Leveraging RAG-Enhanced LLMs for Engineering Education: Design and Evaluation of AI-Tutor

Huiru Xie, Georgia Institute of Technology

Huiru Xie is a Ph.D. student in Electrical and Computer Engineering at Georgia Institute of Technology. She holds an M.S. in Computer Science from Georgia Tech and an M.E. in Electrical Engineering from Southeast University, China. Her research sits at the intersection of Machine Learning, Human-Computer Interaction, and Educational AI, with a focus on building intelligent systems that improve learning experiences and human-technology interaction.

Dr. Liangliang Chen, Georgia Institute of Technology

Liangliang Chen received the Ph.D. degree from the School of Electrical and Computer Engineering at the Georgia Institute of Technology, where he is currently a Postdoctoral Fellow. He received the B.B.A. degree in Business Administration, the B.S. degree in Automation, and the M.Eng. degree in Control Engineering from Harbin Institute of Technology. His research interests include machine learning theory and applications.

Dr. Jacqueline Rohde, Georgia Institute of Technology

Jacqueline (Jacki) Rohde is the Assessment Coordinator in the School of Electrical and Computer Engineering at the Georgia Institute of Technology. Her interests are in sociocultural norms in engineering and the professional development of engineering students.

Weiyu Sun, Georgia Institute of Technology

I'm currently the Ph.D. student at Georgia Institute of Technology majoring in Electrical and Computer Engineering. I earned BE and ME degrees at Nanjing University in China. My research interests/fields include Trustworthy AI, AI4Science, and bio-signal processing.

Dr. Ying Zhang, Georgia Institute of Technology

Dr. Ying Zhang is a Professor and Senior Associate Chair in the School of Electrical and Computer Engineering at the Georgia Institute of Technology. She directs the Sensors and Intelligent Systems Laboratory. Her research focuses on systems-level, interdisciplinary problems across multiple engineering disciplines, including AI-enabled personalized engineering education.

Leveraging RAG-Enhanced Large Language Models for Engineering Education: Design and Evaluation of an AI Tutor

Abstract

The increasing adoption of online and hybrid formats in engineering education has underscored the need for scalable, personalized student support. Although large language models (LLMs) offer a potential solution, they often fall short of engineering pedagogical standards due to hallucinations and misalignment with core instructional principles. In this paper, we present a domain-specific AI tutor that integrates Retrieval-Augmented Generation (RAG) to enhance factual accuracy and provide adaptive instructional scaffolding. The AI tutor supports three learning modes: feedback on student solutions, problem-specific guided support grounded in Self-Regulated Learning (SRL) principles, and open-ended question answering. We deployed the tutor in two undergraduate engineering courses at a large public research-intensive institution in the southeastern United States during the Fall 2025 semester. Using interaction data from 116 students, we evaluated both the quality of interactions and the relationship between students' usage patterns and their learning outcomes. Results demonstrated strong instructional stability, with the tutor achieving over 90% pass rates across key quality metrics. Additionaly, student usage patterns revealed a significant positive association between sustained interaction and improved academic performance. These findings suggest that domain-specific AI tutors, when grounded in rigorous scaffolding principles, can serve as scalable pedagogical interventions that not only answer questions but also potentially shape learning trajectories.

Introduction

The rapid advancement of large language models (LLMs) [1–5] has sparked a technological revolution across diverse sectors, including healthcare [6, 7], legal analysis [8, 9], and software development [10, 11]. Beyond their initial capabilities in text generation, modern LLMs have demonstrated sophisticated reasoning abilities [12], enabling them to process vast amounts of information and perform complex tasks [13, 14].

Educators can leverage LLMs to enhance personalized instruction and facilitate adaptive tutoring [15, 16]. The integration of these models has shown particular success in disciplines that rely on structured logic and text-based representations, such as computer science [17] and mathematics [18]. In programming education, for instance, LLMs effectively function as "pair programmers", assisting beginners with debugging and explaining algorithmic logic [19, 20]. Similarly, in mathematics, LLMs have demonstrated proficiency in solving quantitative problems and guiding students through algebraic derivations [21].

However, applying LLMs to engineering education presents unique challenges compared to pure mathematics or software development. Engineering pedagogy requires not only logical reasoning but also a deep conceptual understanding grounded in physical laws and real-world constraints [22, 23]. Moreover, solving engineering problems often involves multi-step reasoning that integrates heterogeneous information and demands the interpretation of domain-specific visual representations, such as circuit diagrams, an area that remains particularly challenging for text-centric models [24–26].

Another challenge is that generic LLMs are prone to the "Stochastic Parrots" phenomenon [27, 28], producing content that appears plausible but is factually incorrect. In engineering, where precision is critical, such hallucinations can be especially harmful; errors may propagate through multi-step problem-solving processes, potentially misleading students and impairing their conceptual understanding [29, 30].

Finally, there is a fundamental misalignment between the optimization objectives of LLMs and the goals of education. Standard LLMs are typically fine-tuned for "helpfulness", which is often interpreted as providing direct answers to user queries [31]. While this may be efficient for quickly delivering answers, it undermines the learning process by bypassing the cognitive effort essential for skill acquisition [32]. Effective tutoring, by contrast, involves guiding students through the problem-solving process using hints, feedback, and scaffolding rather than simply presenting final solutions. Consequently, current literature lacks robust methodologies for constraining LLMs to prioritize pedagogically sound tutoring strategies over direct information delivery, particularly in the context of engineering education.

Research Goals This study proposes and evaluates an AI tutor specifically designed for engineering education, integrating Retrieval-Augmented Generation (RAG) to ensure factual grounding and incorporating principles of Self-Regulated Learning (SRL) to guide student interaction. Subsequently, we deploy this AI tutor to investigate the following research questions:

1. **Instructional Effectiveness:** Can an RAG-enhanced AI tutor provide valid scaffolding and guidance that aligns with the goals of engineering problem-solving?
2. **System Robustness:** What specific failure modes arise when LLMs are constrained by RAG structures and course-specific documents (e.g., issues of accuracy and relevance)?
3. **Student Engagement:** How do engineering students engage with the designed AI tutor?
4. **User Performance:** Do students who use the AI tutor demonstrate better learning outcomes, such as final exam scores and course grades, compared to those who do not? How are these outcomes related to the frequency of AI tutor use?

Literature Review

Traditional intelligent tutoring systems (ITS) were predominantly rule-based, offering structured and highly reliable instruction [33, 34]. While effective, these systems required substantial

manual effort to incorporate domain rules and pedagogical strategies, thereby limiting their scalability and flexibility across different domains. The emergence of transformer architectures [35] and transformer-based language models [36, 37] has fundamentally reshaped the field of tutoring systems [38–40]. Instead of encoding explicit rules, these models learn patterns from large-scale text corpora, enabling them to generate appropriate responses to open-ended queries. The release of powerful large language models (LLMs), such as GPT-4 [2], marked a significant milestone, demonstrating human-level reasoning on various professional benchmarks in recent years. This flexibility makes them particularly attractive for educational applications, where student questions tend to be highly diverse and unpredictable.

However, the integration of LLM-based tutoring systems remains imbalanced, with engineering disciplines receiving substantially less attention than programming and mathematics education [24]. Programming education has seen extensive LLM integration, from code completion tools to automated feedback systems [41, 42]. Mathematics tutoring has similarly benefited from systems that can parse and solve symbolic problems [43]. In electrical engineering, recent benchmarks such as CIRCUIT [44] and end-to-end solvers [39] have begun to address circuit topology, connectivity, and polarity relationships. Our work contributes to this effort by designing and evaluating an AI tutor with integrated RAG to ground the model's reasoning in course-specific constraints. This approach extends the capabilities of LLMs to support the rigorous, multi-step logical reasoning required in engineering courses such as circuit analysis and signal processing.

Self-Regulated Learning Effective educational technology must be grounded in learning theory. SRL theory provides an important framework for the AI tutoring system, emphasizing that learners should control their learning pace, set goals, and reflect on their understanding [45, 46]. However, the current LLMs often fail to support SRL, as they tend to provide direct answers rather than scaffolding the learning process [32]. To address these issues, recent works have attempted to bridge this gap by designing "Socratic" agents that guide students through multi-turn dialogues [47], or by using game-inspired mechanisms to improve self-regulation skills [48]. These system designs aim to provide on-demand scaffolding, offering temporary support structures that help students progress while maintaining their cognitive engagement. In this work, we take inspiration from this theory and integrate SRL principles into the system's interaction design using a structured scaffolding module with prompt engineering. Rather than acting as a passive answer-providing tool, our AI tutor is designed to avoid giving direct solutions. Instead, it offers adaptive hints and conceptual guidance that encourage students to actively reflect, set goals, and self-correct as they work through problems.

Retrieval-Augmented Generation (RAG) in Education The technical challenge lies in enabling LLMs to provide appropriate scaffolding while simultaneously ensuring factual accuracy. RAG [49] offers a promising approach by grounding model responses in external knowledge sources [50, 51]. Instead of relying solely on parametric knowledge acquired during pretraining, RAG systems retrieve relevant documents and guide generation based on the retrieved context.

