoption.	<p>Establish clear AI policies at institutional and departmental levels.</p> <p>Adapt accreditation and assessment models to accommodate AI-assisted learning.</p> <p>Allocate funding for AI resources and training programs.</p> <p>Encourage cross-disciplinary collaboration to develop best practices for AI in education.</p>	<p>Policy development may lag behind technological advancements.</p> <p>Updating assessment models may face resistance from accreditation bodies.</p> <p>Funding constraints may limit the widespread adoption of AI initiatives.</p>	<p>Provides a structured and responsible approach to AI adoption.</p> <p>Modernizes assessment models to better align with AI-enhanced learning.</p> <p>Encourages sustainable institutional investment in AI technologies.</p>

Acknowledgements

The statements and opinions expressed in this chapter do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

- Ahmed, I., Alvarado, P., Jain, S., Arner, T., Reilley, E., & McNamara, D. (2025). ASU CreateAI Platform: Introduction to CreateAI: Bridging AI Innovation and Accessibility at ASU. In this book.
- CDAO (2025). Responsible AI. <https://www.ai.mil/Initiatives/Responsible-AI/>
- Demarest, C. (2023). Meet Amelia, the US Navy's conversational AI tech-support tool. Defense News, Retrieved from <https://www.defensenews.com/industry/2023/06/13/meet-amelia-the-us-navys-conversational-ai-tech-support-tool/>
- Data, Analytics and Artificial Intelligence Adoption Strategy (2023). <https://www.defense.gov/News/News-Stories/Article/Article/3578219/dod-releases-ai-adoption-strategy/>
- DoD Responsible AI Working Council in accordance with the memorandum issued by Deputy Secretary of Defense Kathleen Hicks on May 26, 2021, Implementing Responsible Artificial Intelligence in the Department of Defense (2022). <https://media.defense.gov/2022/Jun/22/2003022604/-1/-1/0/Department-of-Defense-Responsible-Artificial-Intelligence-Strategy-and-Implementation-Pathway.PDF>
- Jones, N. (2025). AI hallucination can't be stopped- but these techniques can limit their damage. *Nature*. Retrieved from <https://www.nature.com/articles/d41586-025-00068-5>
- Goldberg, B., Owens, K., Gupton, K., Hellman, K., Robson, R., Blake-Plock, S., & Hoffman, M. (2021). Forging competency and proficiency through the synthetic training environment with an experiential learning for readiness strategy. In *Interservice/Industry Training, Simulation, and Education Conference (IITSEC), Orlando, FL*.
- Graesser, A. C., Conley, M. W., & Olney, A. (2012). *Intelligent tutoring systems*. In K. R. Harris, S. Graham, T. Urdan, A. G. Bus, S. Major, & H. L. Swanson (Eds.), *APA educational psychology handbook*, Vol. 3. Application to learning and teaching (pp. 451–473). American Psychological Association. <https://doi.org/10.1037/13275-018>
- Grewal, D., Guha, A., & Becker, M. (2024). AI is Changing the World: Achieving the Promise, Minimizing the Peril. *Journal of Macromarketing*, 44(4), 936-947. <https://doi.org/10.1177/02761467241289573> (Original work published 2024)
- Karimi, E., Smith, J., Billard, R., & Veitch, B. (2024). AI-based adaptive instructional systems for maritime safety training: A systematic literature review. *Discover Artificial Intelligence*, <https://doi.org/10.1007/s44163-024-00153-0>
- Mishra, P., McNamara, D. S., Goodwin, G., & Zapata-Rivera, D. (2025). Large Language Models and Intelligent Tutoring Systems: Conflicting paradigms and possible solutions. In this book.
- Naval Air Warfare Center Training Systems Division (2024). *Research Compendium 2024*. Retrieved from <https://www.navair.navy.mil/nawctsd/sites/g/files/jejdrs596/files/document/%5Bfilename%5D/Research%20Compendium%202024%20NEW%20ABRIDGED%20VERSION%20Standard%20Quality%20508-Compliant.pdf>
- Schroeder, B.L., Hochreiter, J.E., Thayer, S.C., Rivera, J.A. & Van Buskirk, W.L. (2024). Is AI instruction Comparable to human instruction? Designing a pedagogical agent for complex task training. *22nd Triennial Congress of the International Ergonomics Association*.
- Schroeder, B.L., Hochreiter, J.E., Van Buskirk, W.L. (in press). Using artificial intelligence to train human intelligence: Theory and practice in the design of adaptive training systems. In J. Ferraro & P. Mangos (Eds.), *AI and Gamification Technologies for Complex Work*. Taylor & Francis.
- Steinhauser (2024). Assessing Early Successes, Challenges, and the Potential of Generative AI (Artificial Intelligence) in Intelligent Tutoring Systems. Presentation at the 12th Expert Workshop: Assessing Early Successes, Challenges, and the Potential of Generative AI in Intelligent Tutoring Systems

