STEM Education

STEM Education, 6(1): 56–83
DOI: 10.3934/steme.2026004
Received: 18 June 2025
Revised: 10 September 2025
Accepted: 14 October 2025
Published: 21 January 2026

https://www.aimspress.com/journal/steme

Research article

Generative Artificial Intelligence (GenAI) 24x7 Tutor: A simulation of the capability of ChatGPT, Wolfram GPT and Tutor Me GPT to accurately and effectively tutor engineering and math content

Sasha Nikolic*, Bao Anh Vu, Yang Di, Ashley Heath, Son Lam Phung, Xiaoping Lu, David Hastie, Md Rabiul Islam, Le Chung Tran and Brad Stappenbelt

Faculty of Engineering and Information Sciences, University of Wollongong, Wollongong, NSW 2522 Australia

*** Correspondence:** Email: sasha@uow.edu.au.

Academic Editor: Wei Li

Abstract: This study investigates the potential of GenAI tools, specifically ChatGPT-4/4o, Wolfram GPT, and Tutor Me GPT, to function as accessible, on-demand 24/7 tutoring systems for engineering and mathematics education. With increasing interest in personalized learning, GenAI offers the promise of scalable and individualized academic support. However, concerns about hallucinations, erroneous or fabricated outputs common in GenAI, have hindered recommendations for unsupervised educational use. The application of GenAI in specialized tutoring contexts has not been rigorously evaluated for accuracy or pedagogical risk in engineering and mathematics. To address this gap, this research assesses the accuracy and instructional capability of GenAI through a human-led simulation involving three trained research assistants. This approach enables the systematic evaluation of how hallucinations may impact learning outcomes. The GenAI tools were tested across seven engineering and two mathematics subjects, encompassing 35 distinct topics. Results indicate that ChatGPT-4 and Wolfram GPT demonstrated strong performance in tutoring electrical engineering and mathematics, but exhibited limitations in mechanical engineering content. Minor inaccuracies were frequently observed, raising concerns about student reliance without oversight. Nevertheless, notable strengths include GenAI's adaptability to varying student proficiency levels and its structured, step-by-step problem-solving methodology. While GenAI shows promise as a supplementary learning tool, further research is required to improve accuracy and evaluate its long-term pedagogical impact in real-world educational settings. Based on current capabilities, GenAI is best regarded as a supportive aid rather than a replacement for human instruction. This study provides foundational insights for the future integration of GenAI into education, with potential to transform tutoring practices.

Keywords: AI tutoring systems, ChatGPT, engineering education, Generative Artificial Intelligence (GenAI), mathematics, personalized learning

1. Introduction

The advantages of one-on-one tutoring over group instruction in enhancing learning outcomes have been widely discussed for over 40 years [1]. Artificial Intelligence (AI) has long been regarded as a promising solution for delivering personalized learning, offering tailored educational experiences aimed at improving student performance [2,3]. Despite significant technological advances, AI-based tutoring systems have not yet reached a level of sophistication sufficient to replace human tutors [4]. However, the release of Generative Artificial Intelligence (GenAI) through ChatGPT-3.5 in November 2022 introduced new possibilities for personalizing learning experiences using a Large Language Model (LLM) architecture, enabling more natural and intuitive interactions with students [4]. Educators have begun integrating GenAI across various disciplines [5–7], but the absence of institutional training and policy guidance has led to widespread uncertainty within the academic community [8]. For instance, this study was initiated due to a human research ethics committee's hesitation to approve human trials without empirical evidence regarding the potential risks associated with GenAI-assisted learning.

A major limitation of ChatGPT-3.5 was its high rate of hallucination, which compromised accuracy, particularly in engineering and mathematics problem-solving [9,10]. Given that accuracy, especially in error detection and correction, is critical for effective tutoring [11], ChatGPT-3.5 was not yet suitable as a 24x7 tutoring solution. In March 2023, ChatGPT-4 was released and demonstrated significantly improved reliability and accuracy in engineering and mathematical content, especially when integrated with Wolfram GPT [11]. The recent introduction of newer models such as ChatGPT-5 shows further improvements, but hallucinations persist across all models and remain an ongoing challenge [13]. However, GenAI performance varies by subject area [11], and no empirical studies have been conducted to provide robust insights into its capabilities in engineering and mathematics, particularly given the notable limitations of existing benchmarking standards [14]. While established benchmarking methods have merit, they often fail to account for pedagogical impact. For example, would a minor error be recognized by an unsupervised student, or would it hinder their conceptual understanding? Furthermore, there is a positive correlation between student errors and the likelihood of GenAI hallucination [15], potentially exacerbating disadvantages for students already at risk. This study does not aim to benchmark GenAI's ability to produce correct answers, but rather its capacity to guide students toward the correct solution and help them understand the underlying reasoning. The interaction between students and GenAI is therefore central to this investigation.

In an era of increasing demand for personalized learning, this study evaluates ChatGPT-4, ChatGPT-4o, Wolfram GPT, and Tutor Me GPT across five engineering subjects and two mathematics subjects that are foundational to engineering curricula. The objective is to simulate a tutoring experience involving multiple interactions per question, assessing the quality and accuracy of the guidance provided at each step. Data collection is conducted by research assistants with prior tutoring experience, who are familiar with common student errors and misconceptions, enabling authentic simulation of student behavior. This human-based simulation employs a piloted generic prompt [15] that a student can input into any GenAI platform, initiating a personalized learning journey in which they acquire theoretical knowledge and practice problem solving through guided questions. Consequently, this paper addresses two key research questions:

1. How accurate and reliable are the selected Generative AI tools in providing tutoring for engineering and mathematics content, and which tool demonstrates the highest performance?
2. How effectively do the selected Generative AI tools deliver a comprehensive and engaging tutorial experience?

The findings of this study offer insights into the capabilities and potential of GenAI in education. If successful, these results could inform future research on pedagogical effectiveness, prompt engineering, and integration with Intelligent Tutoring Systems (ITS). The implications for the broader integration of AI and GenAI in education are substantial, potentially transforming how personalized tutoring is delivered and expanding access to high-quality education globally.

2. Related work

For several decades, AI has been promoted as a transformative force with the potential to deliver tutoring services through personalized learning experiences [16,17]. This includes equipping educators with data-driven insights into student performance, emotional states, and engagement levels, enabling customized instructional strategies and timely interventions [18]. Tailored instruction has been shown to be more effective than traditional group-based teaching [1]. In large classrooms, AI-based solutions are particularly viable due to their scalability [19]. Early implementations of AI in education, such as ITS, demonstrated the feasibility of simulating one-on-one tutoring environments. ITSs have proven especially effective in domains governed by well-defined rules and procedures, such as mathematics and physics [20]. A meta-analysis by Kulik and Fletcher [21] indicates that students tutored by ITSs often outperform those in conventional instruction-only settings. Nevertheless, these systems remain in a continuous cycle of refinement [17].

ITS integrates principles from three core disciplines: Psychology, Computer Science, and Education [22]. While ITS architectures vary depending on purpose and domain, the fundamental components typically include a domain model, which contains the subject knowledge to be taught; a student model, which tracks the learner's progress and understanding; a tutoring model, which determines the instructional strategies to be applied; and a system controller that coordinates the interactions among the three models [22–24]. Over time, ITSs have evolved significantly. Advances in natural language processing (NLP) and machine learning have enhanced their ability to understand and generate human-like text, resulting in more natural and intuitive user interactions [18].

The introduction of GenAI, particularly Large Language Models (LLMs) such as GPT-3.5, marked a significant advancement in AI's ability to interact naturally and intuitively with students [10]. While much of the discourse around GenAI in education has centered on academic misconduct or cheating [25], it is essential to examine ITS potential role in enhancing learning, particularly in personalized tutoring. Unlike earlier ITS, which relied heavily on predefined rules and structured procedures, LLMs are trained on vast and diverse datasets encompassing a broad range of topics and linguistic patterns, making them highly adaptable [26]. A key advantage of LLMs over traditional ITS is their capacity to engage in open-ended dialogue, addressing a wider variety of queries with nuanced understanding and contextual awareness [27]. Recent efforts to integrate LLMs with ITS have begun to emerge in the literature [27,28].

Although LLMs are trained on extensive datasets to generate diverse outputs, including human-like text, GPT-3.5 has demonstrated limitations in accuracy, frequently producing hallucinations and performing poorly on highly technical or specialized content [29]. In-depth analysis by Nikolic et al. [9] revealed that ChatGPT-3.5 performed variably on engineering and mathematics tasks, often correctly outlining steps but ultimately arriving at incorrect final answers. However, in a follow-up study [11], ChatGPT-4 exhibited significantly improved performance, demonstrating greater accuracy in solving math and engineering problems and producing fewer hallucinations, suggesting its potential suitability as a tutoring tool. As these models continue to evolve, performance is expected to improve; however, hallucinations are likely to persist for the foreseeable future [13]. Therefore, a pedagogical understanding of the implications of such inaccuracies is necessary.

Students have already recognized the potential of GenAI and are increasingly using it to support their learning; however, providing them with structured guidance is in their best interest [30]. Based on the performance capabilities outlined in [11], ChatGPT-4 was selected as one of the primary LLMs for this study. It should be noted that selecting ChatGPT-4 based on accuracy assumes that higher accuracy leads to better pedagogical outcomes. While controlled errors can be beneficial for learning [31,32], the impact of uncontrolled errors in a GenAI tutoring context remains unknown. Furthermore, accuracy issues may affect performance expectancy, which in turn could influence student adoption [33].

On 13 May 2024, ChatGPT-4o was released, with advanced capabilities made freely available to the public [34]. This open access rendered it suitable for inclusion in this study. Other LLMs, such as Copilot and Gemini, were considered but found to perform less effectively on engineering and mathematics content [11], and thus were excluded. However, the work of Nikolic et al. [11] identified that integrating Wolfram GPT, a plugin developed by OpenAI, can enhance accuracy in engineering and mathematical problem-solving. Wolfram Alpha is widely recognized as a powerful computational engine for mathematics [35]. By leveraging Wolfram GPT, users gain access to robust computation, accurate mathematical processing, curated knowledge, real-time data, and visualization through Wolfram Alpha and the Wolfram Language [36]. Consequently, Wolfram GPT was included in this study. Data collection concluded in 2024, and since then, OpenAI has introduced more advanced models that fall outside the scope of this research. At the time of experimentation, however, ChatGPT-4o served as the baseline for the free version, which aligns with the study's target context.