In educational contexts, RAG allows AI tutors to reference specific course materials and textbooks, ensuring alignment with curriculum content [52]. However, RAG systems are not without failure points; valid retrieval does not guarantee valid reasoning, and integrating retrieved snippets

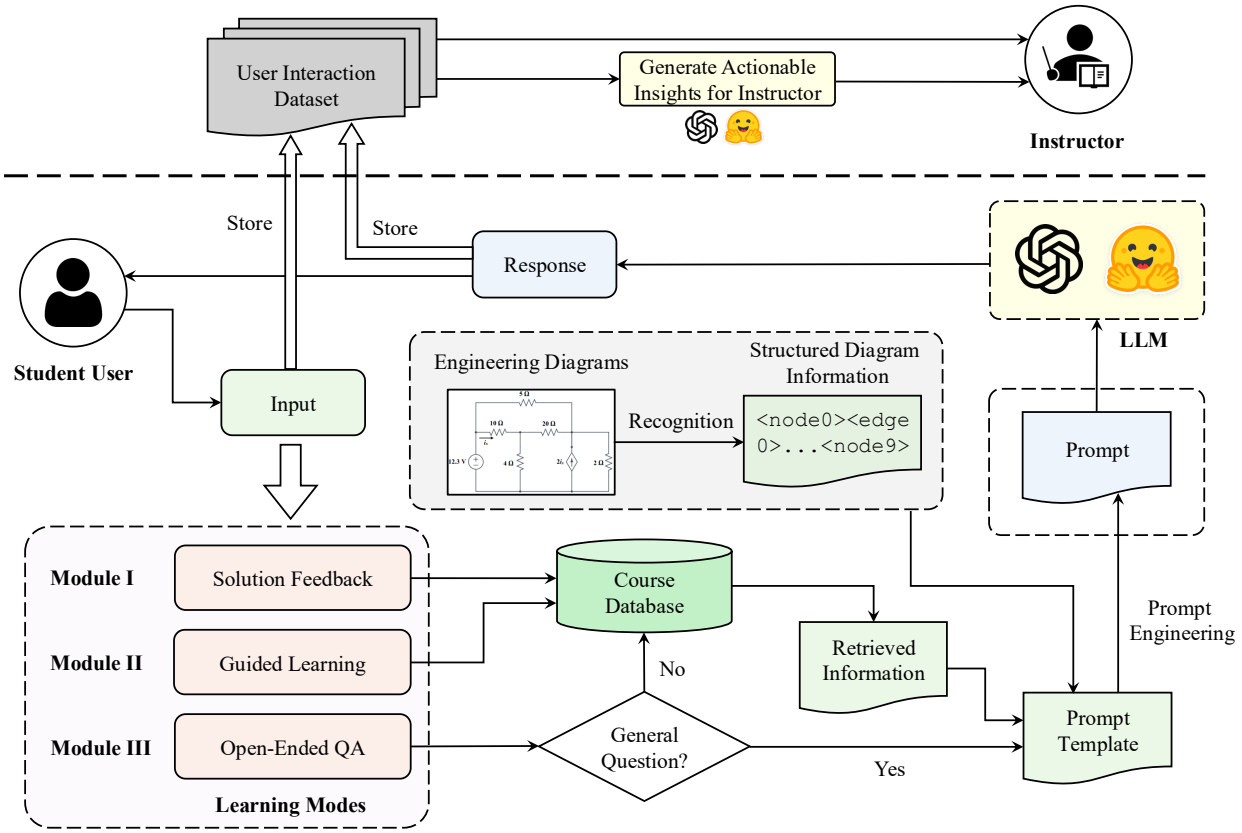


Figure 1: Structure of the RAG-enhanced AI tutor for engineering education

into a coherent pedagogical narrative remains difficult [53]. For instance, standard RAG pipelines may struggle with multi-hop reasoning required for complex engineering problems where information is distributed across different documents or modalities [54].

Our work builds on these advancements by implementing a specialized RAG architecture that indexes verified course materials to reduce the hallucination risks common in standard LLMs. We combine this factual grounding with a pedagogical control layer to explore how a knowledge-enhanced AI tutor can be effectively used to support the precise, curriculum-aligned reasoning needed in university-level engineering education.

AI Tutor Design and Evaluation

AI Tutor Design

The AI tutor pipeline is illustrated in Figure 1. When a student user submits an input, the system retrieves relevant documents based on the selected learning mode and integrates them into a prompt template. Problem-related engineering diagrams are pre-converted into textual descriptions and appended to the LLM prompt to support the different learning modes. With this context-rich prompt, the LLM generates a response, which is delivered to the student and stored in a user

interaction database for further analysis. These analyses can be used to generate actionable insights for instructors, enabling timely instructional support in the classroom.

The AI tutor is designed to support three distinct learning modes: (i) feedback on students' solutions, (ii) problem-specific guided support grounded in self-regulated learning principles, and (iii) open-ended question answering. These three modules address different educational needs, and their core functions are introduced below.

Module I – Feedback on Students' Solutions This module assists students in reviewing their work by analyzing submitted solutions and providing corresponding feedback. Students can either upload their solutions as `.txt` files or type them directly into the interface. The system performs best with LaTeX-formatted submissions, which facilitate clearer processing of mathematical notation. After analyzing a submission, the tutor generates feedback across four dimensions: (i) final answer accuracy and arithmetic correctness, (ii) solution completeness, (iii) methodology appropriateness, and (iv) unit consistency [55]. Each feedback category can be expanded or collapsed via interactive buttons, allowing students to focus on specific areas needing improvement.

Module II – Problem-Specific Guided Support Grounded in Self-Regulated Learning Principles The second module is designed for deeper engagement through interactive dialogs that are linked to specific homework problems and support self-regulated learning. This module employs a scaffolding approach in which the tutor poses guiding questions and offers conceptual cues, rather than providing direct answers. For example, if a student asks, "How do I find the current in this problem?", the tutor might respond with the hint, "The current is the time derivative of charge." The guidance mode is most effective when students engage with the AI tutor as a learning tool, not merely a means to obtain quick answers. Gaining the most benefit from the system requires that student phrase their questions with sufficient detail. As with the solution feedback in Module I, students must use clear formatting in their questions and responses throughout the interaction. To address these usage considerations, we provide students with a tutorial video and analyze the real-world interaction dataset, which contains questions of varying clarity.

Module III – Open-Ended Question Answering The third module enables students to ask questions about general course concepts as "sidebar" conversations. This mode is especially useful for obtaining *quick* clarification on fundamental concepts. For instance, asking "How do I calculate current from total charge?" prompts the system to provide domain-specific guidance that connects theoretical principles with practical problem-solving strategies.

Analysis Provided for Instructor To support instructional decision-making, we leverage LLM-based analysis of student-AI tutor interaction data to identify common knowledge gaps and learning difficulties. The system generates insights that help instructors understand class-wide misconceptions, frequently confused concepts, and areas requiring additional pedagogical attention. This feedback loop enables instructors to adapt their teaching strategies in a timely and data-informed manner.

AI Tutor Deployment

We deployed the AI tutor in the following two courses at a large, public, research-intensive institution in the Southeast United State during the Fall 2025 semester. The project was approved by the institution's Institutional Review Board. Participating students provided informed consent, authorizing the collection and use of their AI tutor interaction records, homework submissions, and grading data for research and publication purposes. The courses where the tutor was implemented are as follows:

- *Circuit Analysis:* This course covers fundamental techniques such as Kirchhoff's circuit laws, operational amplifiers, AC circuits, and frequency response. It includes 11 homework assignments (contain 128 problems totally) and 3 exams over the semester. A total of 72 students from two sections (out of 113 enrolled students), primarily freshmen and sophomores, voluntarily participated in the study from the beginning of the semester. Student performance was measured by the final score, expressed as a percentage out of 100%.

- *Signal Processing:* This course focuses on discrete-time signal processing, filtering, transforms, and computer-based applications. It includes 12 homework assignments (44 problems in total) and 3 exams over the semester. A total of 44 students from one section (out of 166 enrolled students), primarily freshmen and sophomores, voluntarily participated in the study from the beginning of the semester. Student performance was measured using a four-point grading scale. Fall 2025 marked the first semester the AI tutor was deployed in this course. Due to limited project team capacity, tutor enrollment was restricted to specific sections, resulting in fewer student participants in this course than *Circuit Analysis*.

The AI tutor is deployed on the Microsoft Azure platform[1], and students access it via a generated URL following deployment. Figure 2 displays the main page of the AI tutor user interface. The student-tutor interaction data were stored in an SQL database on Microsoft Azure in real time. These records include Chat ID, homework and problem IDs, student IDs, the interaction times, the selected learning mode, student inputs, and the AI tutor's responses. This data enables analysis of students' usage patterns of the AI tutor, the quality of the tutor's responses, and the relationship between students' performance and their frequency of AI tutor usage. Figure 3 presents the evaluation workflow of the AI tutor and its impact on student performance.

Evaluation Metrics

This section first outlines the evaluation metrics used to assess the quality of the AI tutor. The evaluation focuses on the characteristics of AI tutor responses, examined through manual analysis of student-tutor interaction data. In addition, we examine differences in learning outcomes across non-users and groups of AI tutor users with varying levels of usage consistency and practice intensity.

AI Tutor Response Quality To ensure an accurate evaluation of the system, we manually assessed the stored interactions between the AI tutor and students using eight criteria: *Factual Accuracy*, *Judgmental Accuracy*, *Relevance*, *Self-Awareness*, *Response Targeting*, and *Coherence*. For

[1]See `https://portal.azure.com`.