- Stensrud, B. (2024). Artificial intelligence offers unprecedented potential to revolutionise military training, enabling agile and decisive forces. [AI innovation set to revolutionise military training landscape | Shephard](#)
- The EU Artificial Intelligence Act: Up-to-date developments and analyses of the EU AI Act (2025). <https://artificialintelligenceact.eu>
- Wheeler Atkinson, B. F., Pharmed, J., Goldberg, B., Stensrud, B., Mercado, J., Chaparro, B., & Carroll, M. (2023). Virtual Environments for Training: Human Factors Limitations, Challenges, and Opportunities. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 67(1), 1556-1561. <https://doi.org/10.1177/21695067231192611>
- Wu, David. "The Use of AI and AI Algorithms in Financial Markets." *Michigan Journal of Economics*, 9 Mar. 2025, <https://sites.lsa.umich.edu/mjc/2025/03/09/the-use-of-ai-and-ai-algorithms-in-financial-markets/>.
- Zhang, B. H., Lemoine, B., & Mitchell, M. (2018, December). Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 335-340).

BIOGRAPHIES

Editors

Dr. Anne M. Sinatra is a Research Psychologist at the US Army DEVCOM Soldier Center, Simulation and Training Technology Center (STTC) in Orlando, FL. Her research focuses on applying cognitive psychology and human factors principles to computer-based education and adaptive training to enhance learning. She is a member of the research team for the award winning Generalized Intelligent Framework for Tutoring (GIFT) software. She has been the lead editor of the Design Recommendations for Intelligent Tutoring Systems book series since 2019. Dr. Sinatra holds a Ph.D. in Applied Experimental and Human Factors Psychology from the University of Central Florida.

Dr. Vasile Rus is the Jack and Jane Morris Professor of Computer Science and Intelligent Systems at The University of Memphis. He also serves as the Director of the Data Science Center and Program. Dr. Rus' research interests are at the intersection of human, animal, and machine learning; specifically, he is exploring how to use Artificial Intelligence and the data revolution to further our understanding of how people learn, how to improve adaptive instructional systems (AISs), and how to make emerging learning ecologies that include online and blended learning with AISs more effective, efficient, engaging, equitable, relevant, and affordable. Dr. Rus' research has been extensively funded by many federal funding agencies. Currently, he serves as PI on 4 projects funded by NSF and Department of Education and as co-PI on 2 projects, one funded by NSF and one by Department of Defense, for a total amount of funding of more than \$11 million. Other accomplishments include 150+ peer-reviewed publications (conference papers, journal articles, book chapters), 3 best paper awards (5 best paper award nominations, all with his student advisees), winner of several research competitions (e.g., on automated Question Answering), and supervising and graduating 10 PhD students and 37 Masters students.

Dr. Arthur C. Graesser is an Emeritus professor in the Department of Psychology and the Institute of Intelligent Systems at the University of Memphis, as well as an Honorary Research Fellow at University of Oxford. His research interests include question asking and answering, tutoring, text comprehension, inference generation, conversation, reading, problem solving, memory, emotions, artificial intelligence, computational linguistics, and human-computer interaction. He served as editor of the journal *Discourse Processes* and *Journal of Educational Psychology*, as well as presidents of four societies, including Society for Text and Discourse, the International Society for Artificial Intelligence in Education, and the Federation of Associations in the Behavioral and Brain Sciences. He and his colleagues have developed and tested software in learning, language, and discourse technologies, including those that hold a conversation in natural language and interact with multimedia (such as AutoTutor) and those that analyze text on multiple levels of language and discourse (Coh-Metrix and Question Understanding Aid -- QUAID). He has served on four panels with the National Academy of Sciences and four OECD expert panels on problem solving, namely PIAAC 2011 Problem Solving in Technology Rich Environments, PISA 2012 Complex Problem Solving, PISA 2015 Collaborative Problem Solving (chair), and PIAAC Complex Problem Solving 2021.

