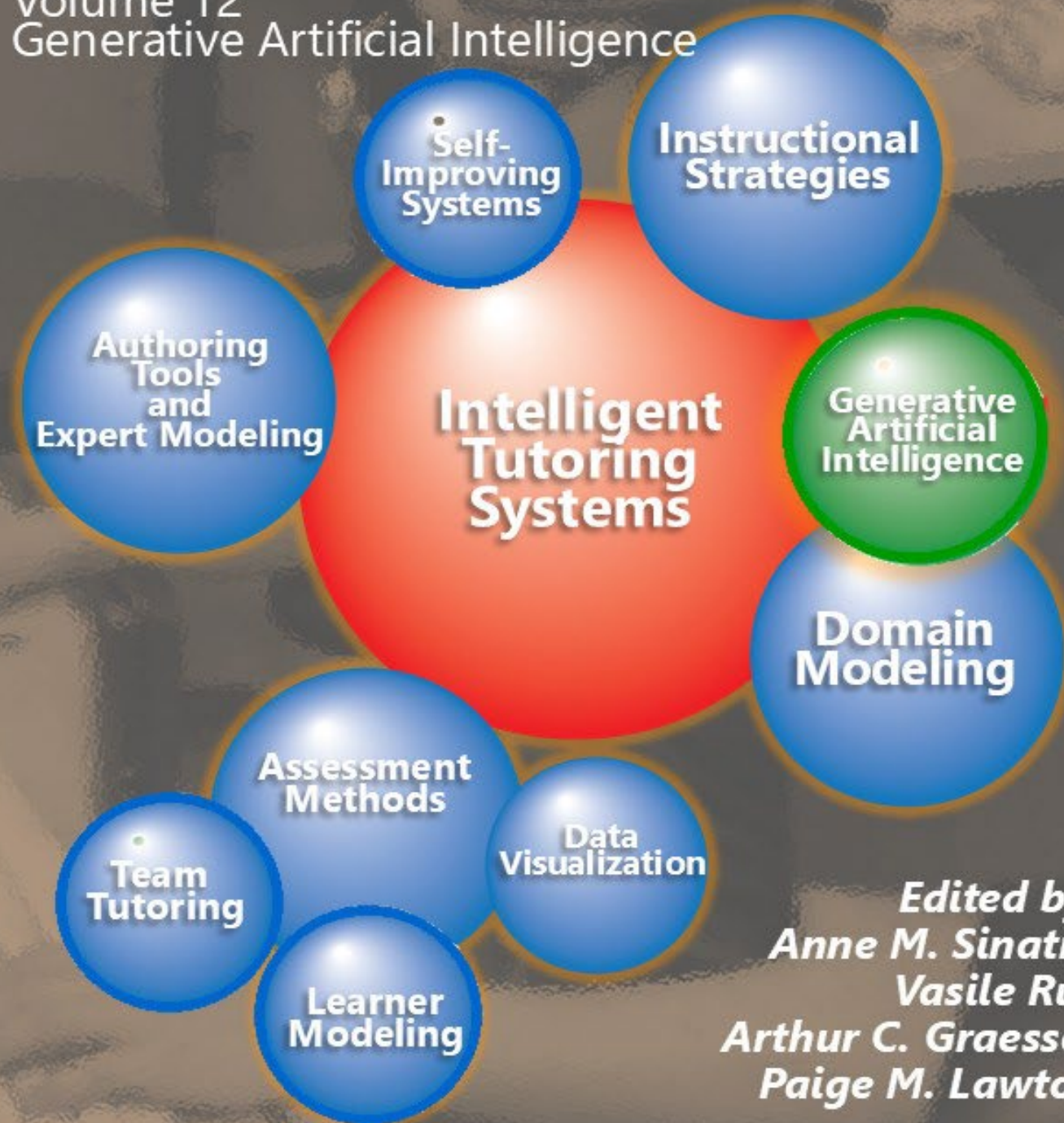


# Design Recommendations for Intelligent Tutoring Systems

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**

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

Punya Mishra<sup>1</sup>, Danielle S. McNamara<sup>1</sup>, Gregory Goodwin<sup>2</sup>, and Diego Zapata-Rivera<sup>3</sup>

Learning Engineering Institute, Arizona State University<sup>1</sup>; US Army Combat Capabilities Development Command (DEVCOM) – Soldier Center<sup>2</sup>; ETS Research Institute<sup>3</sup>

## 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](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., Pislár, 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.