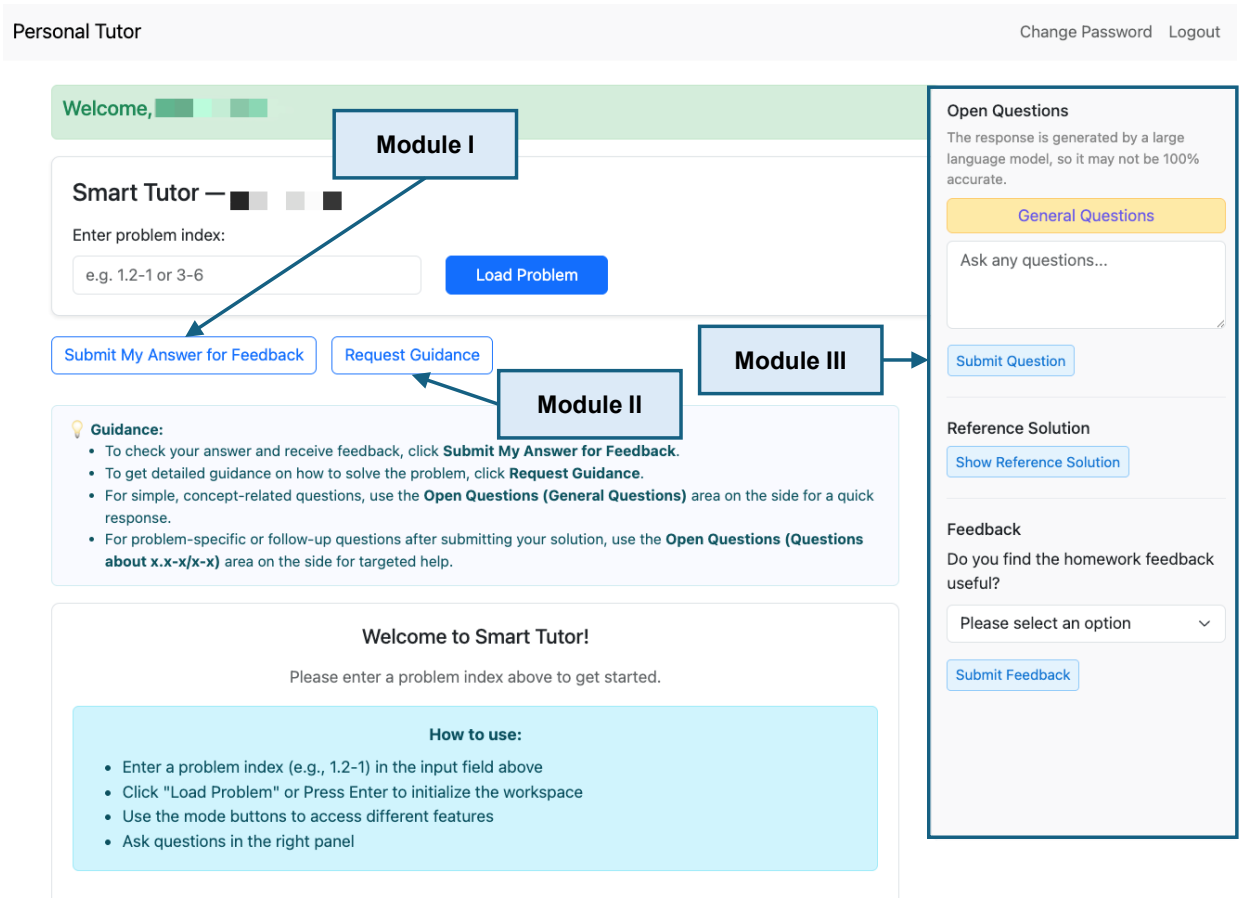


Figure 2: Main page of the AI tutor user interface. The screens displayed after loading a problem and clicking "Submit My Answer for Feedback" (Module I) and "Request Guidance" (Module II) are shown in Figures 9 and 10 in Appendix A.

each criterion, a score of +1 is assigned if it is satisfied; otherwise, the score is 0, resulting in a pass/fail scoring format. The metrics are described in detail below.

- **Factual Accuracy:** Assesses whether the AI tutor's responses are consistent with the problem description and domain-specific knowledge (e.g., circuit principles), and whether they accurately reflect the original problem information.
- **Judgmental Accuracy:** Evaluates the AI tutor's ability to correctly assess the accuracy of student answers, particularly numerical responses. When a student's answer is partially correct, this metric considers whether the tutor clearly identifies which parts are correct and which are incorrect, rather than simply labeling the entire response as "incorrect".
- **Relevance:** Measures whether the AI tutor's response directly addresses the student's specific question. This metric is particularly sensitive to cases where the student asks question A, but the tutor provides a response to question B.
- **Self-Awareness:** Evaluates whether the AI tutor avoids suggesting actions beyond its capabilities, such as requesting that students draw diagrams and send them for feedback.

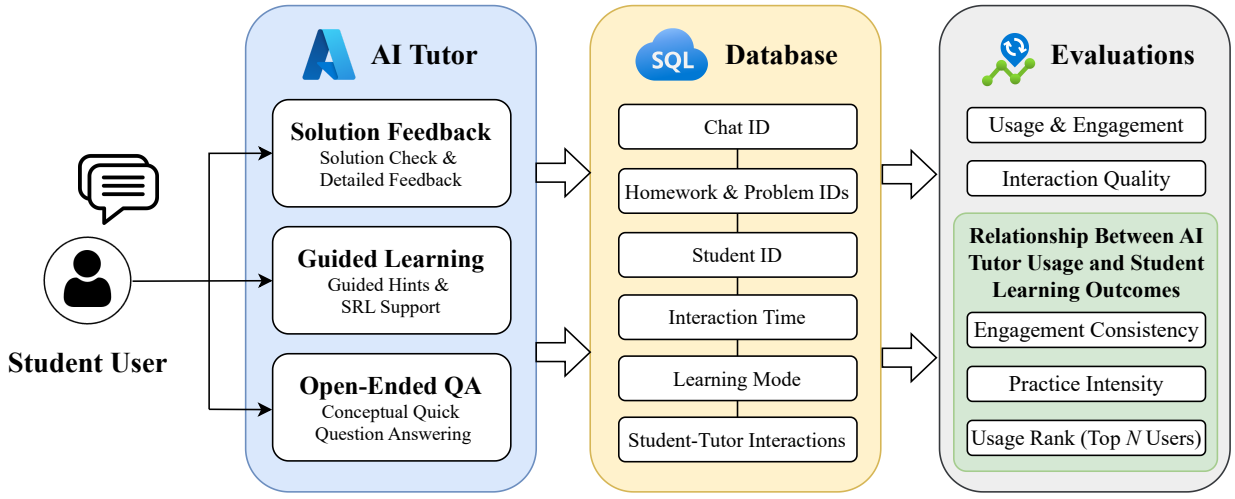


Figure 3: Evaluation workflow of the AI tutor and its relationship to student learning outcomes through usage patterns

- **Response Targeting:** Assesses whether the tutor adapts its responses to the student's current level of understanding, rather than following a fixed response format. For instance, it evaluates whether the tutor avoids repeating concepts the student already knows or providing overly lengthy explanations, especially when the student shows signs of impatience.
- **Coherence:** Examines whether the AI tutor maintains contextual coherence across extended interactions (e.g., beyond five conversational turns), rather than in short, isolated exchanges.

Student Learning Outcome We further employ Glass's Δ [56] to quantify the magnitude of the difference in learning outcomes between the experimental group (with the AI tutor) and the control group (without the AI tutor).[2] Unlike statistical significance measures (e.g., p-values), which are highly dependent on sample size, Glass's Δ offers a standardized measure of the difference between two groups. It is calculated as the difference between the means of the two groups divided by the standard deviation of the control group, as shown in Eq. (1):

$$\Delta = \frac{\bar{x}_{\text{exp}} - \bar{x}_{\text{ctrl}}}{s_{\text{ctrl}}} \tag{1}$$

where $\bar{x}_{\text{exp}}$ and $\bar{x}_{\text{ctrl}}$ represent the mean post-test scores of the experimental and control groups, respectively, and s_{ctrl} is the standard deviation of the control group.

Because Glass's Δ uses the control-group standard deviation as its standardizer, which is a form of standardized mean difference whose magnitude can be interpreted against conventional benchmarks [57]. Under this convention, values of approximately 0.2, 0.5, and 0.8 are commonly described as small, medium, and large, respectively [58]. In the present study, values in

[2]As shown in Figure 6, the score variances between AI tutor users and non-users differ. Therefore, Glass's Δ is used instead of Cohen's d, which assumes equal group variances.

the medium-to-large range suggest comparatively substantial differences in observed learning outcomes between students who engaged with the AI tutor and those who received only standard classroom instruction. It should be noted that Glass's Δ is reported here solely to characterize between-group differences and carries no causal interpretation. Establishing whether AI tutor usage causally affects academic performance would require experimental or quasi-experimental designs in future work.

Results

In the Fall 2025 semester, 72 out of 113 students used the AI tutor in *Circuit Analysis*, while 44 out of 166 students used it in *Signal Processing*.[3] We collected data on a total of 27,242 interactions from students in *Circuit Analysis* (15,809 interactions) and *Signal Processing* (11,433 interactions). An "interaction" is defined as a question-answer dyad, where a student poses a question to the tutor and receives a response. These interactions were linked to a ChatID, which enabled us to capture the back-and-forth conversations with the tutor within a relatively short time period.

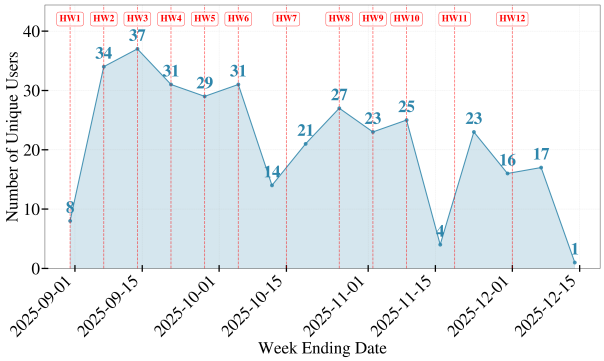
For *Circuit Analysis*, the 72 users averaged 5.23 unique homework assignments per user in terms of AI tutor usage, 34.31 unique problems per user, and a usage span of 52.5 days. For *Signal Processing*, the 44 users averaged 6.00 unique homework assignments per user, 17.94 unique problems per user, and a usage span of 48.9 days.