Dr. Paige M. Lawton is a Research Psychologist at the US Army Combat Capability Development Command – Soldier Center. She holds a Ph.D. in Human Factors Psychology from Embry-Riddle Aeronautical University where her area of focus was on user experience, human error, and system safety. Her current research focuses on developing and evaluating adaptive and intelligent training systems.

Authors

Dr. Tracy Arner serves as the Associate Director of Arizona State University’s Learning Engineering Institute, where she helps lead initiatives that integrate learning sciences, Generative AI, and instructional design to improve student success at scale. She earned her PhD in educational psychology from Kent State University in 2020, building on earlier degrees in educational technology and information technology that grounded her work at the intersection of cognition and technology-enhanced learning. In her prior roles as a K12 technologist, researcher, and postdoctoral scholar, Dr. Arner investigated reading comprehension, misconceptions, game-based learning, and technology-enhanced instruction, contributing to the design and evaluation of intelligent tutoring systems and other evidence-based learning tools. Today, her work focuses on applying learning engineering methods and AI-enabled solutions in real educational contexts, collaborating with faculty, technologists, and external partners to design data-informed interventions that support diverse learners and innovative teaching practices.

Dr. Scott Crossley is a professor of special education whose research focuses on using data science and artificial intelligence to explore cognition, with an emphasis on natural language processing. Dr. Crossley is on the National AI institute of Adult Learning and Online Education (AI-ALOE) leadership team (<https://aialoe.org/>) and runs the Language and Educational Analytics Research (LEAR) Lab (<https://learlab.org/>). His current work focuses on the development of intelligent textbooks, natural language processing tools, and the application of computational tools and machine learning algorithms in learning analytics and text analysis. Dr. Crossley recently served as professor of applied linguistics and computer sciences at Georgia State University. He received his Ph.D. in English and linguistics and M.A. in linguistics from the University of Memphis in 2006 and 2002, respectively. He holds a B.A. in history from California State University of Northridge.

Dr. Paul Deane is a principal research scientist in the Research & Division at ETS (Educational Testing Service). He earned his Ph.D. in linguistics at the University of Chicago in 1987 and is the author of *Grammar in Mind and Brain*, that delves into the interplay of cognitive structures in syntax and semantics, and co-authored "Vocabulary Assessment to Support Instruction." Prior to his tenure at ETS, Deane taught English composition and linguistics at the University of Central Florida from 1986 to 1994. Transitioning to industrial natural language processing from 1994 to 2001, he concentrated on lexicon development, parser design, and semantic information retrieval. Now with ETS, Dr. Deane’s research pursuits involve the formative assessment design in the English language arts, cognitive models of writing skill, automated essay scoring, and vocabulary assessment. Throughout his career, he has contributed to various natural language processing (NLP) and assessment projects, spanning automated item generation, verbal test development tools, collocation error scoring, reading and vocabulary assessment, and automated essay scoring, with a current focus on English language arts skill assessment and writing trait analysis using digital logs and NLP analysis of student texts.

Dr. Peter W. Foltz serves as a Research Professor in the Institute of Cognitive Science at the University of Colorado at Boulder and Executive Director of the NSF Institute for Student-AI Teaming. Dr. Foltz’s research covers machine learning and natural language processing for educational and clinical assessments, large-scale data analytics, cognitive skills in reading and writing, team collaboration, and 21st Century skills learning. Much of his work has focused on NLP techniques for automatically analyzing the meaning of language through writing and speaking. The approaches are used for assessing abilities, for providing

feedback, and for understanding underlying cognitive mechanisms in the brain. Dr. Foltz's expertise and leadership have earned him recognition as a leading authority in educational technology, shaping the future of teaching and learning through his impactful research and scholarship.