Khan Academy is a well-established open educational resource provider with millions of unique users, yet despite its popularity, limited research has been conducted on the effectiveness of its pedagogical approach [37,38]. In March 2023, Khanmigo was introduced, a comprehensive, standalone AI-based tutoring system that integrates GenAI [39]. A less advanced version of Khanmigo [40], later released as the Tutor Me GPT (also known as Khanmigo Lite), was also considered. As Tutor Me GPT has been viewed as a tool with potential to transform learning, similar in function to Wolfram GPT, but requiring further investigation [41], it was included in this study.

Prompt engineering involves modifying input prompts to tailor and optimize the quality and nature of the generated output [9]. Therefore, GenAI must be explicitly prompted to function as a tutor. Unlike the decades of research underpinning ITS, there is limited guidance on effective prompting strategies for educational applications. Several approaches exist, including directly requesting tutor support or hints on specific problems or topics [42], or seeking supportive feedback [43]. The prompt designed for this study employs the Socratic Method, which fosters critical thinking through guided questioning and encourages deeper cognitive engagement [44]. Recently, OpenAI introduced a 'study and learn' feature in ChatGPT that offers a simplified, limited Socratic tutoring experience without requiring users to input a detailed prompt [45].

The scope of this study involves designing a general-purpose prompt to enable GenAI models to simulate an intelligent tutor. The goal is to create a system capable of conversing with students, identifying areas of difficulty, and delivering pedagogically sound instruction through the transfer of theoretical knowledge and the development of problem-solving skills. However, due to the novelty of this approach, research-backed insights are limited, with the most widely recognized prompt framework provided by Mollick and Mollick [46]. Hence, this study is necessary. This simulation aims to assess the capabilities, risks, and opportunities associated with GenAI in tutoring. Once these are established, more in-depth investigations into pedagogical effectiveness, prompt design, and comparisons with traditional ITS can be pursued. As with previous digital innovations, this presents new and promising opportunities for educator collaboration [47].

3. Experiment

Section 3.1 provides a methodological framework overview. Section 3.2 outlines how pilot work informed this study. Section 3.3 describes the simulation process. Section 3.4 details the data collection procedures, and Section 3.5 discusses the study's limitations.

3.1. Methodological framework overview

This study employed a mixed-methods simulation framework consisting of: (i) a standardized tutoring protocol; (ii) a human-in-the-loop role-play design; and (iii) a structured measurement model. These components are elaborated in subsequent sections. As an overview:

Protocol: Each session used a fixed, publicly accessible prompt (Appendix A), a fresh chat history, retained transcripts and notes for audit purposes, and lasted a minimum of 20 minutes. Sessions were conducted across seven subjects (35 topics) and four GenAI tutors (GPT-4, GPT-4o, Wolfram GPT, Tutor Me GPT).

Role-play design: For each topic, two learner profiles were simulated: Case A (average student), alternating between correct and incorrect responses; and Case B (struggling student), consistently providing incorrect responses. Three experienced research assistants (RAs), each with subject expertise and prior tutoring experience, interacted with the GenAI tools and maintained contemporaneous field notes.

Measurement model: Outcomes were recorded using a two-part instrument:

1. Accuracy (primary, objective measure) on a 0–4 scale, anchored by explicit error definitions (none; one minor; few minor; consequential; unacceptable);

2. Tutor Experience, comprising Relevance, Pedagogical Effectiveness, Interactive Engagement, Progression, Contextual Understanding, and Examples/Illustrations (each scored 0–4). The three RAs reviewed each other's outputs to ensure scoring alignment.

Analysis plan: For each topic × model × case combination, item scores and averages were computed. For between-model comparisons across repeated measures (topics), a non-parametric Friedman test was applied.

3.2. The pilot study

The pilot study [15] aimed to evaluate the effectiveness of a generic tutor prompt developed by Mollick and Mollick [46] on ChatGPT-4, along with variations designed to better align with engineering and mathematics content. It was assumed that students would gravitate toward the simplicity of directly asking ChatGPT questions to support their learning. The investigation found that prompt variations had minimal impact on accuracy but slightly influenced user experience. While user experience is subjective, a preferred prompt was selected by the research team based on evaluation results and adopted for the main study (Appendix A). A critical element of the prompt was the instruction '*not to produce diagrams*', as diagrams were consistently inaccurate or irrelevant. Across all prompts, accuracy improved when students provided more correct input [15]. Specifically, ChatGPT-4 made fewer errors when students supplied correct answers to questions.

3.3. Simulations

Building on insights from the pilot study [15], this study simulated two role-playing scenarios. The first scenario (Case A) simulated a student near the middle of the academic distribution. This was achieved by providing correct answers for 50% of responses and incorrect answers for the remaining 50%. The second scenario (Case B) simulated a student who consistently struggled, always providing incorrect answers when prompted. Case B yielded performance data under conditions of maximal student difficulty, while Case A offered insights into the typical experience of an average student. There was no need to test scenarios with consistently correct input, as the pilot demonstrated that model performance improved under such conditions. This design was considered essential to simulate interactions relevant to the most vulnerable learners. The focus was on the quality and dynamics of student–GenAI interactions, rather than the model's standalone accuracy.

Three RAs conducted the simulations, each simulating both cases for their assigned subjects. All were subject matter experts: two were PhD candidates, and one held a doctorate. Each had experience supporting student learning and was familiar with common misconceptions and errors. This authentic expertise guided the simulated interactions. This simulation was necessary to inform future trials with real students. Given well-documented hallucinations [9], the Human Research Ethics Committee required assurance regarding potential impacts on learners. This design ensured that no negative learning consequences would arise for the RAs if the GenAI tutor underperformed.

Three subjects from electrical engineering were selected: Digital Signal Processing (2nd year), Electronics (2nd year), and Power Engineering (3rd year). Two mechanical engineering subjects were included: Engineering Fluid Mechanics (2nd year) and Thermodynamics of Engineering Systems (3rd year). Two core mathematics subjects, foundational to both disciplines, were also selected: Foundations of Engineering Mathematics (1st year) and Advanced Engineering

Mathematics and Statistics (2nd year). For each subject, five random topics were chosen, ensuring diverse performance insights across engineering domains.

3.4. Data collection

Across the seven subjects and 35 topics, the RAs simulated Case A and Case B interactions with ChatGPT-4, ChatGPT-4o, Wolfram GPT, and Tutor Me GPT. ChatGPT-4 was initially selected due to its demonstrated standalone performance [11]. Wolfram GPT was included because prior research indicated it enhances accuracy [12]. ChatGPT-4o was added during data collection, as free public availability offered potential for widespread adoption if results were favorable. Tutor Me GPT was selected due to Khan Academy's extensive investment and resources in advancing technology-based tutoring [40]. Data collection occurred in the second half of 2024.

Within ChatGPT, the 'memory' function was disabled to prevent prior interactions from influencing subsequent sessions. For the same reason, each interaction began in a new prompt session. To initiate a tutorial, the RA copied and executed the pre-designed prompt (Appendix A), beginning a unique learning journey based on the user's specified topic and responses to diagnostic questions. Each GenAI model then delivered theory and practice problems in a personalized manner. It was expected that each practice problem would be broken into individual steps. The RA, acting as a student, provided responses, correct or incorrect, according to the scenario being emulated, incorporating common misconceptions and errors based on their expertise. The following aspects were observed:

- Did the GenAI correctly identify whether the RA provided a right or wrong answer?
- Did the GenAI deliver accurate information in ITS explanations and corrections?
- Did the GenAI provide explanations and insights that were educationally valuable?

Each interaction lasted at least 20 minutes, allowing the RA to engage with multiple examples, explanations, and problems. As such, each simulation was not repeated to calculate an average score. Instead, scores reflected performance across the entire 20-minute session. RAs maintained detailed notes on their observations and reflections.

The evaluation rubric was based on factors identified by Merrill et al. [16], which examined differences between human and computer-based tutoring. Experience was scored from 0 to 4, with 4 being the highest. The rubric had two components. The first was 'accuracy', the most critical and objective measure. A score of 4 was awarded if the GenAI made no errors throughout the interaction. A single minor, inconsequential error resulted in a score of 3. A few minor errors yielded a score of 2, while consequential errors received 1 or 0. Thus, a score of 3–4 was considered necessary to confidently recommend GenAI as a tutoring tool.

Beyond accuracy, the following dimensions were assessed: Relevance to Topic Area, Pedagogical Effectiveness, Interactive Engagement, Progression to More Difficult Concepts, Contextual Understanding, and Use of Examples and Illustrations. Each was scored on a 0–4 scale using descriptive anchors. These metrics are acknowledged as subjective and may reflect personal preferences. Consistency in scoring was evaluated by comparing ratings across the three RAs.

3.5. Limitations

The scope of GenAI models, subjects, topics, and interaction length was determined by balancing achievable insights with available funding. While no interaction was repeated, the minimum 20-minute duration provided substantial data. By simulating two cases per topic (average and struggling students), an effective average score was derived. Although limited in scale, the repetition across multiple topics offered the RAs valuable insights into the risks and opportunities of GenAI tutoring. Given that most sessions encountered at least one hallucination, this sample provided meaningful data on the pedagogical impacts under investigation.

The study employed simulations conducted by RAs rather than real student interactions. This was a prerequisite for obtaining human research ethics approval. While this approach prevents negative learning outcomes, it limits the validity of findings; however, it establishes a foundation for future research.

The RAs are subject matter experts, which may influence both their interactions with GenAI and their evaluation of responses. Additionally, the evaluation rubric includes subjective components that may vary based on individual biases. For this reason, accuracy metrics were separated from other measures. Despite these limitations, the findings offer sufficient insight to justify further research in this area.

GenAI capabilities are evolving rapidly. Between data collection and manuscript preparation, major providers such as OpenAI and Google released more powerful and accurate models. Therefore, the reported findings represent a conservative, worst-case performance scenario.

While this study focuses on GenAI's tutoring capabilities, it is important to acknowledge that human tutors are also fallible and may make errors. However, a comparative analysis of GenAI and human tutor performance is beyond the scope of this study.

4. Results

4.1. Accuracy

Table 1 summarizes the average accuracy performance across Cases A and B for the 7 subjects and 35 topic areas. For each subject, mean scores are reported for each model, with the highest score highlighted in green and the lowest in light red. The results indicate that GenAI models demonstrate substantially higher accuracy in electrical engineering and mathematics compared to mechanical engineering. In electrical engineering, all models achieved scores of at least 3, indicating no more than one minor, non-impactful error, except for Tutor Me, which received an average score of 2.5 in detailed AC (Alternating Current) Analysis. Tutor Me also exhibited the greatest difficulty with mathematics, with four topics scoring below 3. The remaining models performed well, although GenAI tutors (GPT-4) showed limitations in polar coordinates and Fourier Series. These findings suggest strong potential for using ChatGPT-4, GPT-4o, or Wolfram Alpha's GPT-based tools in tutoring applications for these disciplines, based on accuracy. However, the persistence of errors, despite often being minor, remains a critical consideration. The current experimental framework cannot fully assess how such errors might affect independent student learning. Nevertheless, the results support advancing to supervised student trials to evaluate real-world educational impact.