In this section, after examining the detailed usage and engagement patterns of both courses, we first assess interaction quality by sampling 260 recordings from the 27,242 interactions for manual analysis. The sampling was couducted based on ChatID, meaning that all interactions between the tutor and student were pulled while working on a particular homework problem at a particular time. This process allowed us to see a full context of the interaction. The sampling represents less than 1% of the data, which was constrained by the time investment of the evaluator to review the homework problem, read through the full conversation, and evaluate each interaction. Future work will discuss broader evaluation efforts. The samples led to 173 interactions from *Circuit Analysis*, and 87 from *Signal Processing* being evaluated. We then investigate how usage patterns influence student performance by analyzing their final grades.

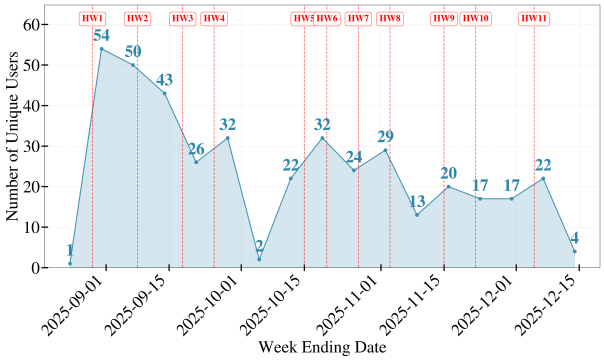
Usage and Engagement The engagement data presented in Figure 4 reveal distinct behavioral patterns across both courses. The weekly active user trends in Figures 4(a) and 4(b) illustrate a classic "novelty effect". Both courses see rapid adoption in the first three weeks as students experiment with the new tool, and a gradual decrease of users is observed post-HW3. However, a sustained user base in both courses indicates a core group of students who have successfully integrated the tool into their regular workflow.

The usage trends in Figures 4(c) and 4(e) indicate that student interaction with the AI tutor in *Signal Processing* was largely utilitarian. Engagement followed a pulsatile pattern: usage spiked sharply in the days leading up to homework deadlines and dropped to near zero immediately af-

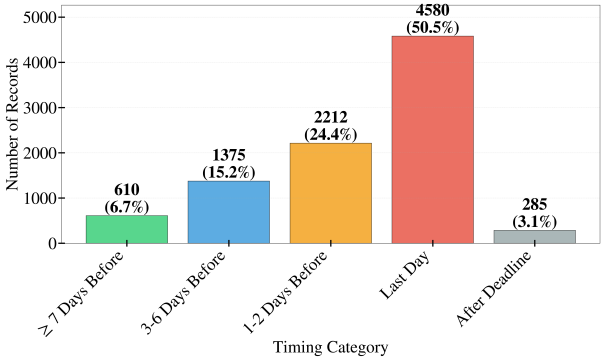
[3]For the *Signal Processing* course, we allocated 50 slots for AI tutor users. Only the first 50 students who volunteered were permitted to register an account.



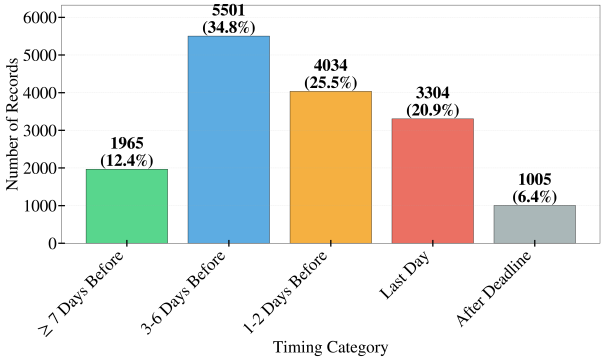
(a) Weekly active users in *Signal Processing*



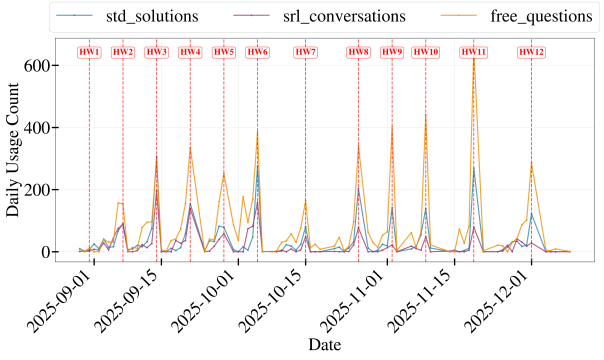
(b) Weekly active users in *Circuit Analysis*



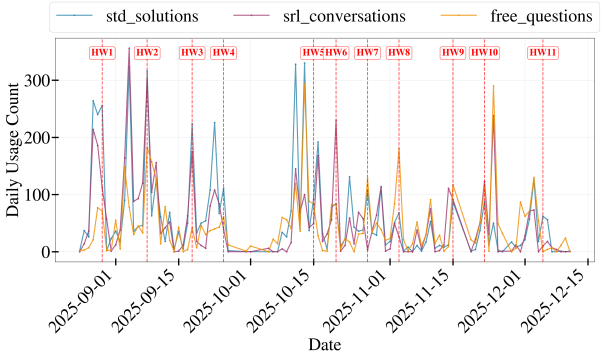
(c) User engagement patterns in *Signal Processing*



(d) User engagement patterns in *Circuit Analysis*



(e) Daily usage trends by interaction type in *Signal Processing*



(f) Daily usage trends by interaction type in *Circuit Analysis*

Figure 4: Usage patterns in *Signal Processing* (44 unique users) and *Circuit Analysis* (72 unique users). (a–b) Weekly active users, illustrating retention trends; (c–d) Submission timing distributions relative to assignment deadlines; (e–f) Daily usage volumes by interaction type: *std_solution* (Module I – Solution Feedback), *srl_conversations* (Module II – Guided Learning), and *free_questions* (Module III – Open-Ended Question Asking).

terward. More than 70 percent of all queries occurred within the final two days before assignment submission. The trends in *Circuit Analysis* (Figures 4(d) and 4(f)) are less pronounced because the two sections had different homework deadlines[4]. Even so, the decline in usage between HW4 and HW5 suggests a similar deadline-driven pattern. At the same time, Figures 4(e) and 4(f) add nuance. Each colored line represents one of the three module types, and students engaged with all types across the semester. This distribution suggests that students were not only seeking final answers but also using the tutor to support deeper conceptual learning.

Assessment Analysis I: Interaction Quality The detailed distribution of those sample interactions is shown in Table 1. The distribution of pass/fail results across six metrics is illustrated in Figure 5. Our AI tutor demonstrates high proficiency in *Self-Awareness* (98.46%), *Relevance* (98.08%), and *Coherence* (97.69%), indicating strong conversational stability.

Table 1: Distribution of student-AI interactions across different learning modes by course

Course	Solution Feedback	Guided Learning	Open-Ended Question Asking
Circuit Analysis	67	76	30
Signal Processing	15	37	35

However, the analysis also highlights three primary areas for improvement: *Guidance Targeting*, *Factual Accuracy*, and *Judgmental Accuracy*, with failure rates of 7.31%, 5.77%, and 5.38%, respectively. *Guidance Targeting* is the most significant issue, suggesting that the AI tutor occasionally struggles to adapt to the student's dynamic level of understanding, leading to either repetitive questioning or overly complex explanations. Additionally, detailed analysis indicates that the primary causes of failures in *Factual Accuracy* and *Judgmental Accuracy* stem from the inherent limitations of LLMs in performing numerical verification and domain-specific fact-checking without the support of external tools.

Assessment Analysis II: AI Tutor User Performance Usage logs revealed diverse engagement patterns, ranging from brief trials to consistent use throughout the semester. To investigate the relationship between the consistency and intensity of AI tutor interaction and student performance, we categorized students into four groups: Non-users, *Occasional* users, *Moderate* users, and *Frequent* users. Because the two courses differed in the number of assignments and problems, we applied percentile-based thresholds to facilitate cross-course comparison.

We examined three aspects of student engagement: (1) Engagement Consistency, defined as the number of homework assignments for which students used the tool; (2) Practice Intensity, measured by the number of problems students attempted; and (3) High Usage Characteristics, referring to performance patterns among the most active users, as identified by cumulative usage ranks. This multidimensional approach provides a clearer understanding of how both the frequency and depth of AI tutor use relate to student performance.

[4]In Figure 4, the later deadline is used for each assignment.