Dr. Carol Forsyth is a research scientist at Educational Testing Service Research Institute, where she focuses on innovations in digital assessment and learning environments. Her work applies theoretically grounded educational data mining, computational linguistics, and advanced methodologies—including Generative AI and conversational models—to improve the effectiveness and adaptability of technology-enhanced assessments and instructional systems. Forsyth's research spans areas such as conversation-based assessment, serious games, collaborative problem solving, and simulation-based learning, contributing to both applied solutions and the broader scientific understanding of assessment and educational measurement.

Dr. Benjamin Goldberg is a Senior Scientist and Technical Lead for the Adaptive & Intelligent Training Systems Team at the US Army's DEVCOM – Soldier Center. His research focuses on adaptive experiential learning with an emphasis on simulation-based environments and leveraging data and Artificial Intelligence to create personalized experiences that drive accelerated competency development. Dr. Goldberg holds a Ph.D. in Modeling & Simulation from the University of Central Florida and is well published across several high-impact journals and proceedings, including *IEEE Transactions of Learning Technologies*, the *Journal of Artificial Intelligence in Education*, and *Computers in Human Behavior*.

Dr. Gregory Goodwin is a Branch Chief and senior research scientist at the Combat Capabilities Development Command, Soldier Center in Orlando, FL. Prior to that, he worked in academia. After working in academia, Dr. Goodwin has spent the last twenty years working for the Army researching ways to improve training methods and technologies. He holds a Ph.D. in Psychology from Binghamton University and an M.A. in Psychology from Wake Forest University.

Mr. Langdon Holmes is a PhD student in Psychological Sciences at Vanderbilt University, where he studies analytics, second language acquisition, and the application of transformer-based models to educational data. Working in the Learning Analytics Lab, Holmes conducts research that leverages computational linguistics, large language models, and quantitative methods to better understand how people learn language and how AI can support assessment and instruction. His work contributes to advances in educational measurement, adaptive learning systems, and the analysis of learner behavior using cutting-edge AI techniques.

Mr. Siddharth Jain serves as AI Engineering Lead, AI Acceleration at Arizona State University, where he focuses on advancing artificial intelligence initiatives and integrating cutting-edge AI and robotics technologies into research and real-world applications. With a Master of Science in Robotics and Autonomous Systems, Jain is dedicated to exploring deep learning, autonomous systems, and edge computing to drive innovation across AI projects and institutional goals. His work reflects a commitment to fostering AI research excellence and practical deployment of intelligent systems.

Dr. Richard N. Landers is the John P. Campbell Distinguished Professor of Industrial-Organizational Psychology at the University of Minnesota. He is Principal Investigator for TNTLAB (Testing New Technologies in Learning, Assessment and Behavior), where his research concerns the use of innovative technologies like games, gamification, machine learning, artificial intelligence, and virtual reality, to improve psychometric assessment, employee selection, adult learning, and research methods. He is a Fellow of the Society for Industrial and Organizational Psychology, American Psychological Association, and Association for Psychological Science. His work appears primarily in psychology journals, such as *American Psychologist*, *Journal of Applied Psychology*, *Journal of Business and Psychology*, *Industrial and Organizational Psychology Perspectives*, and *Psychological Methods*, and in interdisciplinary human-computer interaction outlets, including *Computers in Human Behavior*, *International Journal of Human-*

Computer Studies and Simulation & Gaming. He is currently an associate editor for the International Journal of Selection and Assessment and the incoming editor-in-chief for Technology, Mind and Behavior. He is author of a statistics textbook, a research methods textbook, and has developed two edited scholarly volumes: Social Media in Employee Selection and the Cambridge Handbook of Technology and Employee Behavior.

Dr. James C. Lester is the Goodnight Distinguished University Professor in Artificial Intelligence and Machine Learning at North Carolina State University. He is the Director of the Center for Educational Informatics and the Director of the National Science Foundation AI Institute for Engaged Learning. His research centers on transforming education with artificial intelligence. His current work ranges from AI-driven narrative-centered learning environments and virtual agents for learning to multimodal learning analytics and sketch-based learning environments. He received a National Science Foundation CAREER Award, four Best Paper Awards, and the International Federation for Autonomous Agents and Multiagent Systems Influential Paper Award. At North Carolina State University, he has been recognized with the Alumni Association Outstanding Research Award, the Outstanding Teacher Award, and the Alexander Quarles Holladay Medal for Excellence. His research is supported by the National Science Foundation, the US Department of Education's Institute of Education Sciences, the National Institutes of Health, and the Army Futures Command. He has served as Editor-in-Chief of the International Journal of Artificial Intelligence in Education. He is a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI).