Table 1. Average performance in relation to accuracy (4 is highest).

Subject	#	Topic	4	4o	Wolf	TMe
Digital Signal Processing	1	Convolution and LTI System	3.5	3	3	3
	2	Impulse Response and FIR	3.5	3.5	3.5	3
	3	Difference Equation	3	3.5	3.5	4
	4	Z Transform	3	3.5	3.5	3
	5	Power Series Method	3.5	3	3.5	3
		Average by Model	**3.3**	**3.3**	**3.4**	**3.2**
Electronics	1	Semiconductor Diodes	3.5	3	4	3
	2	Zener Diodes and Application	3	4	4	3
	3	Detailed AC Analysis	4	3.5	4	2.5
	4	Junction FET	4	4	4	3
	5	Metal Oxide Semiconductor	4	3.5	3.5	3
		Average by Model	**3.7**	**3.6**	**3.9**	**2.9**
Power Engineering	1	Electrical Transmission	3.5	3.5	3.5	3.5
	2	AC Power Calculations	4	3.5	3	3
	3	Power System Components	4	3.5	3.5	3
	4	Distribution System Operation	4	3.5	4	3
	5	Safety Customer Install	3.5	4	3.5	3
		Average by Model	**3.8**	**3.6**	**3.5**	**3.1**

Subject	#	Topic	4	4o	Wolf	TMe
Engineering Fluid Mechanics	1	Fluid Pressure on Circle, Square, Triangle	4	0	2.5	4
	2	Bernoulli's principle	1	4	3.5	3.5
	3	Turbine Power	4	4	4	4
	4	Fluid Momentum	1	1	1	1
	5	Fluid Friction in Pipes	3	2.5	1	2.5
		Average by Model	**2.6**	**2.3**	**2.4**	**3**
Thermodynamics of Engineering Systems	1	Rankin Cycle	4	3.5	3	4
	2	Psychrometric Chart	4	4	1.5	2
	3	Two Stage Gas Turbine	0	1.5	0.5	0
	4	Polytropic Processes	0	3.5	2	3.5
	5	Refrigeration cycles	1	1.5	1.5	1.5
		Average by Model	**1.8**	**2.8**	**1.7**	**2.2**
Foundations of Engineering Mathematics	1	Integration	4	4	3.5	4
	2	Taylor Series	4	4	4	3
	3	First Order ODEs	3.5	4	3.5	4
	4	Volume of Solids	3.5	2.5	3.5	2.5
	5	Polar Coordinates	2.5	3.5	4	4
		Average by Model	**3.5**	**3.6**	**3.7**	**3.5**
Advanced Engineering Mathematics and Statistics	1	Partial Derivatives	3	4	3.5	4
	2	Multivariable Chain Rule	4	2.5	4	2.5
	3	Double Integrals	3	3	3.5	1
	4	Laplace Transform	3	4	4	4
	5	Fourier Series	2	3.5	4	2.5
		Average by Model	**3**	**3.4**	**3.8**	**2.8**

However, with respect to mechanical engineering content, the average model accuracy fell below expectations, with most scores below 3. A score of 2 indicates that several inaccuracies were present, increasing the likelihood that students may encounter errors. If students are unaware of these inaccuracies, their learning could be adversely affected. Therefore, student engagement with these topics should proceed with considerable caution. These findings suggest that further evaluation should be conducted in a controlled setting, as students may be capable of identifying and correcting such errors when properly guided.

A Friedman Test [48] was employed to assess differences in "treatments" (the GenAI models) across repeated "measures" (the different topics). The test revealed no statistically significant differences in the accuracy of the GenAI models. This outcome was anticipated due to the limited sample size and the coarse scoring scale used.

4.2. Tutor experience

Table 2 summarizes the average performance in relation to tutor experience across Cases A and B, encompassing the 7 subjects and 35 topic areas. As previously, average scores for each model are provided per subject, with the highest score highlighted in green and the lowest in light red.

Unlike accuracy, which can be objectively evaluated against the rubric, tutor experience is inherently subjective and influenced by the RAs' individual learning preferences and academic backgrounds. For instance, the electrical engineering and mathematics RAs reported similar experience scores, which contrasted with those of the mechanical engineering RA. To account for potential bias, all three RAs reviewed one another's prompt outputs to determine whether differences stemmed from personal preference or topic-specific characteristics. Following discussion, it was concluded that variations in experience were attributable to the nature of the subject matter. Specifically, electrical engineering and mathematics topics benefited from detailed explanations and real-world connections, features more frequently exhibited by ChatGPT-4, whereas such elaboration occasionally overcomplicated mechanical engineering content, which is already strongly grounded in real-world and relatable contexts.

The results indicate that scoring differences were marginal, suggesting that any of the models could serve effectively as tutoring systems, with the lowest average score being 2.97. ChatGPT-4/4o typically provided detailed responses characterized by analytical depth, while Wolfram and Tutor Me delivered more concise answers; nonetheless, all three systems emphasized well-scaffolded learning approaches.

Table 2. Average performance in relation to tutor experience (4 is highest).

			Average Performance			
			4	4o	Wolf	TMe
Digital Signal Processing	1	Relevance to Topic Area	3.4	3.3	3.4	2.9
	2	Pedagogical Effectiveness	3.4	3.4	3.4	2.9
	3	Interactive Engagement	3.6	3.4	3.3	2.9
	4	Progression to More Difficult Concepts	3.3	2.9	3.2	3
	5	Contextual Understanding	3.3	3.3	3.4	3
	6	Use of Examples and Illustrations	3.4	3.3	3.6	3.4
			3.4	**3.27**	**3.38**	**3.02**
Electronics	1	Relevance to Topic Area	3.3	3.3	3.2	2.9
	2	Pedagogical Effectiveness	3.5	3.4	3.3	3.1
	3	Interactive Engagement	3.8	3.3	3.3	3.1
	4	Progression to More Difficult Concepts	3	3.1	3	2.8
	5	Contextual Understanding	3.6	3.2	3.4	3.1
	6	Use of Examples and Illustrations	3.1	3.2	3.4	2.9
			3.38	**3.25**	**3.27**	**2.98**
Power Engineering	1	Relevance to Topic Area	3.4	3.5	3.2	3.2
	2	Pedagogical Effectiveness	3.3	3.2	3.4	3.1
	3	Interactive Engagement	3.5	3.5	3.2	3.3
	4	Progression to More Difficult Concepts	3.2	3.1	2.9	3
	5	Contextual Understanding	3.4	3	3.3	3.1
	6	Use of Examples and Illustrations	3	3	3.2	2.7
			3.3	**3.22**	**3.2**	**3.07**

			Average Performance			
			4	4o	Wolf	TMe
Engineering Fluid Mechanics	1	Relevance to Topic Area	3.3	4	4	3.8
	2	Pedagogical Effectiveness	2.9	3.6	3.4	3.8
	3	Interactive Engagement	3.6	4	3.6	3.8
	4	Progression to More Difficult Concepts	2.5	2.9	2.8	2.4
	5	Contextual Understanding	3.5	3.5	3.8	3.4
	6	Use of Examples and Illustrations	2.9	3.8	3.6	3.6
			3.12	**3.63**	**3.53**	**3.47**
Thermodynamics of Engineering Systems	1	Relevance to Topic Area	3.7	3.8	3.8	3.8
	2	Pedagogical Effectiveness	3.2	3.2	3.6	4
	3	Interactive Engagement	3.5	4	3.3	4
	4	Progression to More Difficult Concepts	3	2.4	3.8	4
	5	Contextual Understanding	3.6	3.4	3.4	4
	6	Use of Examples and Illustrations	2.9	3.3	3.6	4
			3.32	**3.35**	**3.58**	**3.97**
Foundations of Engineering Mathematics	1	Relevance to Topic Area	3.9	4	4	4
	2	Pedagogical Effectiveness	4	3.2	3.2	3.8
	3	Interactive Engagement	4	3.2	3.1	3.2
	4	Progression to More Difficult Concepts	3.8	2.5	3.5	1.6
	5	Contextual Understanding	4	2.8	3.2	2.8
	6	Use of Examples and Illustrations	4	2.6	2.6	2.4
			3.95	**3.05**	**3.27**	**2.97**
Advanced Engineering Mathematics and Statistics	1	Relevance to Topic Area	3.9	3.7	4	4
	2	Pedagogical Effectiveness	3.6	2.6	2.5	3.2
	3	Interactive Engagement	4	3.2	2.2	3.2
	4	Progression to More Difficult Concepts	3.7	3	3.5	3.2
	5	Contextual Understanding	4	3.4	3	3
	6	Use of Examples and Illustrations	3.7	2.7	3	2.8
			3.82	**3.1**	**3.03**	**3.23**

5. Discussion

Section 5.1 presents an overview of the identified strengths, Section 5.2 discusses the identified weaknesses, Section 5.3 outlines the identified opportunities, and Section 5.4 provides a summary of key findings.

5.1. Identified strengths

Implementation of GenAI tutors, specifically GPT-4, GPT-4o, Wolfram, and Tutor Me, revealed both shared strengths and distinct capabilities in enhancing student learning across various engineering and mathematics disciplines. A common strength was their ability to deliver personalized and adaptive learning experiences by tailoring instructional content to students' varying levels of understanding. This adaptability enables accurate assessment of foundational knowledge, offering step-by-step guidance for beginners while posing more challenging problems for advanced learners, thereby promoting academic growth and accommodating diverse educational needs. These findings confirm the personalization benefits reported in prior review studies [10,49].

All four GenAI tutors demonstrated effective pedagogical methodologies. They guided students through problem-solving via small, incremental steps and encouraged reflection on theoretical concepts and their interrelationships. This approach fosters deeper engagement with core material and supports coherent cognitive processing during learning. The GenAI tutors performed particularly well when addressing foundational and entry-level content, effectively responding to questions centered on basic principles and theories. GPT-4 emphasized explaining the intuition behind mathematical formulas and connecting them to real-world applications, whereas GPT-4o, Wolfram, and Tutor Me adopted a more hands-on strategy: presenting example problems, explaining the relevant steps or formulas, and then prompting students to apply the method to similar cases.

The GenAI tutors excelled in adapting to individual student proficiency levels by providing personalized instructional support calibrated to academic background. They adjusted the complexity and direction of content based on inquiries about students' prior knowledge. For instance, in signal processing, they provided clear, step-by-step explanations of convolution operations for beginners while posing advanced synthesis questions for higher-level learners, thus supporting academic progression across skill levels. However, for more complex topics requiring the integration of multiple concepts within a single solution, accuracy became inconsistent. These systems also performed effectively in knowledge retention assessments. For example, they evaluated students' understanding of fundamental distinctions, such as between FIR (finite impulse response) and IIR (infinite impulse response) systems, through targeted questioning.