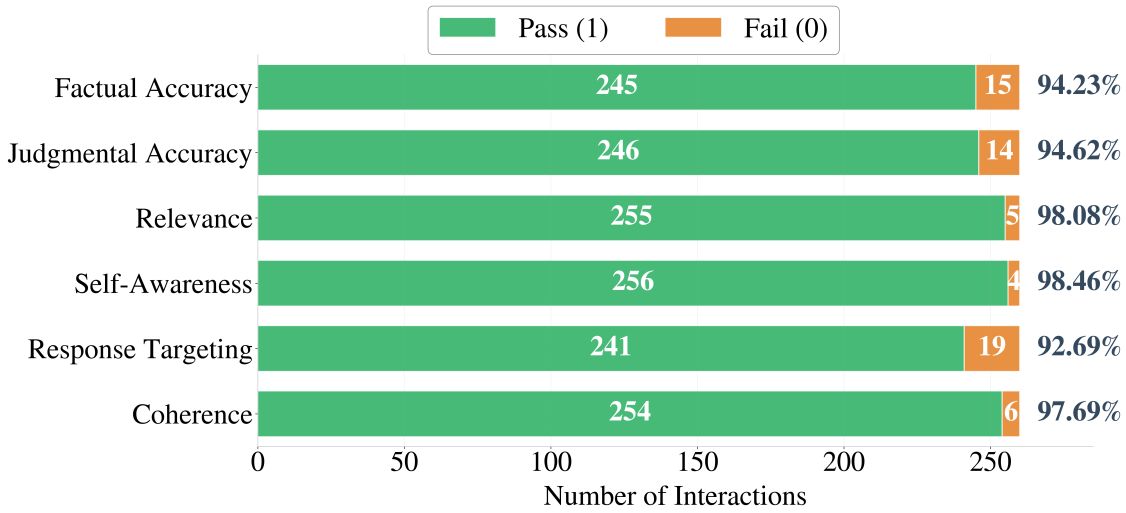


Figure 5: The distribution of metrics for quality evaluation

Engagement Consistency (Homework Count)

We grouped students based on the number of homework assignments for which they used the AI tutor. The thresholds were set at approximately 20% and 75% of total assignments (there were 11 homework assignments for *Circuit Analysis* and 12 for *Signal Processing*): *Occasional* users ($\leq$ 20% coverage, 1 or 2 HWs), *Moderate* users (20%–75% coverage, 3–8 HWs), and *Frequent* users ($>$ 75% coverage, 9+ HWs).

Table 2 and Figure 6 illustrate the relationship between the consistency of AI tutor usage and student performance. Table 2 presents the sample size (N) for each group, the standard deviation (SD) of scores, the difference (Diff) between each group's average score and that of non-users, and the corresponding values of Glass's Δ. Figure 6 presents box plots showing the distribution of final scores or grades across the different usage groups. In each box plot, the box represents the interquartile range (IQR), spanning from the 25th to the 75th percentile, with the horizontal green dashed line inside the box indicating the mean value. The whiskers extend to the minimum and maximum values within 1.5 times the IQR, and circles denote outliers beyond this range. (This definition applies to all subsequent tables and figures of a similar format.)

In *Circuit Analysis*, grades exhibited a monotonic association with the consistency of AI tutor usage. While *Occasional* users showed a negligible difference relative to non-users (Glass's $\Delta = 0.09$), larger standardized differences were observed for *Moderate* ($\Delta = 0.18$) and *Frequent* ($\Delta = 0.32$) users. These patterns indicate that higher levels of sustained usage are associated with higher academic performance.

However, *Signal Processing* exhibited a different pattern. *Occasional* users had lower average grades (mean = 2.67) than non-users (mean = 2.94), indicating a negative association between limited AI tutor usage and academic performance in this group. In contrast, positive standardized differences were observed for *Moderate* (Glass's $\Delta = 0.24$) and *Frequent* ($\Delta = 0.33$) users. Taken together, these results suggest that performance in *Signal Processing* varies systematically with the

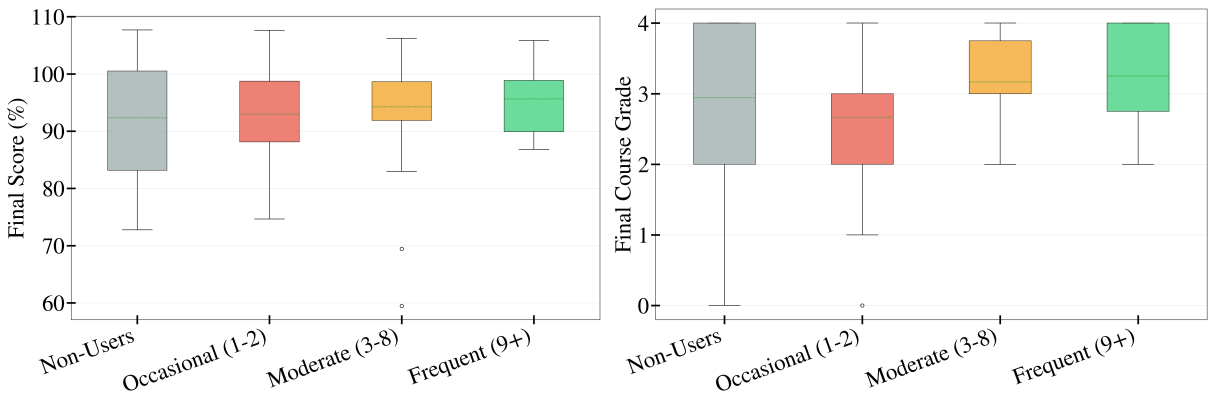


Figure 6: Final Grades by Engagement Consistency (Based on Homework Count). Left: *Circuit Analysis* (Final Score as a Percentage); Right: *Signal Processing* (Four-Point Grading Scale).

Table 2: Relationship between engagement consistency (Homework Count) and final grades

	Circuit Analysis					***Signal Processing***				
Usage Group	**Final %**	**SD**	**N**	**Diff**	**Δ**	**Grade**	**SD**	**N**	**Diff**	**Δ**
Non-Users	92.31	10.39	41	—	—	2.94	0.93	122	—	—
Occasional	93.22	8.44	20	+0.91	0.09	2.67	1.05	18	–0.27	–0.30
Moderate	94.23	8.93	36	+1.92	0.18	3.17	0.69	6	+0.23	0.24
Frequent	**95.63**	5.83	16	**+3.32**	0.32	**3.25**	0.83	20	**+0.31**	0.33

level of AI tutor engagement, with more consistent usage aligning with higher observed outcomes.

Practice Intensity (Problem Count)

We also examined the relationship between the number of problems attempted using the AI tutor and student outcomes. Usage thresholds were defined at approximately 10% and 45% of the total available problems: *Occasional* users attempted $\leq$10%, *Moderate* users attempted 10%–45%, and *Frequent* users attempted more than 45%. Whereas the previous analysis characterized engagement in terms of the number of homework assignments involving tutor use, the present analysis focuses on the proportion of problems within each assignment that were attempted with the tutor.

The results in Table 3 indicate that practice volume is differently associated with performance across the two courses. In *Circuit Analysis*, students who attempted the largest number of problems (>59 problems) exhibited the highest observed grades (96.78%, Glass's $\Delta = 0.42$). Figure 7 further shows that this high-intensity group displayed a more concentrated score distribution, characterized by lower variability in performance. In *Signal Processing*, a slightly different association pattern was observed. Students in the *Moderate* group (5–20 problems) exhibited the highest average performance (mean grade = 3.38, Glass's $\Delta = 0.47$), marginally exceeding that of the *Frequent* group (mean grade = 3.25). In contrast, both groups demonstrated substantially higher performance than *Occasional* users (Glass's $\Delta = -0.48$). Collectively, these results de-

scribe an association between problem attempt volume and performance in this course that is not strictly monotonic, although a general trend is evident, with higher observed outcomes among students at moderate levels of engagement.

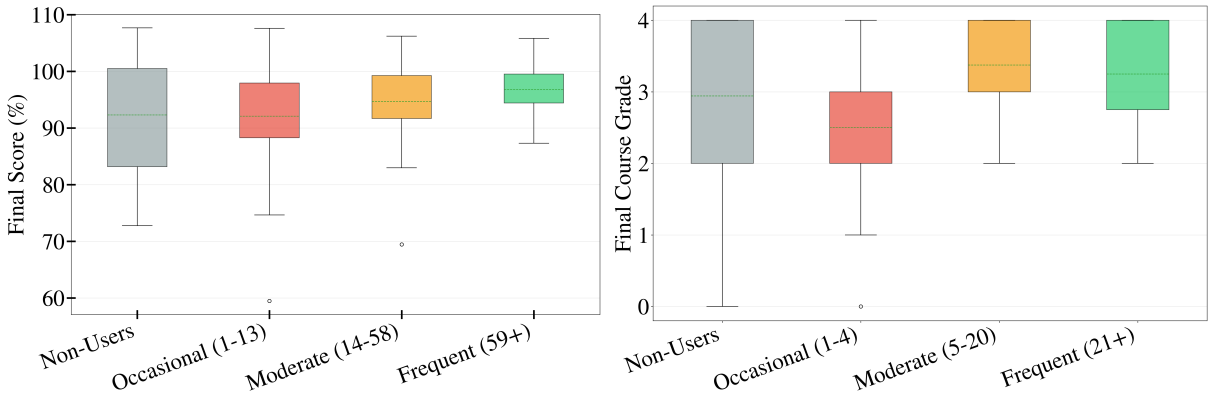


Figure 7: Final Grades by Practice Intensity (Based on Problem Count). Left: *Circuit Analysis* (Final Score as a Percentage); Right: *Signal Processing* (Four-Point Grading Scale). In *Circuit Analysis*, higher practice intensity groups exhibit a more concentrated distribution of final grades. In *Signal Processing*, differences in average performance are observed across practice intensity groups, with the highest mean values appearing in the *Moderate* group.

Table 3: Relationship between Practice Intensity (Problem Count) and Final Grades

	Circuit Analysis					***Signal Processing***				
Usage Group	**Final %**	**SD**	**N**	**Diff**	**Δ**	**Grade**	**SD**	**N**	**Diff**	**Δ**
Non-Users	92.31	10.39	41	—	—	2.94	0.93	122	—	—
Occasional	92.23	9.87	26	–0.09	–0.01	2.50	1.00	16	–0.44	–0.48
Moderate	94.68	7.60	30	+2.37	0.23	**3.38**	0.70	8	**+0.44**	**0.47**
Frequent	**96.78**	5.09	16	**+4.47**	**0.42**	3.25	0.83	20	+0.31	0.33

Usage Ranks and Performance

To characterize the outcomes associated with the highest levels of platform interaction, we analyzed performance based on cumulative usage ranks. Unlike the previous groupings defined by fixed thresholds, this analysis focuses on the most active student cohorts by absolute volume: the *Top 30*, *Top 20*, and *Top 10* users ranked by the number of logged interactions. This approach allows us to observe how performance characteristics shift as the definition of the "high-usage" group becomes increasingly selective. Table 4 provides the corresponding statistical summary, while Figure 8 illustrates the distributional patterns across usage cohorts.