Dr. Mihai Lintean serves as Lead Research and Development Engineer Scientist at EdClub Inc. He received his PhD at the University of Memphis, with focus on natural language processing, educational data mining, and computational models of learning. His research develops and applies machine learning and NLP techniques to understand human communication, automated assessment, and intelligent tutoring systems, with particular emphasis on measuring semantic similarity, discourse processes, and adaptive learning technologies. Lintean's work bridges theoretical advances in computational linguistics with practical tools that support writing evaluation, learner feedback, and data-driven educational outcomes.

Dr. Danielle S. McNamara is a Professor in the Department of Psychology, Director of the Science of Learning and Educational Technology (SoLET) Lab, and Executive Director of the Learning Engineering Institute at Arizona State University. She is an international expert in the fields of cognitive and learning sciences, comprehension, writing, natural language processing (NLP), intelligent tutoring, and artificial intelligence in education. She and her team have developed several intelligent tutoring systems, including iSTART and Writing Pal, for comprehension and writing instruction and practice. Research on these technologies have explored methods for improving student engagement via gamebased practice, enhanced adaptability functions, and assessed the feasibility and usability of these systems in real world settings such as high school classrooms. Dr. McNamara has also led the development and testing of multiple NLP tools that have been used in various contexts and learning environments involving essay writing, reading comprehension, second language learning, and creativity. Such tools allow for quick, efficient, and reliable analyses of large text corpora, which is particularly relevant and valuable when analyzing big data. She is an APS, AERA, and ST&D fellow, Founding Editor of APA Technology, Mind, & Behavior, and served as associate editor for the Cognitive Science Journal, Journal of Educational Psychology, Learning & Instruction, International Journal of Artificial Intelligence in Education, and founding AE of Topics in Cognitive Science.

Dr. Wookhee Min is a Senior Research Scientist in the Department of Computer Science at NC State University. His research focuses on adaptive learning technologies, with an emphasis on game-based learning environments, multimodal learning analytics, user modeling, and natural language processing. Min serves as the Managing Director of the NSF AI Institute for Engaged Learning (EngageAI) and is a member of the Center for Educational Informatics. He is a Co-Principal Investigator on several federally funded

projects, including AI Play, which introduces artificial intelligence to middle school students and teachers through workshops, camps, and school-based programs while examining pathways to engage rural students in AI careers; ExplainIt, which seeks to transform undergraduate STEM education through an explanation-based classroom response system that provides real-time support to students and instructors; and CompGen, which develops machine learning-driven, competency-based scenario generators that deliver tailored synthetic training experiences in support of competency-based experiential learning.

Dr. Punya Mishra is Director of *Innovative Learning Futures* at the *Learning Engineering Institute* and Professor at the *Mary Lou Fulton College of Teaching & Learning Innovation* at *Arizona State University*. As director, he provides strategic vision around research, program development, and public engagement. He is internationally recognized for his work in educational technology, creativity, and the application of design to educational innovation. A *TED-Ed educator* and *AERA Fellow* he is ranked in the top 2% of scientists worldwide, #44 (#4 in psychology) among scholars with the biggest influence on practice and policy. He is a podcaster, award-winning instructor, engaging public-speaker and an accomplished visual artist.

Dr. Bradford Mott is a Senior Research Scientist in the Center for Educational Informatics and an Adjunct Assistant Professor in the Department of Computer Science at North Carolina State University. His interdisciplinary research sits at the intersection of artificial intelligence, human-computer interaction, and education, with a focus on AI-driven learning technologies. He designs, develops, and studies narrative-centered game-based learning environments, intelligent tutoring systems, and student modeling approaches that support engagement, problem solving, and computational thinking. Dr. Mott's work has contributed to award-winning educational and interactive technologies, and his research has informed advances in K-12 AI and computer science education, adaptive learning systems, and human-AI collaborative environments.