Despite using identical prompts, the GenAI tutors exhibited distinct characteristics. In electronics, Wolfram focused primarily on computational aspects, helping students improve their mathematical proficiency; GPT-4o and Tutor Me emphasized integrating computation with conceptual understanding to ensure students grasped the physical significance behind calculations; GPT-4 assessed comprehensive understanding by situating questions within real-world application contexts. GPT-4's creative use of examples and illustrations enhanced conceptual relevance. For instance, Figure 1 illustrates how GPT-4, when explaining first-order separable ordinary differential equations (ODEs), prompted students to generate real-world examples, thereby integrating theoretical knowledge with practical application and increasing learning relevance.

Can you think of any situations where this kind of differential equation might be useful, based on the behavior of y as x changes?

I'm not sure

No worries, I can give you a few examples where differential equations like the one we've discussed are quite useful:

1. **Population Dynamics**: Differential equations are often used to model population growth or decline. For instance, $y = Ae^{kt}$ can describe an exponentially growing population where y is the population at time t, A is the initial population, and k is the growth rate. If the population is growing without any constraints (like food or space), k would be positive, leading to exponential growth.
2. **Radioactive Decay**: This type of equation also models the decay of radioactive substances. Here, A would represent the initial amount of a radioactive material, and k would be a negative constant that describes the rate of decay. Over time, the quantity of the substance decreases exponentially.
3. **Interest Calculation in Finance**: Exponential functions describe the growth of investments under continuously compounded interest. If A represents the principal amount, and k the annual interest rate, the amount of money grows exponentially over time.

Each of these scenarios can be modeled using an equation similar to $y = Ae^{kt}$, where the behavior of y as x (or t, often used for time) changes is crucial for predicting outcomes and making decisions.

Figure 1. Example of ChatGPT-4 connecting theory to practice. This example illustrates how the ChatGPT tutor links differential equations to real-world applications. It also demonstrates how opportunities for deeper cognitive engagement may be bypassed by the user.

Similar efforts to connect theory with practice are evident in GPT-4o. For instance, discussing the impact of different conductor materials on power transmission enables students to observe the direct application of theoretical knowledge in practical engineering contexts. This approach broadens students' understanding and reinforces their comprehension of various factors influencing power systems. Flexibility in tiered instruction represents a distinctive pedagogical strength of Wolfram, ensuring that each student is appropriately challenged and supported through differentiated instruction tailored to varying proficiency levels.

Overall, GPT-4, GPT-4o, Wolfram, and Tutor Me exhibit substantial capabilities in delivering personalized and adaptive tutoring experiences, particularly in foundational topics and introductory content. Their strengths include effective learning scaffolding, guided problem-solving, and adaptive instruction aligned with individual student needs. However, as content complexity increases, their effectiveness varies, with some models excelling in fostering conceptual understanding while others, such as Wolfram, are better suited for detailed procedural practice. This mirrors real-world tutoring scenarios, where different human tutors employ distinct pedagogical approaches. Nevertheless, switching between generative AI (GenAI) tutors is significantly more convenient than changing human tutors. Therefore, engaging with multiple GenAI tutors to access diverse instructional strategies may offer a more holistic and realistic learning experience. Furthermore, strategic adjustments to prompts could enhance personalization to better align with individual student requirements.

5.2. Identified weaknesses

In the previous section, several strengths of GenAI tutors were identified, particularly their capacity to personalize the learning experience. However, the technology still exhibits limitations. Engineering and mathematics content is highly equation-driven, and the current communication medium, text-based chat, poses a significant constraint on user experience. Although potential workarounds exist, such as writing equations in code format and rendering them via a LaTeX plugin, these methods are not typical for most students. This chat-based interaction limits students' ability to present intermediate steps when solving mathematical problems. Consequently, the responsibility for decomposing problems into manageable steps and guiding students through each stage falls primarily on the GenAI tutor. While GenAI tutors are generally capable of breaking down solution processes, two issues arise in their implementation: first, individual steps may still require extensive internal calculations that are impractical to input via chat; second, when a student completes an initial step, the GenAI tutor often completes the remaining steps rather than continuing to guide the student. For example, consider the case in Figure 2 involving the calculation of Fourier series coefficients. Although the steps are correctly outlined, it is impractical for the student to input responses for all three steps within the chat interface, as each step involves solving integrals with multiple intermediate operations. The student may instead submit only the result of the first step, as illustrated in Figure 3.

Practice Question

Let's try one more problem to make sure you've got it.

Consider the function $f(t) = t$ over the interval $[-L, L]$. Write the general form of its Fourier series.

Step 1: Find a_0.

Step 2: Find a_n.

Step 3: Find b_n.

Go ahead and set up the integrals for each of these steps. What do you get for each coefficient?

Figure 2. An example of a question that may challenge students when interacting with generative AI (GenAI) through text-only input. In this case, a problem requiring the calculation of Fourier Series coefficients is presented. Although step-by-step instructions are provided for the student to complete, entering responses for all three steps within a single chat interface is impractical due to input constraints and lack of structured response support.

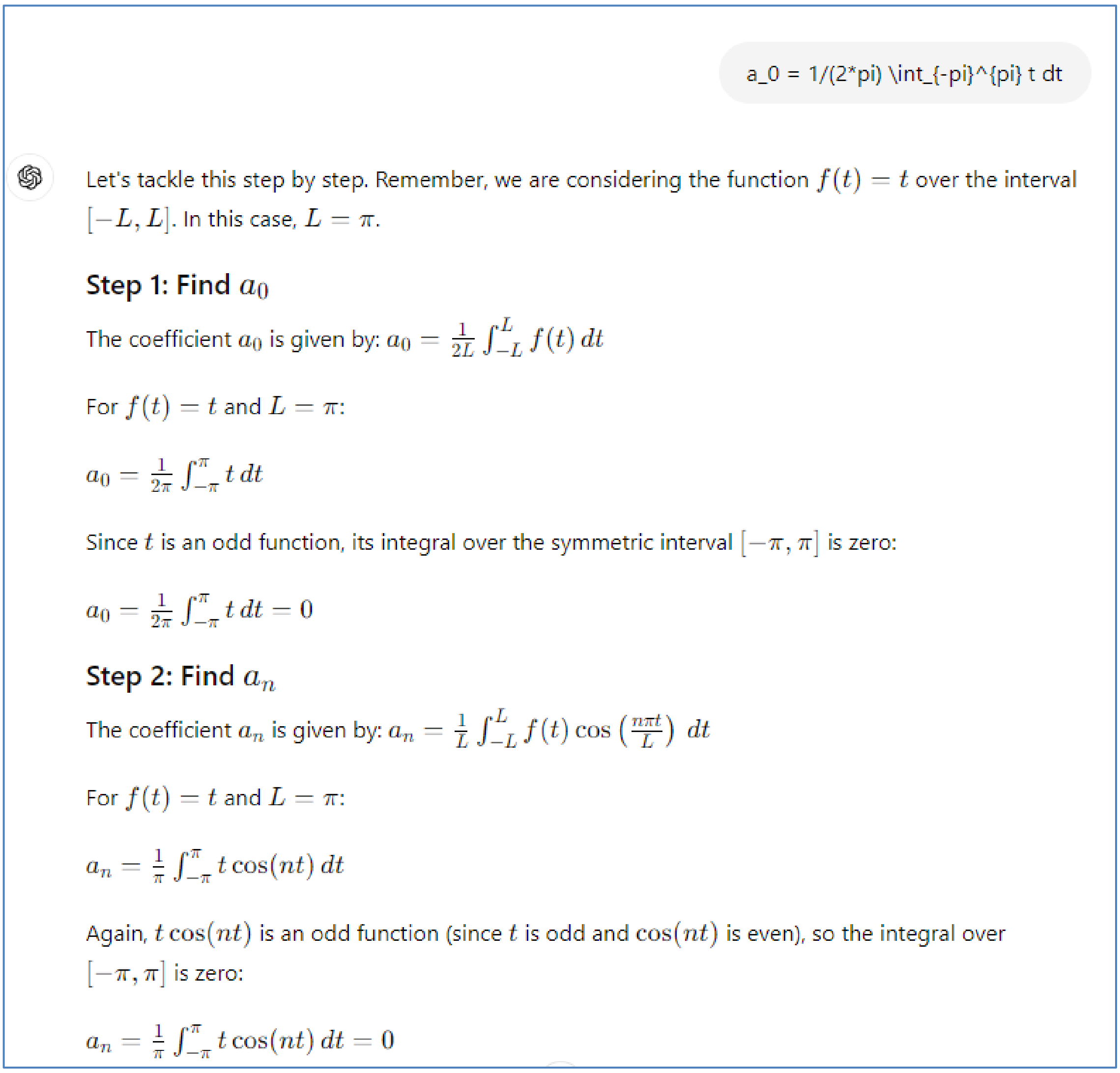

Figure 3. Building upon Figure 2, this illustrates a challenge in aligning student inputs with responses from GenAI tutors.

However, the student then encounters a second issue: rather than verifying the student's solution or guiding the next step based on the input, the GenAI generates complete solutions for all remaining steps. Although the explanations are clear and detailed, this approach is unhelpful, as students receive neither feedback on their existing work nor the opportunity to attempt the remaining steps independently. The tendency for GenAI tutors to provide the entire solution at once, after only one or two steps, occurs consistently across all platforms. This pattern poses challenges for learning. While students may review the solution mentally for confirmation, this passive engagement does not support effective learning or retention. Instead, students must perform as much of the work as possible through active practice [16], and the learning process requires cognitive effort and productive struggle [50].

The required effort may stem from students' inability to fully trust the solutions provided by GenAI tutors, as shown in Table 1, necessitating additional effort to verify correctness. Even with explicit instructions, some students may overlook or struggle with this verification process [51]. A potential workaround involves two-stage verification, such as cross-checking results with Wolfram or other more reliable computational tools [11]; however, this may undermine the purpose of using a GenAI tutor. One of the fundamental advantages of a tutor is minimizing the risk of acquiring incorrect knowledge or adopting flawed learning strategies [16]. This underscores the potential value of this study in assessing the likelihood of exposure to inaccurate knowledge on a topic-specific basis. Beyond the topics examined here, teaching staff should evaluate subject areas independently before implementing a GenAI tutor for their students.