Two findings emerge from this analysis. First, average performance peaks among the most selective cohorts. In *Circuit Analysis*, the Top 10 users achieved scores 4.51 percentage points higher than non-users ($\Delta = 0.43$), while in *Signal Processing*, they scored 0.56 grade points higher ($\Delta = 0.60$), approaching a medium-to-large effect size. These values are comparable to or exceed those observed in the *Frequent* usage categories from the previous analyzes, confirming that

students at the extreme end of the engagement spectrum consistently demonstrate strong academic outcomes.

What's more, Figure 8 reveals a substantial reduction in performance variance among highly active users. In both courses, the non-user group displays wide interquartile ranges with significant low-end outliers—scores approaching 20% in *Circuit Analysis* and a grade of 0.0 in *Signal Processing*. In contrast, the top user cohorts (Top 30 through Top 10) exhibit tightly clustered distributions, effectively eliminating the lower performance tail. This variance compression suggests that intensive platform engagement is associated not only with higher average achievement but also with more stable performance outcomes, potentially serving as a protective factor against academic failure.

Table 4: Usage Rank and Performance: Final Grades of Top Usage Users

	Circuit Analysis			***Signal Processing***		
Cohort	**Final %**	**Diff**	**Δ**	**Grade**	**Diff**	**Δ**
Non-Users	92.31	—	—	2.94	—	—
Top 30 Users	96.09	+3.77	0.36	3.20	+0.26	0.28
Top 20 Users	95.83	+3.52	0.33	3.25	+0.31	0.33
Top 10 Users	**96.83**	**+4.51**	**0.43**	**3.50**	**+0.56**	**0.60**

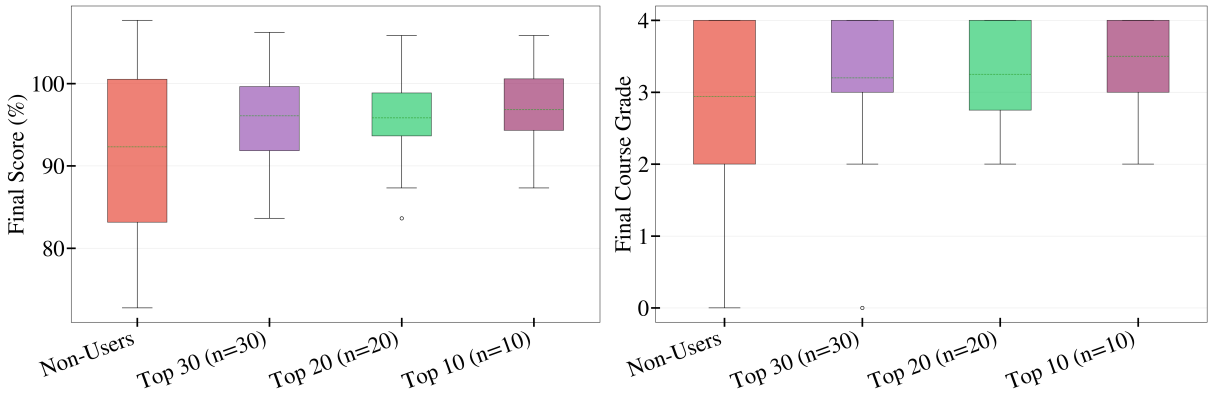


Figure 8: Final Grades by Usage Rank. Left: *Circuit Analysis* (Final Score as a Percentage); Right: *Signal Processing* (Four-Point Grading Scale). Note the compression of the interquartile range (box height) among high-usage groups, indicating increased performance consistency and reduced variance.

Discussions

Reliability and the Scaffolding Trade-off The study demonstrates that while the RAG-enhanced AI tutor can effectively provide scaffolded learning guidance, achieving 100% precision in responses remains challenging. Although the system exhibited strong instructional stability, with high scores in *Self-Awareness* (98.46%) and *Coherence* (97.69%), failure rates in *Guidance Targeting* (7.31%) and *Factual Accuracy* (5.77%) highlight ongoing limitations. A detailed analysis

of issues related to *Judgment Accuracy* indicates that current LLMs, even when grounded by RAG, occasionally fail to rigorously verify numerical engineering solutions. These findings underscore the continued necessity of integrating a "human-in-the-loop" paradigm into university-level engineering AI tutors to ensure reliability and fairness in student support.

Engagement Density and Learning Outcomes Our analysis shows a positive relationship between AI tutor usage density and academic performance, although this relationship is non-linear. While the data suggests that higher interaction frequency correlates with improved learning outcomes, we identified two critical patterns characterizing this relationship. First, there exists a minimum engagement threshold: Occasional usage is not enough to activate the tool's scaffolding benefits. Second, beyond the association with higher average scores, increased usage intensity is accompanied by a significant reduction in performance variation. This suggests that sustained interaction is associated with more consistent academic achievement across the active cohort.

One notable observation is that mere access to the AI tutor is not associated with better performance. In both *Circuit Analysis* and *Signal Processing*, students who used the tool occasionally performed similarly to the non-users, and sometimes worse. Occasional users covered less than 20% of assignments or attempted only a few problems. In *Signal Processing*, these users actually had a lower grade (2.67 or 2.50) compared to non-users (2.94). For students surpassing this engagement threshold, we observed a clear positive gradient. In most cases, high engagement with the AI tutor is strongly linked to higher academic performance. However, we must interpret this correlation with caution. Students who use the AI tutor extensively may already have higher motivation, better self-regulation skills, or stronger learning habits. Thus, the AI tutor may function as a facilitator for students who are already engaged, rather than being the sole driver of improvement. Separating the tool's direct impact from pre-existing student characteristics would require randomized controlled trials in future work.

Beyond mean score differences, the distributional analysis (Figures 6–8) reveals a clear association between usage intensity and variance reduction. As usage intensity increased, the dispersion of grades, as measured by the interquartile range, narrowed substantially, indicating more consistent performance among frequent users. This contraction in variability suggests that sustained engagement with the AI tutor is associated with a more stable performance baseline. By providing on-demand instructional scaffolding, the AI tutor may help stabilize learning outcomes and mitigate the risk of severe underperformance. Notably, high-frequency user groups exhibited few, if any, low-end outliers, with virtually no failing grades observed. These findings have important implications for educational equity, as they indicate that consistent tool usage may help prevent extreme academic lag and elevate the lower bound of student achievement.

Limitations Our findings are subject to several limitations. First, the system's performance depends heavily on the quality and completeness of the indexed course materials; gaps in the retrieval corpus are directly correlated with an increased risk of hallucinations. Second, the system currently performs best with structured inputs, such as LaTeX-formatted equations, which may impose an additional workload on students. Third, the novelty effect observed during the first three weeks suggests that long-term retention strategies are needed to transform initial curiosity into sustained learning habits. Fourth, the study was limited to two electrical engineering courses at a single

institution, and generalizing these findings to other engineering disciplines will require further investigation.

Additionally, a key limitation concerns the interpretation of the observed association between AI tutor usage and academic performance. Although the analysis reveals a strong positive relationship, selection bias cannot be ruled out: students who are already high-performing or highly motivated may be more likely to adopt and consistently use the AI tutor. Future work will work to address this limitation, as described below.

Future Directions To address these limitations and enhance both the technical robustness and pedagogical utility of the system, we propose several directions for future work:

- **Enhancing System Accuracy and Reliability:** To reduce hallucination risks, future iterations should expand the retrieval corpus to include more comprehensive course materials, such as worked examples and common misconceptions. Additionally, the system will be augmented with external computational tools (such as a Python sandbox) to verify student calculations step-by-step against ground-truth solvers.

- **Reducing Barriers through Multimodal Input:** The current requirement for LaTeX formatted input may create accessibility challenges for some students. To address this, we are actively developing OCR capabilities that will allow students to upload images of handwritten work. This enhancement is expected to increase accessibility, particularly for students who are mathematically proficient but less familiar with LaTeX syntax.

- **Sustaining Engagement and Personalizing Guidance:** The novelty effect observed during the first three weeks suggests that initial curiosity alone does not lead to sustained learning benefits. To promote long-term engagement, we will implement a dynamic user state tracker that estimates each student's knowledge level and engagement patterns based on interaction history. This will enable two key improvements: (1) personalized strategies, such as timely reminders and progress tracking, to encourage consistent usage, and (2) adaptive guidance that delivers direct hints to students facing difficulties and higher-level conceptual support to advanced learners.

- **Validating Generalizability and Causal Effects:** To evaluate the broader applicability of our findings, the system should be deployed across additional engineering disciplines and institutions with diverse student populations. To work towards identifying possible causal effects, future work will leverage experimental or quasi-experimental designs that control for prior academic achievement.

- **Expanding Evaluation:** Manually evaluating interactions with human reasoning, while important, can be complemented by using artificial intelligence to evaluate at a larger scale (tens of thousands of responses) that is beyond human capabilities in a reasonable time frame .

Conclusions

This study describes the design and evaluation of a RAG-enhanced AI tutor for engineering education and examines how students engaged with it in authentic course settings. Interaction patterns were largely deadline-driven and stabilized after an initial novelty period. Across six qualitative evaluation metrics, the tutor demonstrated strong instructional reliability, with pass rates above 90 percent for factual accuracy, judgment, relevance, self-awareness, targeting, and coherence, although these is still room for improvement. In terms of student outcomes, occasional use was associated with minimal differences relative to non-use, whereas sustained engagement corresponded with higher observed performance and reduced score variability. These patterns suggest that consistent AI-supported scaffolding may help stabilize performance across a diverse student population.