Dr. Benjamin Nye is the Director of Learning Science at the University of Southern California, Institute of Creative Technologies (ICT). Ben's research tries to remove barriers to development and adoption of adaptive and interactive learning technology so that they can reach larger numbers of learners. Dr. Nye's research has been recognized for excellence in adaptive and intelligent tutoring systems, cognitive agents, and realistic behavior in training simulations. His research is on scalable learning technologies and design principles that promote learning, with the goal of making effective learning tools more broadly available.

Dr. Andrew M. Olney served as Professor in both the Institute for Intelligent Systems and Department of Psychology at the University of Memphis. Dr. Olney received a B.A. in Linguistics with Cognitive Science from University College London in 1998, an M.S. in Evolutionary and Adaptive Systems from the University of Sussex in 2001, and a Ph.D. in Computer Science from the University of Memphis in 2006. He is the former editor of the *Journal of Educational Data Mining* (2017-2022) and former Director/Associate Director of the Institute for Intelligent Systems at the University of Memphis (2006-2017), where he oversaw tenure track faculty, staff, and approximately 40 faculty affiliates. Dr. Olney has over 100 publications in the fields of artificial intelligence, education, and psychology, with a particular focus on using natural language processing to assess and enhance learning from text and learning through conversation. He has been an investigator on \$18M in federal grants (\$6M as PI) in these fields. His primary research interests are in natural language interfaces. Specific interests include language and knowledge acquisition (both human and AI) with applications in learning and performance support systems.

Dr. Elizabeth Reilley is the Executive Director, AI Acceleration, with Enterprise Technology at Arizona State University. Previously at ASU, she worked in the area of Data and Analysis also within Enterprise Technology. Elizabeth has 15 years of experience in higher education working in Academic Affairs, Information Technology, and Strategy and Policy. She has a PhD in Educational Research, Measurement and Evaluation and an MBA from the University of North Carolina at Greensboro and a BA in Philosophy and Community and Justice Studies from Guilford College.

Dr. Rebecca (Becky) Robinson is a Research Psychologist with the Institutional Research and Assessment Division (IRAD), Vice Provost for Academic Affairs (VPAA), at the Army University. She holds a PhD in Experimental Psychology from the University of Texas at Arlington where she conducted studies in social influence and behavior, attitudes, decision-making, personality/individual differences, and survey development/psychometrics. Dr. Robinson's current efforts include improving self-regulated learning skills in early military education, developing program evaluations, assessing online learning tools, and providing technical advisory services for research functions.

Dr. Ramona Shires is a Research Scientist and Senior Instructional Systems Designer at Aptima, Inc., where she conducts research at the center of cognitive science, human performance, and instruction systems design. Her work focuses on applying rigorous empirical methods and theoretical models to understand and enhance decision-making, training effectiveness, and adaptive performance in dynamic environments. Shires's expertise includes instructional analyses and design, human-computer interaction, and AI engineering, supporting solutions that improve safety, performance, and resilience in defense and other high-stakes operational domains. She is committed to advancing research that bridges human cognition and technological innovation to solve real-world training challenges.