All GenAI tutors performed most poorly on mechanical engineering content, where they were most likely to produce incorrect solutions. For instance, in calculating pressure drop in a water pipe, the correct answer was provided in only 3 out of 8 cases. The appropriate formula for this calculation is the implicit Colebrook-White equation, or approximations such as the Haaland or Swamee-Jain equations. In some instances, a valid equation was applied in an incorrect context (e.g., given the water velocity, a different formula was required). In others, an entirely incorrect equation was used. Figure 4 illustrates an example in which GPT provides an erroneous equation for a thermodynamics problem involving the Brayton cycle. The efficiency equation in GPT's response resembles the correct version from the Wikipedia page, but temperature (T) is incorrectly substituted in place of pressure (P).

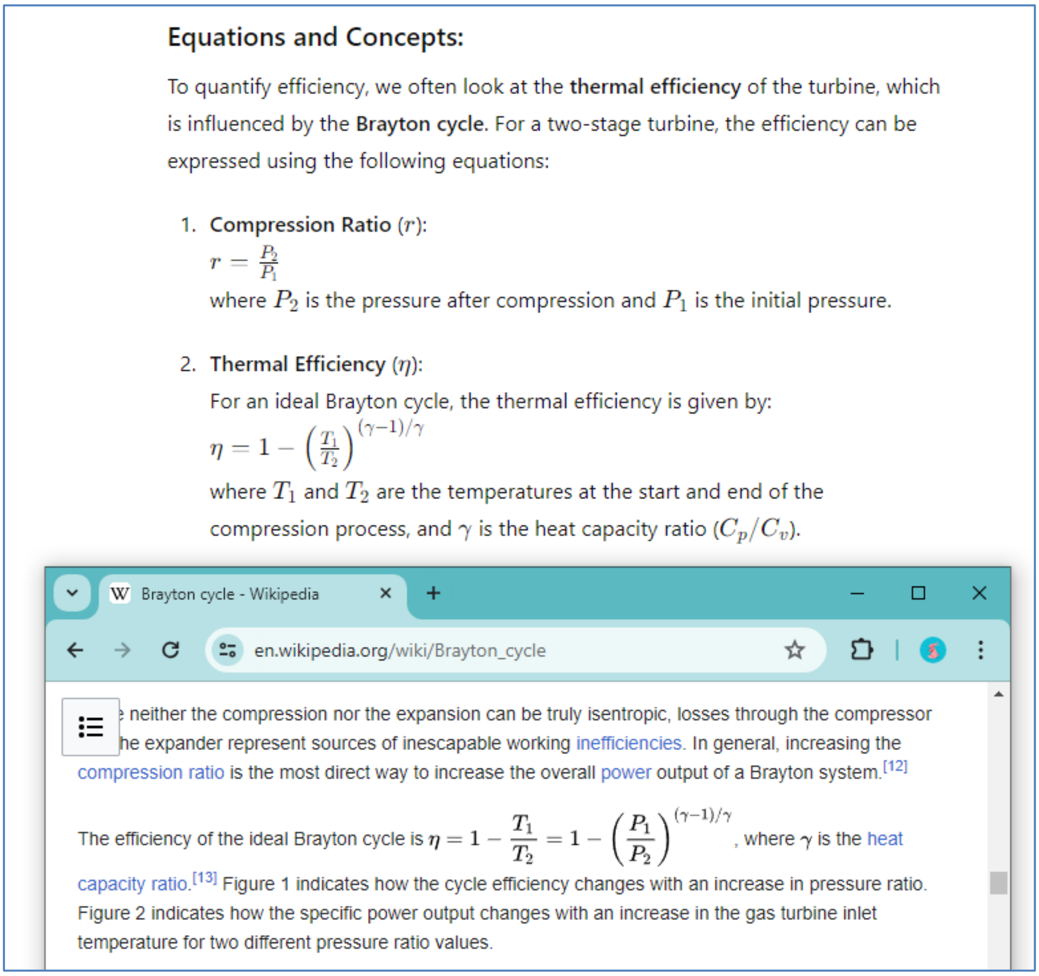

Figure 4. Example of wrongly substituting variable T and P in a given formula for a thermodynamics question relating to the Brayton cycle.

For many of the topics assessed, as indicated in Table 1 by values between 3 and 4, the most common issues were minor rather than serious errors. Typically, if students correctly followed the provided explanations, an answer that was not fully simplified was classified as a minor issue rather than a significant error. For instance, Figure 5 illustrates an integration calculation performed by Tutor Me, in which the final result was algebraically incomplete. Such omissions, while not mathematically incorrect, may lead to student confusion if the expected form of the solution is not clearly specified.

3. For $1 < t \leq 2$:
   - During this interval, $h(t-\tau)$ shifts further, and the overlap changes as both parts of the triangular pulse $x(\tau)$ contribute.
   - The integral splits into two parts: $y(t) = \int_{t-1}^{1} \tau \cdot 1\, d\tau + \int_{1}^{t} (2-\tau) \cdot 1\, d\tau$
   - Evaluate the first integral: $\int_{t-1}^{1} \tau\, d\tau = \left[\frac{\tau^2}{2}\right]_{t-1}^{1} = \frac{1}{2} - \frac{(t-1)^2}{2}$
   - Evaluate the second integral: $\int_{1}^{t} (2-\tau)\, d\tau = \left[2\tau - \frac{\tau^2}{2}\right]_{1}^{t} = (2t - \frac{t^2}{2}) - (2 \cdot 1 - \frac{1^2}{2}) = (2t - \frac{t^2}{2}) - (2 - \frac{1}{2}) = 2t - \frac{t^2}{2} - \frac{3}{2}$
   - Combine the results: $y(t) = \left(\frac{1}{2} - \frac{(t-1)^2}{2}\right) + \left(2t - \frac{t^2}{2} - \frac{3}{2}\right) = \frac{1}{2} + 2t - \frac{t^2}{2} - \frac{3}{2} - \frac{(t-1)^2}{2}$
   - Simplify the combined result: $y(t) = 2t - \frac{t^2}{2} - \frac{1}{2} - \frac{(t^2-2t+1)}{2}$
   - Further simplification: $y(t) = 2t - \frac{t^2}{2} - \frac{t^2}{2} + t - \frac{1}{2}$

Figure 5. Example of Tutor Me failing to provide the best-simplified answer, which could lead to student misunderstanding.

Another limitation identified during pilot testing [15] was the unreliability of diagram generation by GenAI tutors. Generated diagrams were typically irrelevant and contained incorrect labels or inaccurate information. As a result, the prompt was updated to explicitly prohibit diagram generation. A significant drawback of this restriction arises in contexts requiring geometrical interpretation. This limitation is partially mitigated by the GenAI tutor's capacity to generate detailed verbal descriptions and analogies. For instance, when learning about the volume of solids of revolution, a request for a visual description of the method yielded a clearly articulated response from ChatGPT-4, as illustrated in Figure 6. Nevertheless, a diagram would be substantially more effective in conveying the methodology described here.

When GenAI tutors were tasked with generating charts, output accuracy varied by topic. Performance was generally strong for mathematics-based charts: equations could be accurately plotted during tutorial discussions or upon request, and required regions could be shaded appropriately. Difficulties emerged with specialized chart types. In one test case, GenAI tutors were asked to demonstrate the use of a psychrometric chart. While the explanation of how these charts function was accurate and thorough, the support provided is limited for users unfamiliar with such charts or those who rely more heavily on visual learning.

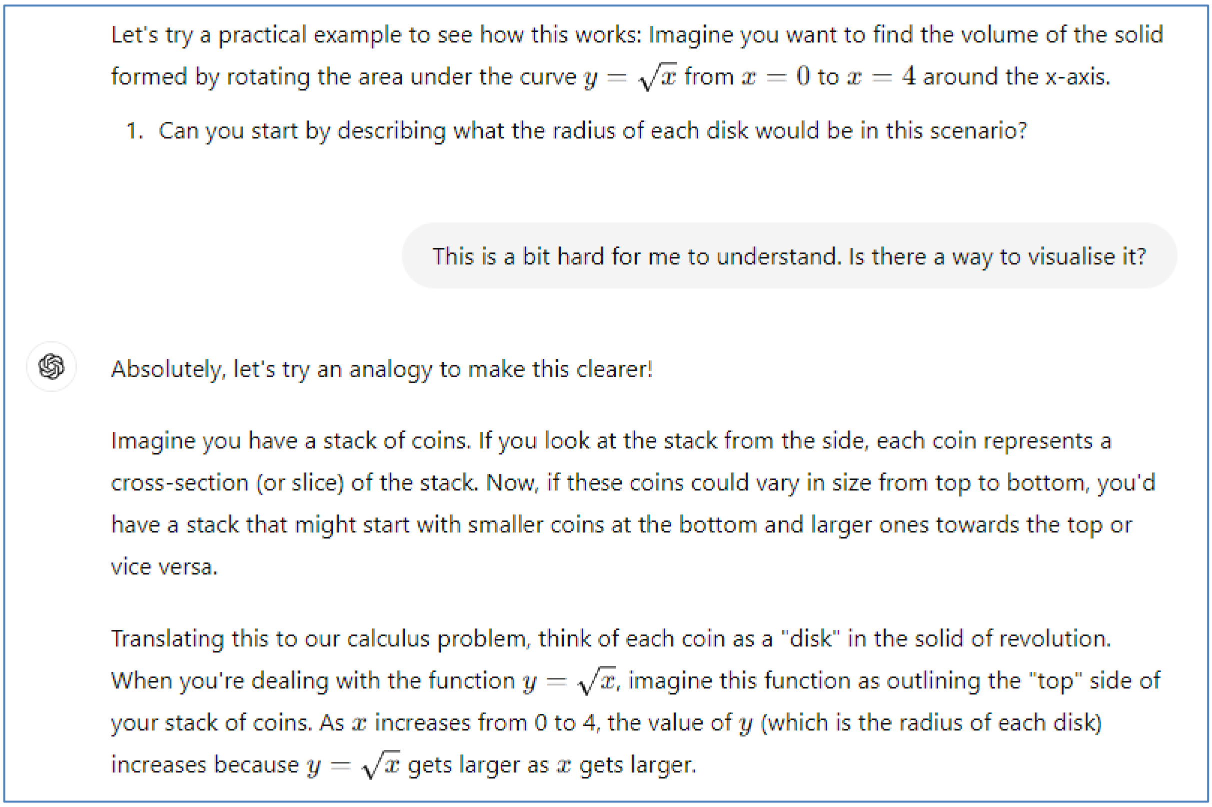

Figure 6. Example of GenAI describing a visualization in words rather than a diagram. While capable of textual description, this approach is less effective than directly providing a diagram. The prompt was specifically designed to prevent diagram generation due to their potential inaccuracy.