For engineering educators, these findings provide evidence that retrieval-augmented AI tutors can offer scalable conceptual support while maintaining instructional consistency. For engineering education researchers, the results highlight important considerations for studying engagement intensity, utilitarianism, and interaction quality when evaluating AI-based interventions.

Future work will address both technical and methodological limitations. Technically, we will improve retrieval and incorporate computational verification to strengthen accuracy, while adding OCR-based multimodal input to reduce access barriers. Methodologically, controlled study designs and expanded automated evaluation will help establish stronger evidence of impact and improve the strength of future investigations.

Acknowledgments

The authors appreciate the support provided by the School of Electrical and Computer Engineering and the College of Engineering at the institution where the study was conducted, as well as support from the Bill Kent Family Foundation. We appreciate the collaborative relationship with the course instructors. The authors would also like to acknowledge the assistance of ChatGPT in polishing the language of this paper.

References

[1] OpenAI, *Introducing chatgpt*, 2022.

[2] OpenAI et al., *GPT-4 Technical Report*, 2023.

[3] A. Chowdhery et al., "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.

[4] H. Touvron et al., "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[5] G. Team et al., "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.

[6] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature medicine*, vol. 29, no. 8, pp. 1930–1940, 2023.

[7] K. Singhal et al., "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[8] D. M. Katz, M. J. Bommarito, S. Gao, and P. Arredondo, "Gpt-4 passes the bar exam," *Philosophical Transactions of the Royal Society A*, vol. 382, no. 2270, p. 20 230 254, 2024.

[9] I. Cheong, K. Xia, K. K. Feng, Q. Z. Chen, and A. X. Zhang, "(a) i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024, pp. 2454–2469.

[10] D. Nam, A. Macvean, V. Hellendoorn, B. Vasilescu, and B. Myers, "Using an llm to help with code understanding," in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, 2024, pp. 1–13.

[11] X. Hou et al., "Large language models for software engineering: A systematic literature review," *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 8, pp. 1–79, 2024.

[12] Z. Ke et al., "A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems," *arXiv preprint arXiv:2504.09037*, 2025.

[13] J. Wei et al., "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682*, 2022.

[14] W. X. Zhao et al., "A survey of large language models," *arXiv preprint arXiv:2303.18223*, vol. 1, no. 2, 2023.

[15] E. Kasneci et al., "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and individual differences*, vol. 103, p. 102 274, 2023.

[16] D. Baidoo-Anu and L. O. Ansah, "Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning," *Journal of AI*, vol. 7, no. 1, pp. 52–62, 2023.

[17] M. Kazemitabaar, J. Chow, C. K. T. Ma, B. J. Ericson, D. Weintrop, and T. Grossman, "Studying the effect of ai code generators on supporting novice learners in introductory programming," in *Proceedings of the 2023 CHI conference on human factors in computing systems*, 2023, pp. 1–23.

[18] S. Frieder, "Mathematical capabilities of chatgpt," *arXiv preprint arXiv:2301.13867*, 2023.

[19] J. Finnie-Ansley, P. Denny, B. A. Becker, A. Luxton-Reilly, and J. Prather, "The robots are coming: Exploring the implications of openai codex on introductory programming," in *Proceedings of the 24th Australasian computing education conference*, 2022, pp. 10–19.

[20] B. A. Becker, P. Denny, J. Finnie-Ansley, A. Luxton-Reilly, J. Prather, and E. A. Santos, "Programming is hard-or at least it used to be: Educational opportunities and challenges of ai code generation," in *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, 2023, pp. 500–506.

[21] A. Lewkowycz et al., "Solving quantitative reasoning problems with language models," *Advances in neural information processing systems*, vol. 35, pp. 3843–3857, 2022.

[22] J. Qadir, "Engineering education in the era of chatgpt: Promise and pitfalls of generative ai for education," in *2023 IEEE global engineering education conference (EDUCON)*, IEEE, 2023, pp. 1–9.

[23] W. Sun, L. Chen, Y. Cai, H. Xie, Y. Zeng, and Y. Zhang, "Edu-circuit-hw: Evaluating multimodal large language models on real-world university-level stem student handwritten solutions," *arXiv preprint arXiv:2602.00095*, 2026.

[24] S. S. Nguyen and J. Kittur, "Review of current and potential uses of large language models in engineering," in *Frontiers in Education*, Frontiers Media SA, vol. 10, 2025, p. 1 649 650.

[25] L. Chen, H. Xie, Z. Qin, Y. Guo, J. Rohde, and Y. Zhang, "Enhancing large language models for automated homework assessment in undergraduate circuit analysis," in *2025 IEEE Frontiers in Education Conference (FIE)*, Nashville, TN, USA, 2025, pp. 1–9.

[26] L. Skelic, Y. Xu, M. Cox, W. Lu, T. Yu, and R. Han, *Circuit: A benchmark for circuit interpretation and reasoning capabilities of llms*, 2025.

[27] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 610–623.

[28] E. M. Bender, E. Costello, K. Lee, R. Farrow, and G. Ferreira, "Unsafe ai for education: A conversation on stochastic parrots and other learning metaphors," *Journal of Interactive Media in Education*, vol. 2025, no. 1, 2025.

[29] S. Laato, B. Morschheuser, J. Hamari, and J. Björne, "AI-Assisted Learning with ChatGPT and Large Language Models: Implications for Higher Education," in *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, Jul. 2023, pp. 226–230.

[30] X. Liu et al., "Know3-rag: A knowledge-aware rag framework with adaptive retrieval, generation, and filtering," *arXiv preprint arXiv:2505.12662*, 2025.

[31] L. Ouyang et al., "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[32] S. Sonkar, K. Ni, S. Chaudhary, and R. Baraniuk, "Pedagogical alignment of large language models," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 13 641–13 650.

[33] A. Pereira, A. Gomes, and T. Primo, "Evaluation of Human-Artificial Hybrid Tutoring System for Mediation of Engagement in E-Learning," *Revista Brasileira de Informática na Educação*, vol. 33, p. 216, Apr. 2025.

[34] K. R. Koedinger, A. T. Corbett, and C. Perfetti, "The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning," en, *Cognitive Science*, vol. 36, no. 5, pp. 757–798, 2012.

[35] A. Vaswani et al., "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[36] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[37] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., "Improving language understanding by generative pre-training," 2018.

[38] R. FattahiBavandpour, "Advancing education with large language models: A systematic review of potential, limitations, and business opportunities," 2024.

[39] L. Chen, W. Sun, H. Xie, Y. Cai, and Y. Zhang, "Enhancing large language models for end-to-end circuit analysis problem solving," *arXiv preprint arXiv:2512.10159*, 2025.

[40] L. Chen, H. Xie, J. Rohde, and Y. Zhang, "Wip: Large language model-enhanced smart tutor for undergraduate circuit analysis," in *2025 IEEE Frontiers in Education Conference (FIE)*, Nashville, TN, USA, 2025, pp. 1–5.

[41] S. Boguslawski, R. Deer, and M. G. Dawson, "Programming education and learner motivation in the age of generative AI: Student and educator perspectives," *Information and Learning Sciences*, vol. 126, no. 1-2, pp. 91–109, Jul. 2024.

[42] H. Jin, S. Lee, H. Shin, and J. Kim, "Teach AI How to Code: Using Large Language Models as Teachable Agents for Programming Education," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI '24, New York, NY, USA: Association for Computing Machinery, May 2024, pp. 1–28.

[43] Z. Levonian et al., "Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference," *arXiv preprint arXiv:2310.03184*, 2023.

[44] L. Skelic, Y. Xu, M. Cox, W. Lu, T. Yu, and R. Han, "Circuit: A benchmark for circuit interpretation and reasoning capabilities of llms," *arXiv preprint arXiv:2502.07980*, 2025.

[45] C. Dignath and G. Büttner, "Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and secondary school level," en, *Metacognition and Learning*, vol. 3, no. 3, pp. 231–264, Dec. 2008.

[46] B. J. Zimmerman, "Attaining self-regulation: A social cognitive perspective," in *Handbook of self-regulation*, Elsevier, 2000, pp. 13–39.

[47] J. Liu et al., "Socraticlm: Exploring socratic personalized teaching with large language models," in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, ser. NIPS '24, Vancouver, BC, Canada: Curran Associates Inc., 2024.

[48] W. Ge et al., "Srlagent: Enhancing self-regulated learning skills through gamification and llm assistance," *arXiv preprint arXiv:2506.09968*, 2025.

[49] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[50] Y. Gao et al., "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, vol. 2, no. 1, 2023.

[51] M. Cheng et al., "A survey on knowledge-oriented retrieval-augmented generation," *arXiv preprint arXiv:2503.10677*, 2025.

[52] D. Thüs, S. Malone, and R. Brünken, "Exploring generative ai in higher education: A rag system to enhance student engagement with scientific literature," *Frontiers in Psychology*, vol. 15, p. 1 474 892, 2024.

[53] S. Barnett, S. Kurniawan, S. Thudumu, Z. Brannelly, and M. Abdelrazek, "Seven failure points when engineering a retrieval augmented generation system," in *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, 2024, pp. 194–199.

[54] M. A. Shavaki, P. Omrani, R. Toosi, and M. A. Akhaee, "Knowledge Graph Based Retrieval-Augmented Generation for Multi-Hop Question Answering Enhancement," en-US, in *2024 15th International Conference on Information and Knowledge Technology (IKT)*, Dec. 2024, pp. 78–84.