Dr. Robert A. Sottolare is the Vice President of Training & Simulation Solutions at Accelint LLC, formerly known as Soar Technology, LLC (SoarTech). He was previously SoarTech's Director of Learning Sciences and joined SoarTech in 2018 after completing a 35-year federal career in both Army and Navy training science and technology organizations. As of 2026, Dr. Sottolare has over 40 years of experience as an engineer, scientist, professor, and program manager for training science and technology projects in industry (7 years and counting), government (35 years), and academia (4 years). He has published over 300 peer-reviewed technical papers as part of his research and earned over 3600 citations (h-index = 31; i10 index = 69; see Google Scholar). His research and development experience covers networked simulations, distributed simulations, distributed teams, adaptive training methods, artificial intelligence methods for training, testing, and analysis, simulation augmentations, and simulation interface design. He is an IEEE Senior Member. Dr. Sottolare's achievements include several edited books including the first six volumes of the Design Recommendations for Intelligent Tutoring Systems book series. He is the father of the award-winning Generalized Intelligent Framework for Tutoring (GIFT), an open-source adaptive training architecture, and in 2009 earned a patent (U.S. Patent 7,525,735) for his work on high resolution, head mounted projection displays. Dr. Sottolare has been honored as the recipient of the US Army Meritorious Civilian Service Award (2018) for Research & Technical Excellence, the Army Research Laboratory's Commander's Award for Civilian Service, the National Training and Simulation Association's (NTSA's) Governor's Award for Lifetime Achievement in Modeling & Simulation (2015), the US Army Research Development & Engineering Command's (2012; inaugural recipient) Modeling & Simulation Lifetime Achievement Award, and the US Army Achievement Medal for Civilian Service (2008) for Research Excellence. In 2019, NTSA recognized Drs. Sottolare, Goldberg, Brawner & Sinatra for their research and development of the Generalized Intelligent Framework for Tutoring (GIFT) with its Education & Human Performance Team Award. In 2020, Drs. Sottolare & DeFalco were the recipients of I/IITSEC's Best Tutorial Award. He was honored as an Interservice/Industry Training, Simulation & Education Conference (I/IITSEC) fellow (NTSA's highest honor) in 2025. He earned his Doctorate in Modeling & Simulation with a focus in Intelligent Systems at the University of Central Florida and he is a graduate of the Defense Acquisition University's Advanced Program Manager's Course.

Dr. Andy Smith is a Research Scientist in the Center for Educational Informatics at North Carolina State University. He earned his Ph.D. in Computer Science from North Carolina State University. His research focuses on utilizing artificial intelligence and machine learning for applications such as adaptive training and learning environments with an emphasis on user modeling, game-based learning, and educational data

mining. Dr. Smith brings 15 years of experience in AI research and software development, including 5 years of software development experience in industry.

Dr. Randall Spain is a Research Scientist at the US Army Combat Capabilities Development Command (DEVCOM) Soldier Center, Simulation and Training Technology Center (STTC). He received his Ph.D. and MS degrees in Human Factors and Experimental Psychology from Old Dominion University. His research focuses on designing, developing, and evaluating adaptive training technologies with a particular emphasis on investigating data-driven models of coaching and feedback to support team training in synthetic training environments, using natural language processing methods to support team communication analytics, and investigating UI/UX principles for intelligent user interfaces. Prior to joining the DEVCOM-Soldier Center, Dr. Spain was a Research Scientist in the Center for Educational Informatics at North Carolina State University where he led research sponsored by the National Institute of Standards and Technology (NIST), the US Air Force, and the US Army Research Laboratory evaluating AI-driven training and learning technologies.

Dr. William Swartout is Research Professor Emeritus in the USC Computer Science department. At USC he was a co-founder of the Institute for Creative Technologies and served as its Chief Science Officer and Chief Technology Officer. He was also a founding co-Director of the Center for Generative AI and Society. Swartout has been involved in cutting edge research and development of artificial intelligence systems throughout his career. In 2009, Swartout received the Robert Engelmere Award from the Association for the Advancement of Artificial Intelligence (AAAI) for seminal contributions to knowledge-based systems and explanation, groundbreaking research on virtual human technologies and their applications, and outstanding service to the artificial intelligence community. Swartout is a Fellow of the AAAI, has served on their Board of Councilors and is past chair of the Special Interest Group on Artificial Intelligence (SIGART) of the Association for Computing Machinery (ACM). He has served as a member of the Air Force Scientific Advisory Board, the Board on Army Science and Technology of the National Academies and the JFCOM Transformation Advisory Group. Prior to helping found the ICT in 1999, Swartout was the Director of the Intelligent Systems Division at the USC Information Sciences Institute. His particular research interests include virtual humans, natural language processing, particularly explanation and text generation, knowledge acquisition, knowledge representation, and intelligent computer based education. He received his Ph.D. and M.S. in computer science from MIT and his bachelor's degree from Stanford University.

Ms. Lisa Townsend is a Senior Research Psychologist at the US Army Combat Capabilities Development Command Soldier Center, Simulation & Training Technology Center. She has a M.S. in Industrial/Organizational Psychology and a BA in Psychology, from the University of Central Florida (UCF). She has worked on many diverse teams including those within Research and Development, Technology Transfer, Instructional Systems Design, and Human Systems Integration. Ms. Townsend's areas of expertise involve team training, Front End Analysis (FEAs), Training Systems Analyses (TSAs), Instructional Systems Design (ISD), Training Effectiveness Evaluations (TEEs), and the development of training and organization related metrics. Her efforts in these areas have spanned across Services and platforms.