A commonly observed issue is ChatGPT's tendency, particularly in GPT-4, to repeat large portions of text verbatim. For instance, after defining all terms in an equation, ChatGPT may restate the entire equation and definitions in the subsequent response. This redundancy can be tedious to read and may significantly impair the student's experience, potentially leading to disengagement from learning.

As outlined in Section A, the GenAI tutors provided distinct learning experiences, resulting in varied strengths and weaknesses. During preliminary assessments of student readiness, both GPT-4 and Wolfram GPT emphasized foundational knowledge. While beneficial for novices, this approach may not enrich students with strong prior knowledge, potentially causing them to progress slowly on advanced topics or lose interest. In contrast, Tutor Me GPT attempts to promote deeper learning by posing complex questions even to students with weak foundations, a strategy that risks overwhelming them and reducing learning efficiency.

Despite their individual strengths, GPT-4, GPT-4o, Wolfram GPT, and Tutor Me GPT exhibit limitations in instructional depth, conceptual understanding, and practical application. Although these models attempt to personalize instruction based on student backgrounds, they often fail to leverage this information for targeted follow-up, resulting in superficial content that does not adequately address individual learning needs. GPT-4, GPT-4o, and Wolfram GPT, while effective at extending basic concepts, frequently provide insufficient explanations, making it difficult for students with weaker foundations to grasp advanced material.

While thorough in its approach, Tutor Me GPT employs complex teaching strategies and occasionally provides incorrect answers, which can overwhelm students unfamiliar with the subject. For example, it guided students through a challenge question but produced a wrong result due to misprocessing "I", leading to a missing coefficient of 1/2. Furthermore, Tutor Me GPT's emphasis on computational exercises often fails to explore complex concepts, such as system stability in signal processing, resulting in a surface-level understanding of critical topics. Similarly, Wolfram GPT's frequent prioritization of computation over conceptual insight may limit students' ability to understand underlying principles and apply knowledge in real-world contexts, thereby hindering holistic development. An example illustrating differences in conceptual insight between GPT-4o and Wolfram GPT is shown in Figure 7. After completing a calculation review, GPT-4o assessed students on related conceptual understanding. In contrast, Wolfram GPT presented additional computational problems to further test performance.

ChatGPT-4o Using the power loss formula $P = I^2 R$:
1. Current I: Given as $100\,\mathrm{A}$.
2. Resistance R: Calculated as $10\,\Omega$.

Plugging in these values, what is the power loss P over the 100 miles of transmission line?

Student

$$P = 100 \cdot 100 \cdot 10 = 100000\,\mathrm{W}$$

ChatGPT-4o Perfect! The power loss P over the 100 miles of transmission line is $100^2 \times 10 = 100000\,\mathrm{W}$, or $100\,\mathrm{kW}$.
Summarizing the Calculation:
1. Current (I): 100 A.
2. Resistance (R): 10 Ω.
3. Power Loss (P): $100^2 \cdot 10 = 100000\,\mathrm{W}$.

Let's review:
1. Why is it important to use high voltage for long-distance transmission?
2. How do step-up and step-down transformers work in the transmission process?

Wolfram GPT Excellent! Your calculation for the magnitude of the voltage drop is correct. Here's the breakdown: $|V| = \sqrt{(\mathrm{Re}(V))^2 + (\mathrm{Im}(V))^2}$
Given: $\mathrm{Re}(V) = 10000\,\mathrm{V}, \mathrm{Im}(V) = 5000\,\mathrm{V}$
Calculate the squared components: $(\mathrm{Re}(V))^2 = (10000)^2 = 100000000\,\mathrm{V}^2$
$(\mathrm{Im}(V))^2 = (5000\pi)^2 = 25000000\pi \approx 25000000 \times 9.8696 = 246740000\,\mathrm{V}^2$
So,

$$|V| = \sqrt{100000000 + 246740000} = \sqrt{346740000} \approx 18584.62\,\mathrm{V}$$

Now, let's think about what this result tells us. The voltage drop across the transmission line is substantial, which shows the importance of managing transmission line parameters to minimize losses.

Review:
1. High voltage reduces current, thereby reducing I^2R losses.
2. Transmission lines have inherent resistance and inductance that contribute to voltage drops.
3. Complex impedance combines resistance and reactance, and the voltage drop can be calculated using Ohm's Law.

Practice Question
Let's apply this to another scenario. Suppose you have a transmission line with the following parameters:
- Resistance $R = 0.1\,\Omega/\mathrm{km}$
- Inductance $L = 0.5\,\mathrm{mH/km}$
- The line is 50 km long, and the current is 600 A at a frequency of 60 Hz.

Figure 7. An example of the differences in conceptual insights provided by ChatGPT-4o and Wolfram GPT.

While comprehensive, GPT-4's integration of theoretical concepts with practical applications may present excessive cognitive load for students with weaker foundational knowledge, potentially diminishing motivation and reducing learning efficiency. Similarly, both GPT-4o and Tutor Me GPT adequately address computational procedures and fundamental concepts but insufficiently emphasize practical applications, which may hinder the development of students' analytical and problem-solving abilities when confronting real-world engineering challenges.

Overall, these generative AI (GenAI) models frequently struggle to balance theoretical instruction with practical application, computational precision, and conceptual depth, resulting in inconsistent learning experiences. The computational emphasis of Wolfram GPT, the limited practical context in Tutor Me GPT, and the theoretical bias of GPT-4o collectively contribute to variability in instructional depth and coherence. To address these limitations, students would benefit from strategically aligning their individual learning profiles with the specific capabilities of each model.

5.3. Identified opportunities

Significant opportunities exist to enhance the GenAI tutors evaluated in this study. Some improvements could be achieved through refined prompting strategies, a promising avenue for future research, while others require direct technological advancements. A key opportunity involves developing adaptive learning pathways that dynamically adjust content complexity based on initial diagnostic assessments, ensuring students receive personalized support aligned with their specific learning needs. Enhanced diagnostic tools could further refine these pathways by accurately identifying knowledge gaps and enabling more targeted instructional interventions.

Integrating theoretical knowledge with practical applications, where GPT-4 demonstrated particular strength, represents another critical area for improvement. Incorporating real-world case studies and authentic problem-solving scenarios would effectively bridge the gap between abstract concepts and applied skills, especially in disciplines such as Power Engineering, where understanding real engineering contexts significantly enhances learning outcomes. Strengthening the integration of computational and conceptual instruction with practical applications in both Tutor Me GPT and GPT-4o would better equip students for real-world problem-solving. The efficacy of GenAI in real-world engineering applications has been demonstrated in project-based engineering work [52].

Striking an appropriate balance between computational exercises and conceptual instruction is essential. Wolfram GPT and Tutor Me GPT could be improved by incorporating more advanced content and complex theoretical concepts, thereby fostering deeper understanding beyond basic calculations. Strategically combining the strengths of different models, such as Wolfram GPT's computational accuracy with GPT-4's conceptual insights, could yield a more effective pedagogical approach. Additionally, optimizing tiered instruction methods could provide appropriately differentiated challenges for students across varying skill levels. For instance, integrating sophisticated conceptual discussions alongside computational tasks can simultaneously support advanced learners and reinforce foundational knowledge for others.

Ultimately, synthesizing the pedagogical strengths of each model, the creative examples from GPT-4, the computational rigor from Wolfram GPT, and the practical application focus from Tutor Me GPT, could produce a more balanced, engaging, and effective GenAI tutoring experience. This integration presents a substantial opportunity for future research and development.

5.4. Summary

Table 3 summarizes the observations and reflections provided by the three research assistants. It highlights the strengths, weaknesses, and opportunities associated with each GenAI tutor. This data can inform improvements to prompting strategies or guide the design of future model iterations.

Table 3. Strengths, weaknesses and opportunities for ChatGPT-4/4o, Wolfram GPT and Tutor Me GPT.

GenAI Tutor	Strengths	Weaknesses	Opportunities
ChatGPT-4	- Excellent in explaining intuition behind formulas and real-life applications. - Creative use of examples and illustrations for teaching complex concepts. - Effective at adapting to student levels and maintaining student engagement through personalized support.	- Prone to inaccuracies in calculations, often using incorrect formulas or context. - Repetition of content can be tedious and distracting. - Sometimes drifts off-topic or includes irrelevant content.	- Enhance integration of creative examples with problem-solving guidance to deepen conceptual understanding. - Enhance diagnostic tools to identify student knowledge gaps more accurately.
ChatGPT-4o	- Deep understanding of student needs, allowing for more tailored assessments and personalized learning. - Consistently high-quality, adaptable teaching strategies across different learning levels.	- Overemphasis on basic concepts may not challenge advanced students, leading to boredom. - May fail to fully utilize personalized data in subsequent instruction, limiting deeper engagement.	- Introduce integrated teaching methods that combine theory with practical applications, enhancing real-world problem-solving skills. - Create more balanced teaching by integrating advanced content with computational exercises to challenge students effectively.
Wolfram GPT	- Strong in computational teaching, providing detailed example-based solutions that aid comprehension. - Effective at multi-level problem-solving guidance and consistent solution quality.	- Focus on computation over conceptual understanding can limit students' grasp of underlying principles. - Limited in demonstrating practical application and advanced concepts, especially for students needing higher-order thinking development. - Heavy focus on computation may lead to superficial learning and neglect of conceptual depth.	- Strengthen integration of theoretical and practical instruction, incorporating real engineering problems to improve students' application skills.
Tutor Me GPT	- Personalized training through detailed questioning and targeted assessment of course concepts. - Effective at testing students' understanding and providing targeted feedback for improved learning.	- Complexity in questioning can overwhelm students with weaker foundations, affecting their learning efficiency. - Over-reliance on basic computational exercises limits exposure to advanced topics, affecting students' deep understanding of subjects like system stability.	- Enhance examination of complex concepts and practical applications to provide a more balanced learning experience across student capabilities. - Introduce more advanced content and layered teaching methods to cater to students' individual learning paths and increase their problem-solving abilities in real-world contexts.

6. Conclusions

This study addressed two research questions. The first question was, ‘How accurate and reliable are the selected Generative AI tools in providing tutoring for engineering and mathematics content, and which tool demonstrates the highest performance?’ The findings indicate that accuracy is highly dependent on subject matter and specific topics, suggesting variability in the underlying training data and methodologies. In most cases, at a minimum, the GenAI tutors introduced at least one minor, non-consequential error, requiring students to identify and correct it. Given the limited scope of data collection, the probability of such errors occurring appears high. Since hallucinations have not yet been fully mitigated, it is critical to consider how students can develop awareness of these inaccuracies. With accuracy falling short of 100%, the central issue becomes how effectively students can manage this inherent risk. Would students be willing to use these tools knowing they might be guided toward incorrect solutions? The traditional role of a tutor is to minimize the risk of acquiring faulty knowledge [16]. If hallucinations are eventually eliminated, the balance of risk may shift, potentially making human tutors the greater source of error.