[55] L. Chen, Z. Qin, Y. Guo, J. Rohde, and Y. Zhang, "Benchmarking large language models on homework assessment in circuit analysis," *International Journal of Artificial Intelligence in Education*, pp. 1–62, 2025.

[56] G. V. Glass, "9: Integrating findings: The meta-analysis of research," *Review of Research in Education*, vol. 5, no. 1, pp. 351–379, 1977.

[57] C. Andrade, "Mean difference, standardized mean difference (smd), and their use in meta-analysis: As simple as it gets.," *The Journal of clinical psychiatry*, vol. 81, no. 5, 20f13681–20f13681, 2020.

[58] J. Cohen, *Statistical power analysis for the behavioral sciences*. routledge, 2013.

[59] J. A. Svoboda and R. C. Dorf, *Introduction to Electric Circuits*. John Wiley & Sons, 2013.

A AI Tutor User Interface Details

Figures 9 and 10 show the screens displayed after loading a problem and clicking "Submit My Answer for Feedback" and "Request Guidance", respectively.

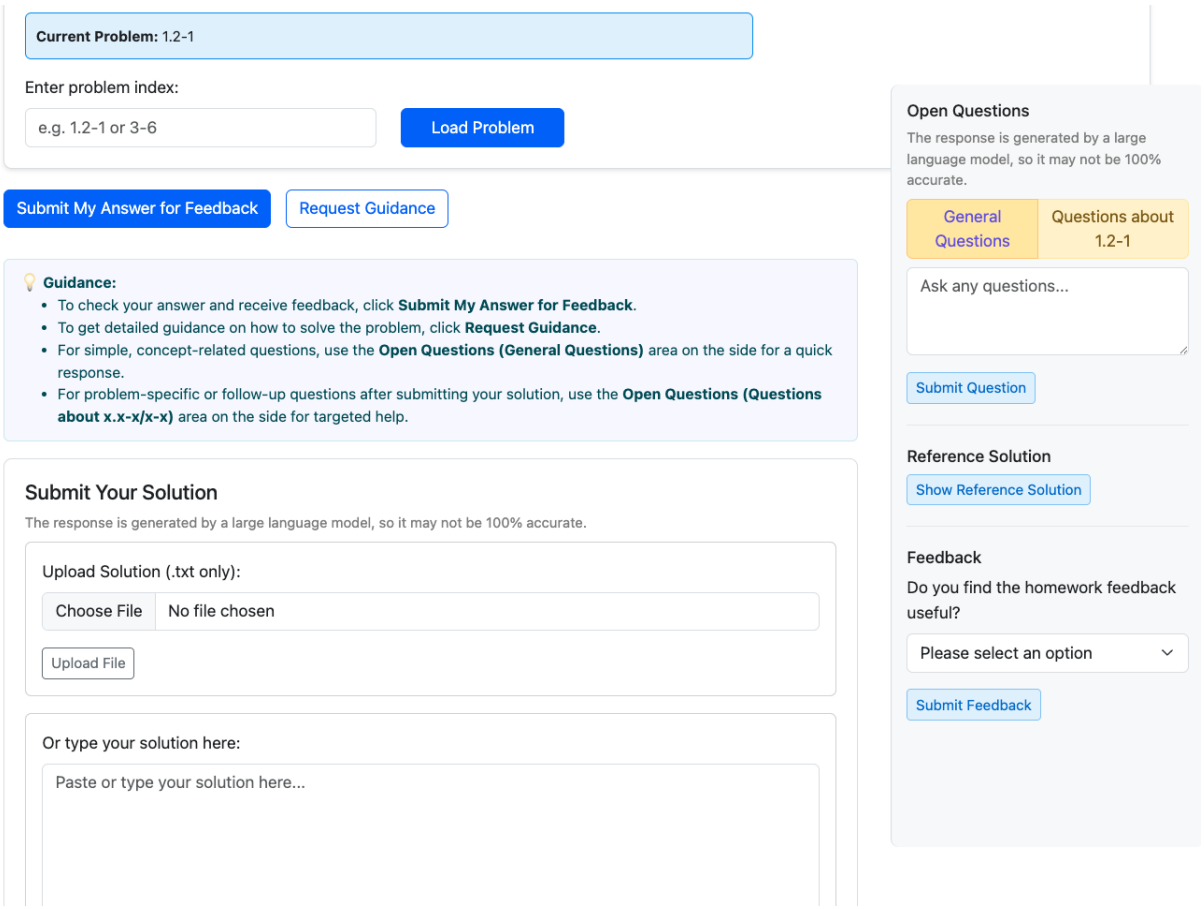


Figure 9: User interface of the AI tutor: screen shown after loading a problem and clicking "Submit My Answer for Feedback".

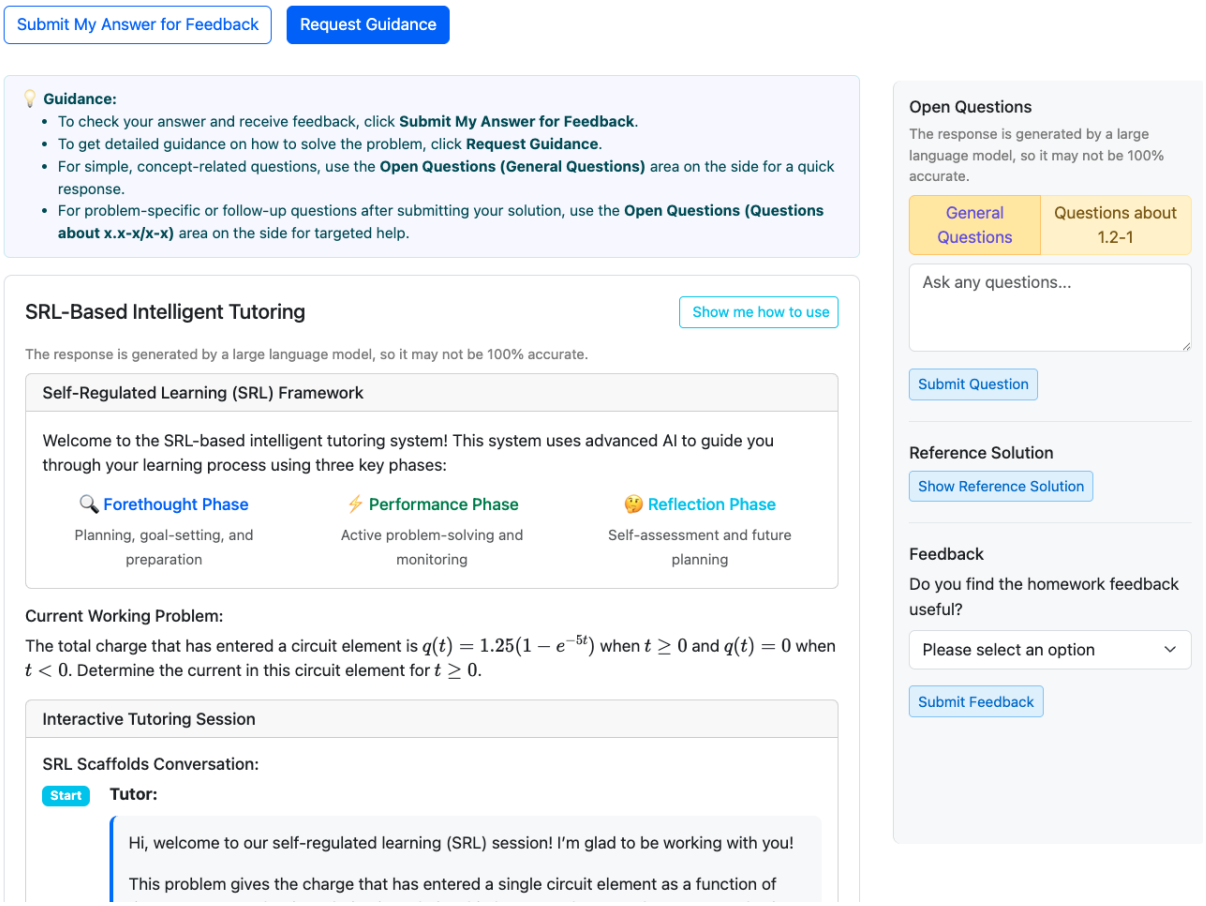


Figure 10: User interface of the AI tutor: screen shown after loading a problem and clicking "Request Guidance". The current problem displayed is adapted from Problem 1.2-1 in the textbook [59].

B Course Details

Tables 5 and 6 present the grading breakdown and homework details for the *Circuit Analysis* course.

Component	Weight	Full Score
Homework	15%	101.36 (126.36)
Experiments	5%	100
Class Participation	15%	100
Quiz 1	20%	100
Quiz 2	20%	100
Final Exam	25%	100
Total	100%	100.20 (103.95)

Table 5: *Circuit Analysis* Course Grading Breakdown. The full score for the homework component is the average total score across 11 assignments. The value in parentheses indicates the average full score including optional problems, as detailed in Table 6.

Assignment Index	Problem Number	Full Score
1	10 (5)	100 (150)
2	10 (4)	100 (140)
3	10 (0)	100 (100)
4	10 (3)	100 (130)
5	12 (1)	120 (130)
6	10 (2)	100 (120)
7	10 (3)	100 (130)
8	8 (2)	100 (125)
9	7 (3)	105 (150)
10	6 (1)	90 (105)
11	10 (1)	100 (110)
Total	103 (25)	1115 (1390)

Table 6: *Circuit Analysis* Homework Information. Some assignments include optional problems. In the "Problem Number" column, numbers in parentheses indicate the count of optional problems. In the "Full Score" column, values in parentheses represent the total possible score including optional problems.