Dr. Shawn Weil served as Chief Growth Officer (CGO) of Aptima Inc., where he held the responsibility of leading Aptima's corporate strategic planning activities, overseeing market expansion, and aligning Aptima's technical offerings with critical real-world needs. He provided senior-level mentorship about customer engagement, technical management, and business process development. As a technical leader, Dr. Weil has served as principal investigator on a variety of Department of Defense research and development programs for DARPA, the Office of Naval Research, the Air Force Research Laboratory, and other organizations. His technical areas of expertise include human-AI teaming, command and control, and collective performance assessment. He also works as a senior mentor for Aptima's principal investigators

and project managers. Dr. Weil received a PhD in Cognitive/Experimental Psychology from The Ohio State University with specializations in cognitive systems engineering, quantitative psychology, and psycholinguistics, and a BA in Psychology and Music from Binghamton University (SUNY). He is a member of the Human Factors and Ergonomics Society, the American Psychological Association, the National Defense Industrial Association, and the Cognitive Science Society.

Dr. Diego Zapata-Rivera is distinguished presidential appointee at ETS in Princeton, NJ. He earned a Ph.D. in computer science (with a focus on artificial intelligence in education) from the University of Saskatchewan in 2003. His research at ETS has focused on the areas of innovations in reporting systems and technology-enhanced assessment including work on adaptive learning and assessment environments, conversation-based assessment, caring assessment, and game-based assessment. His research interests also include Bayesian student modeling, open student models, conversation-based tasks, virtual environments, authoring tools, and program evaluation. Dr. Zapata-Rivera has produced more than 150 publications including edited volumes, journal articles, book chapters, and technical papers. He has been a committee member and organizer of international conferences and workshops in his research areas. He is a co-PI and research co-director of the INVITE AI Institute (invite.illinois.edu). He has been a committee member and organizer of international conferences and workshops in his research areas. He was elected as a member of the International AI in Education Society Executive Committee (2022–2027) and selected to be an IEEE Education Society Distinguished Lecturer (2024-2025). He is a member of the editorial board of User Modeling and User-Adapted Interaction and an associate editor of The International Journal of Artificial Intelligence in Education (IJAIED) and AI for Human Learning and Behavior Change. He is a former associate editor of IEEE Transactions on Learning Technologies. Zapata-Rivera has been invited to contribute his expertise to projects sponsored by the National Research Council, the National Science Foundation, NASA, and the US Army Research Laboratory.

Design Recommendations for Intelligent Tutoring Systems

Volume 12
Generative Artificial Intelligence

Design Recommendations for Intelligent Tutoring Systems (ITSS) explores the impact of intelligent tutoring system design on education and training. Specifically, this volume examines “Generative Artificial Intelligence”. The Design Recommendations book series examines tools and methods to reduce the time and skill required to develop ITSS with the goal of improving the Generalized Intelligent Framework for Tutoring (GIFT). GIFT is a modular, service-oriented architecture developed to capture simplified authoring techniques, promote reuse and standardization of ITSS along with automated instructional techniques and effectiveness evaluation capabilities for adaptive tutoring tools and methods.

About the Editors:

- **Dr. Anne M. Sinatra** is a Research Psychologist at US Army Combat Capabilities Development Command (DEVCOM) – Soldier Center and works on the Generalized Intelligent Framework for Tutoring (GIFT). She has been lead editor on the Design Recommendations for Intelligent Tutoring Systems book series since 2019.
- **Dr. Vasile Rus** is a Professor in the Department of Computer Science at the University of Memphis with a joint appointment in the Institute for Intelligent Systems.
- **Dr. Arthur C. Graesser** is an Emeritus Professor in the Department of Psychology and the Institute of Intelligent Systems at the University of Memphis and is a Honorary Research Fellow at the University of Oxford.
- **Dr. Paige M. Lawton** is a Research Psychologist at US Army Combat Capabilities Development Command (DEVCOM) – Soldier Center.

A Volume in the Adaptive Tutoring Series