Could awareness of potential inaccuracies foster greater vigilance, encouraging students to follow and understand each step carefully to avoid internalizing incorrect information? Laboratory studies suggest that embedding error detection and correction techniques into learning processes can enhance the student experience [53], raising the possibility that the current limitations in accuracy might, under certain conditions, serve as a pedagogical advantage. While RAs classified many of the observed errors as inconsequential, it remains unclear whether students would perceive them similarly. This highlights the need for a scaffolded study that extends the simulation process under supervised conditions to better understand student engagement before broader implementation is recommended (if at all, until accuracy improves).

Potential short-term strategies could increase confidence in the use of these tools. For instance, since Wolfram GPT has been shown to answer computational questions more reliably than GPT-4 in direct comparison [11], students could copy and paste exercises from tutor mode into Wolfram GPT to verify whether the final answers align. While this approach does not guarantee correctness, even a minor discrepancy could prompt caution and help develop students’ error detection skills.

The second research question was, ‘How effectively do the selected Generative AI tools deliver a comprehensive and engaging tutorial experience?’ The study found that all GenAI tutors performed reasonably well, each exhibiting distinct strengths and weaknesses. This variation is comparable to human tutors or teachers, who also differ in pedagogical approaches and delivery styles. Just as human educators can be trained to improve their methods, the insights from this study can inform researchers and students in refining prompts or guiding future model development. Across all subjects and assessments, ChatGPT-4 was observed to provide the most effective tutorial experience, whereas Tutor Me GPT performed the weakest; however, individual student experiences may vary.

Although the RAs assigned scores indicating positive learning experiences, the simulation offers no direct evidence of GenAI’s actual effectiveness in supporting learning. It assumes that human-to-machine knowledge transfer is efficient, yet evidence suggests this may not hold true due to factors such as lack of empathy, impacts on short- and long-term memory consolidation, and the promotion of multitasking behaviors [54]. Once again, this underscores the necessity of further scaffolded simulation studies before wider adoption.

In summary, this work demonstrates that the application of GenAI as a tutor remains a work in progress. It suggests that, in the near term, GenAI can support existing learning practices but should not replace them. The findings show considerable promise and justify further investigation into the educational impacts of GenAI tutoring. While multimodal GenAI tutors, incorporating voice, enhanced visuals, and handwritten input capabilities, may represent a turning point, this study establishes a relevant benchmark against which future advancements can be measured.

Author contributions

S.N. conceptualized the study, designed the methodology, undertook the core analysis, and wrote the original draft. B.A.V. completed the simulation tasks, data collection, contributed to the first draft, and performed the statistical analysis. Y.D. and A.H. completed the simulation tasks, data collection, and contributed to the first draft. All other authors provided subject-specific insights across all aspects, ensured alignment of simulations with subject content, and supported the writing of multiple revisions. All authors reviewed and approved the final manuscript.

Use of Generative-AI tools declaration

ChatGPT-4/4o, Wolfram GPT, and Tutor Me GPT were used in this study as outlined in the methodology. Additionally, ChatGPT-4o was used to succinctly reframe certain sentences or ideas generated by the authors. Grammarly was employed for grammar correction. All AI-generated content was critically reviewed and verified by the authors, who take full responsibility for the final manuscript.

Acknowledgments

We thank the reviewers for their constructive feedback. This work is an initiative of the Australasian Artificial Intelligence in Engineering Education Centre (AAIEEC).

We thank the University of Wollongong for supporting this research through a 2024 Learning & Teaching Innovation Grant.

Conflict of interest

The authors declare no conflicts of interest in this paper.

Prof. Sasha Nikolic is an editorial board member for STEM Education and was not involved in the editorial review or the decision to publish this article.

Ethics declaration

Human research ethics approval was not required for this work.

References

1. Bloom, B.S., The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher*, 1984, 13(6): 4–16. https://doi.org/10.3102/0013189x013006004
2. Kim, W.H. and Kim, J.H., Individualized AI Tutor Based on Developmental Learning Networks. *IEEE Access*, 2020, 8: 27927-37. https://doi.org/10.1109/ACCESS.2020.2972167

3. Nwana, H.S., Intelligent Tutoring Systems: An Overview. *Artificial Intelligence Review*, 1990, 4(4): 251–277. https://doi.org/10.1007/BF00168958
4. Mollick, E., *Co-Intelligence: Living and Working with AI*, London: WH Allen, 2024.
5. Belkina, M., Daniel, S., Nikolic, S., Haque, R., Lyden, S., Neal, P., et al., Implementing Generative AI (GenAI) in Higher Education: A Systematic Review of Case Studies. *Computers and Education: Artificial Intelligence*, 2025, 100407. https://doi.org/https://doi.org/10.1016/j.caeai.2025.100407
6. Ahmed, Z., Shanto, S.S. and Jony, A.I., Potentiality of Generative AI Tools in Higher Education: Evaluating ChatGPT's Viability as a Teaching Assistant for Introductory Programming Courses. *STEM Education*, 2024, 4(3): 165–182. https://doi.org/10.3934/steme.2024011
7. Eddine, R.J., Gide, E. and Al-Sabbagh, A., Generative AI in Higher Education: A Cross-Sector Analysis of ChatGPT's Impact on STEM, Social Sciences, and Healthcare. *STEM Education*, 2025, 5(5): 757–801. https://doi.org/10.3934/steme.2025035
8. Nikolic, S., Wentworth, I., Sheridan, L., Moss, S., Duursma, E., Jones, R.A., et al., A Systematic Literature Review of Attitudes, Intentions and Behaviours of Teaching Academics Pertaining to AI and Generative AI (GenAI) in Higher Education: An Analysis of GenAI Adoption Using the Utaut Framework. *Australasian Journal of Educational Technology*, 2024, 40(6): 56–75. https://doi.org/10.14742/ajet.9643
9. Nikolic, S., Daniel, S., Haque, R., Belkina, M., Hassan, G.M., Grundy, S., et al., ChatGPT Versus Engineering Education Assessment: A Multidisciplinary and Multi-Institutional Benchmarking and Analysis of This Generative Artificial Intelligence Tool to Investigate Assessment Integrity. *European Journal of Engineering Education*, 2023, 48(4): 559–614. https://doi.org/10.1080/03043797.2023.2213169
10. Cong-Lem, N., Soyoof, A. and Tsering, D., A Systematic Review of the Limitations and Associated Opportunities of ChatGPT. *International Journal of Human–Computer Interaction*, 2025, 41(7): 3851–66. https://doi.org/10.1080/10447318.2024.2344142
11. Haynes, K. and Brendle, J., Cross-Age Math Tutoring of Kindergarten and First Grade Students by Middle School Tutors. *International Journal of Education in Mathematics, Science and Technology*, 2019, 7(3): 238–250.
12. Nikolic, S., Sandison, C., Haque, R., Daniel, S., Grundy, S., Belkina, M., et al., ChatGPT, Copilot, Gemini, Scispace and Wolfram Versus Higher Education Assessments: An Updated Multi-Institutional Study of the Academic Integrity Impacts of Generative Artificial Intelligence (GenAI) on Assessment, Teaching and Learning in Engineering. *Australasian Journal of Engineering Education*, 2024, 29(2): 126–53. https://doi.org/10.1080/22054952.2024.2372154
13. Kalai, A.T., Nachum, O., Vempala, S.S. and Zhang, E., *Why Language Models Hallucinate*, 2025. Available from: https://cdn.openai.com/pdf/d04913be-3f6f-4d2b-b283-ff432ef4aaa5/why-language-models-hallucinate.pdf.
14. Zhang, H., Da, J., Lee, D., Robinson, V., Wu, C., Song, W., et al., A Careful Examination of Large Language Model Performance on Grade School Arithmetic. *Advances in Neural Information Processing Systems*, 2024, 37: 46819–46836.
15. Nikolic, S., Heath, A., Vu, B.A., Daniel, S., Alimardani, A., Sandison, C., et al., Prompt

Potential: A Pilot Assessment of Using Generative Artificial Intelligence (ChatGPT-4) as a Tutor for Engineering and Maths. *52nd Annual Conference of the European Society for Engineering Education (SEFI)*, 2024. http://dx.doi.org/10.2139/ssrn.5121997
16. Merrill, D.C., Reiser, B.J., Ranney, M. and Trafton, J.G., Effective Tutoring Techniques: A Comparison of Human Tutors and Intelligent Tutoring Systems. *The Journal of the learning sciences*, 1992, 2(3): 277–305. https://doi.org/10.1207/s15327809jls0203_2
17. Feng, S., Magana, A.J. and Kao, D., A Systematic Review of Literature on the Effectiveness of Intelligent Tutoring Systems in STEM. In *2021 IEEE Frontiers in Education Conference (FIE)*, 2021, 1–9. https://doi.org/10.1109/FIE49875.2021.9637240
18. Lin, C.C., Huang, A.Y. and Lu, O.H., Artificial Intelligence in Intelligent Tutoring Systems toward Sustainable Education: A Systematic Review. *Smart Learning Environments*, 2023, 10(1): 41. https://doi.org/10.1186/s40561-023-00260-y
19. Liu, H., Zhang, Y. and Jia, J., The Design of Guiding and Adaptive Prompts for Intelligent Tutoring Systems and Its Effect on Students' Mathematics Learning. *IEEE Transactions on Learning Technologies*, 2024, 17: 1379–89. https://doi.org/10.1109/TLT.2024.3382000
20. VanLehn, K., Siler, S., Murray, C. and Baggett, W.B., What Makes a Tutorial Event Effective? *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, 2022, 1084–1089. https://doi.org/10.4324/9781315782416-195
21. Kulik, J.A. and Fletcher, J.D., Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. *Review of educational research*, 2016, 86(1): 42–78. https://doi.org/10.3102/0034654315581420
22. Singh, N., Gunjan, V.K. and Nasralla, M.M, A Parametrized Comparative Analysis of Performance between Proposed Adaptive and Personalized Tutoring System "Seis Tutor" with Existing Online Tutoring System. *IEEE Access*, 2022, 10: 39376–39386. https://doi.org/10.1109/ACCESS.2022.3166261
23. Aleven, V., Rowe, J., Huang, Y. and Mitrovic, A., Domain Modeling for AIED Systems with Connections to Modeling Student Knowledge: A Review. *Handbook of Artificial Intelligence in Education*, 2023, 127–169. https://doi.org/10.4337/9781800375413.00017
24. Romoser, M., An Autonomous Intelligent Driving Simulation Tutor for Driver Training and Remediation: A Concept Paper. In *Driving Assessment Conference*, 2011. https://doi.org/10.17077/drivingassessment.1438
25. Nikolic, S., Ros, M., Al-Abdeli, Y.M. and Fairweather, H., Beyond Assessment Security: A Critical Policy Analysis of Four Alternative Strategies to Uphold Academic Integrity and Adopt the GenAI Transformation of Teaching and Learning for an Accredited Engineering Degree. *STEM Education*, 2025, 5(4): 564–586. https://doi.org/10.3934/steme.2025027
26. Rudolph, J., Tan, S. and Tan, S., ChatGPT: Bullshit Spewer or the End of Traditional Assessments in Higher Education?. *Journal of Applied Learning and Teaching*, 2023, 6(1): 342–363. https://doi.org/https://doi.org/10.37074/jalt.2023.6.1.9
27. Vujinović, A., Luburić, N., Slivka, J. and Kovačević, A., Using ChatGPT to Annotate a Dataset: A Case Study in Intelligent Tutoring Systems. *Machine Learning with Applications,* 2024, 16: 100557. https://doi.org/https://doi.org/10.1016/j.mlwa.2024.100557
28. Piro, L., Bianchi, T., Alessandrelli, L., Chizzola, A., Casiraghi, D., Sancassani, S., et al., Mylearningtalk: An LLM-Based Intelligent Tutoring System. In *International Conference on*

Web Engineering, 2024, 428–431. https://doi.org/10.1007/978-3-031-62362-2_39
29. Mustapha, K.B., Yap, E.H. and Abakr, Y.A., Bard, ChatGPT and 3dGPT: A Scientometric Analysis of Generative AI Tools and Assessment of Implications for Mechanical Engineering Education. *Interactive Technology and Smart Education*, 2024, 21(4): 588–624. https://doi.org/10.1108/ITSE-10-2023-0198
30. Ghassemi, M., Birhane, A., Bilal, M., Kankaria, S., Malone, C., Mollick, E., et al., ChatGPT One Year On: Who Is Using It, How and Why?. *Nature*, 2023, 624(7990): 39–41. https://doi.org/10.1038/d41586-023-03798-6
31. Chen, X., Mitrovic, A. and Mathews, M., Learning from Worked Examples, Erroneous Examples, and Problem Solving: Toward Adaptive Selection of Learning Activities. *IEEE Transactions on Learning Technologies*, 2020, 13(1): 135–49. https://doi.org/10.1109/TLT.2019.2896080
32. Nikolic, S., Suesse, T.F., Grundy, S., Haque, R., Lyden, S., Hassan, G.M., et al., Laboratory Learning Objectives: Ranking Objectives across the Cognitive, Psychomotor and Affective Domains within Engineering. *European Journal of Engineering Education*, 2024, 49(3): 457–473. https://doi.org/10.1080/03043797.2023.2248042
33. Honig, C., Rios, S. and Desu, A., Generative AI in Engineering Education: Understanding Acceptance and Use of New GPT Teaching Tools within a UTAUT Framework. *Australasian Journal of Engineering Education*, 2025, 30(1): 80–92. https://doi.org/10.1080/22054952.2025.2467500.
34. OpenAI. "Hello GPT-4o." OpenAI. https://openai.com/index/hello-gpt-4o/ (accessed 13/05/2024.
35. Necesal, P. and Pospıšil, J., Experience with Teaching Mathematics for Engineers with the Aid of Wolfram Alpha. *Proceedings of the World Congress on Engineering and Computer Science*, 2012, 1: 271–274.
36. ChatGPT Gets Its “Wolfram Superpowers”!, 2023, accessed 23/03. Available from: https://writings.stephenwolfram.com/2023/03/chatgpt-gets-its-wolfram-superpowers/.
37. Kelly, D.P. and Rutherford, T., Khan Academy as Supplemental Instruction: A Controlled Study of a Computer-Based Mathematics Intervention. *The International Review of Research in Open and Distributed Learning*, 2017, 18(4). http://orcid.org/0000-0002-6463-7022
38. Vidergor, H.E. and Ben-Amram, P., Khan Academy Effectiveness: The Case of Math Secondary Students' Perceptions. *Computers & Education*, 2020, 157: 103985. https://doi.org/https://doi.org/10.1016/j.compedu.2020.103985
39. KhanLabs. "World-class AI for education." https://www.khanacademy.org/khan-labs (accessed 21/03/23).
40. Khan Academy. "Tutor Me." https://chatgpt.com/g/g-hRCqiqVlM-tutor-me (accessed 01/07/2024).
41. Scarlatos, A., Editorial Overview: Special Issue on Artificial Intelligence in Education. *Journal of Educational Technology Systems*, 2024, 52(3): 299–300. https://doi.org/10.1177/00472395241236997
42. Pardos, Z.A. and Bhandari, S., ChatGPT-Generated Help Produces Learning Gains Equivalent to Human Tutor-Authored Help on Mathematics Skills. *Plos one*, 2024, 19(5): e0304013. https://doi.org/10.1371/journal.pone.0304013

43. Frankford, E., Sauerwein, C., Bassner, P., Krusche, S. and Breu, R., AI-Tutoring in Software Engineering Education. *Proceedings of the 46th International Conference on Software Engineering: Software Engineering Education and Training*, 2024, 309–319. https://doi.org/10.1145/3639474.3640061
44. Nahar, N., Exploring Student Perspectives on AI-Generated Feedback Using a Socratic Method Chatbot. *Journal of learning development in higher education*, 2025, 37: 1–11. https://doi.org/10.47408/jldhe.vi37.1724
45. OpenAI. "Introducing study mode." https://openai.com/index/chatgpt-study-mode/ (accessed 29/07/25).
46. Mollick, E. and Mollick, L., General Tutor - GPT4. Available from: https://www.moreusefulthings.com/student-exercises (accessed 01/02/24).
47. Gregory, S., O'Connell, J., Butler, D., McDonald, M., Kerr, T., Schutt, S., et al., New Applications, New Global Audiences: Educators Repurposing and Reusing 3D Virtual and Immersive Learning Resources. *ASCILITE Annual Conference,* 2015. https://doi.org/10.14742/apubs.2015.980
48. Friedman, M., The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, 1937, 32(200): 675–701. https://doi.org/10.1080/01621459.1937.10503522
49. Liu, J., Wang, C., Liu, Z., Gao, M., Xu, Y., Chen, J., et al., A Bibliometric Analysis of Generative Ai in Education: Current Status and Development. *Asia Pacific Journal of Education*, 2024, 44(1): 156–75. https://doi.org/10.1080/02188791.2024.2305170
50. Shein, E., The Impact of AI on Computer Science Education. *Communications of the ACM*, 2024, 13–15. https://doi.org/10.1145/3673428
51. Alimardani, A., Borderline Disaster: An Empirical Study on Student Usage of GenAI in a Law Assignment. *IEEE Transactions on Technology and Society*, 2025, 1–10. https://doi.org/10.1109/TTS.2025.3540978
52. Nikolic, S., Quince, Z., Lindqvist, A.L., Neal, P., Grundy, S., Lim, M., et al., Project-Work Artificial Intelligence Integration Framework (PAIIF): Developing a CDIO-Based Framework for Educational Integration. *STEM Education*, 2025, 5(2): 310–32. https://doi.org/10.3934/steme.2025016
53. Nikolic, S., Ritz, C., Vial, P.J., Ros, M. and Stirling, D., Decoding Student Satisfaction: How to Manage and Improve the Laboratory Experience. *IEEE Transactions on Education*, 2014, 58(3): 151–58. https://doi.org/http://dx.doi.org/10.1109/TE.2014.2346474
54. Horvath, Jared Cooney, 3 Critical Problems GenAI Poses for Learning. *Harvard Business Publishing*, 2024. Available from: https://hbsp.harvard.edu/inspiring-minds/the-limits-of-gen-ai-educators-in-higher-ed?cid=organic-social%7Clinkedin%7C2024-08-19-4-the-limits-of-gen-ai-educators-in-higher-ed%7Cnone%7Cthought-lead-inspiring-minds%7Cunknown%7Cinspiring-minds-article%7Caug2024.

Appendix A

The GenAI Tutor Prompt:

You are an upbeat, encouraging tutor who will be helping a university student with the topic of [...] in the subject [...]. Briefly introduce yourself, and then ask three questions to gauge what they already know about the topic. Wait for a response. Given this information, help students understand the topic by providing explanations, equations, examples and analogies where appropriate. Keep your responses short. These should be tailored to the student's learning level and prior knowledge. Then give the student a related question to work through. The question should test the students' understanding. Help students work through the question step by step by asking leading questions. Do not provide immediate answers or solutions to problems. Ask the student to explain their thinking. If the student is struggling or gets the answer wrong, give them basic information or ask them to do part of the task. If the student struggles, then be encouraging and give them some hints. Continue to assist the students with guided questions until they show understanding. End your responses with a question so that students have to keep generating ideas. Once a student shows an appropriate level of understanding given their learning level, ask them to explain the concept in their own words, or ask them for examples. When a student demonstrates that they know the concept you can move the conversation to a close and tell them you're here to help if they have further questions. *Never provide diagrams*.

Note: In the first line of the prompt, the context was explicitly provided to the GenAI tools to improve efficiency. As per the pilot [15], the prompt can be designed so that the GenAI tool asks the user for the lesson context. The prompt is easily changed to accommodate the pedagogical experience desired by the user.

Author's biography

Dr. Sasha Nikolic is an Associate Professor in the School of Engineering at the University of Wollongong with a PhD in Engineering Education. He currently serves as President for both the Australasian Association of Engineering Education (AAEE) and the Australasian Artificial Intelligence in Engineering Education Centre (AAiEEc), where he leads multi-institutional initiatives on Generative AI.

Bao Anh Vu is a PhD student and Associate Lecturer in Applied Statistics at the University of Wollongong. She is also a research fellow at the Australian National University.

Yang Di is a PhD student and Research Assistant in Computer Engineering at the University of Wollongong.

Dr. Ashley Heath is an Associate Lecturer in Mechanical Engineering at the University of Wollongong.

Dr. Son Lam Phung is a Professor in the School of Engineering at the University of Wollongong.

Dr. Xiaoping Lu is an Associate Professor in the School of Mathematics and Physics at the University of Wollongong.

Dr. David Hastie is a Senior Lecturer in the School of Engineering at the University of Wollongong.

Dr. Md Rabiul Islam is an Associate Professor in the School of Engineering at the University of Wollongong.

Dr. Le Chung Tran is an Associate Professor in the School of Engineering at the University of Wollongong.

Dr. Brad Stappenbelt is a Senior Lecturer in the School of Engineering at the University of Wollongong.