Retrieval-Augmented Tutoring for Algorithm Tracing and Problem-Solving in AI Education

Mragisha Jain[1], Tirth Bhatt[1], Griffin Pitts[1], Aum Pandya[1], Peter Brusilovsky[2], Narges Norouzi[3], Arto Hellas[4], Juho Leinonen[4], Bita Akram[1*]

[1]North Carolina State University, [2]University of Pittsburgh,
[3]University of California, Berkeley, [4]Aalto University,

***Correspondence:** bakram@ncsu.edu

Abstract

Students learning algorithms often need support as they interpret traces, debug reasoning errors, and apply procedures across unfamiliar problem instances. In this paper, we present KITE (Knowledge-Informed Tutoring Engine), a Retrieval-Augmented Generation (RAG)-based intelligent tutoring system designed to serve as a classroom teaching assistant for algorithmic reasoning and problem-solving tasks. KITE uses an intent-aware Socratic response strategy to tailor support to different student needs, responding with targeted hints, guiding questions, and progressive scaffolding intended to strengthen students' algorithmic problem-solving ability. To keep responses aligned with course content, KITE uses a multimodal RAG pipeline that retrieves relevant information from course materials. We evaluate KITE using three forms of assessment: RAGAs-based metrics for response grounding and quality, expert evaluation of pedagogical quality, and a simulated student pipeline in which a weaker language model interacts with KITE across two-turn dialogues and produces revised answers after receiving feedback. Results indicate that KITE produces contextually grounded and pedagogically appropriate responses. Further, using simulated students, KITE's feedback helped the student models produce more accurate follow-up responses on procedural and tracing questions, suggesting that its scaffolding can support algorithmic problem-solving. This work contributes a tutoring architecture and an evaluation approach for assessing retrieval-grounded explanations and scaffolded problem-solving feedback.

1 Introduction

Large language models (LLMs) such as ChatGPT are now widely used by students for learning support, including explanation, feedback, and problem-solving (Pitts et al., 2025b; Pitts and Motamedi, 2026). Students often value these tools because they provide immediate access to assistance when instructors or teaching assistants are unavailable (Pitts et al., 2025b). Although these tools make information more accessible, prior work raises concerns that students may accept AI-generated responses without sufficient evaluation, especially when those responses appear complete and confident (Essel et al., 2024; Pitts et al., 2025c, 2026). In education, this can lead students to bypass the reasoning processes that assignments are designed to develop (Pitts et al., 2025c, 2026). These concerns highlight the need for LLM-based systems that provide timely, course-grounded information while delivering pedagogically appropriate support that helps students reason through learning tasks.

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) offers a promising approach for building course-grounded tutoring systems by allowing LLM responses to draw on curated instructional materials. This grounding can reduce unsupported or course-inconsistent claims and help align explanations with the concepts, terminology, and conventions used in a course. However, strong retrieval does not by itself ensure effective tutoring. Even when a system retrieves relevant material, it may still provide responses that are too direct, insufficiently instructional, or mismatched to the student's immediate learning need. Prior work on intelligent tutoring systems suggests that effective support depends on both the accuracy of the information provided and how assistance is delivered, including when to offer direct explanation, feedback, or more guided support (Koedinger and Aleven, 2007).

Socratic tutoring offers one way to address this challenge by guiding learners through targeted questions, prompts, and progressive hints instead of immediately providing full solutions. This approach is grounded in cognitive apprenticeship and guided facilitation (Collins et al., 1989; Hmelo-Silver and Barrows, 2006), and has been used in dialogue-based tutoring systems such as AutoTutor (Graesser et al., 1999). However, integrating

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 841–850

Socratic guidance into retrieval-grounded tutoring systems remains an open design problem: a course-specific tutor must stay faithful to instructional materials while also providing feedback that fits the type of problem the student is trying to solve.

In this paper, we present **KITE** (Knowledge-Informed Tutoring Engine), a RAG-based intelligent tutoring system that connects students to relevant course materials while using intent-aware tutoring strategies to support different forms of help-seeking. KITE uses a multi-stage multimodal retrieval pipeline to locate relevant instructional content and an intent-aware response strategy to determine how that content should be used in the response. For questions that require direct explanation, KITE provides responses aligned with retrieved course materials. For procedural, debugging, validation, and tracing questions, KITE provides targeted feedback, guiding questions, and progressive hints to support student reasoning. To evaluate KITE, we first assess its retrieval-grounded outputs for non-procedural questions using RAGAS-based metrics for grounding, relevance, and response quality. We then evaluate procedural and tracing questions through a simulated student pipeline in which a weaker language model revises its answers after receiving KITE's feedback. Finally, human experts assess the resulting interactions to judge feedback quality and whether the revised answers show improvement. This work contributes (1) KITE, an intent-aware tutoring system that combines multimodal retrieval with pedagogical support, and (2) an evaluation of its retrieval-grounded responses and scaffolded feedback using automated metrics, simulated students, and expert evaluation. We assess two research questions: **RQ1:** How well does KITE produce grounded, course-relevant responses for non-procedural student questions? and **RQ2:** To what extent does KITE's feedback support improved responses on procedural and tracing questions?

2 Related Work

RAG-based educational assistants have been used for a range of instructional purposes, including interactive learning support, content generation, and large-scale course deployment (Li et al., 2025). Across these systems, grounding LLM responses in course materials has generally improved factual accuracy compared to unaugmented models. For example, KAG (Hasan et al., 2025) reports Precision@5 of 0.85 and a 34% reduction in student task completion time, while MoodleBot (Neumann et al., 2024) achieves 88% accuracy on course-related queries. However, these systems primarily function as direct question-answering tools and do not adapt their responses to different forms of student help-seeking.

Although RAG can improve factual accuracy, deployment studies suggest that course-grounded assistants also need evaluation in instructional workflows. In one classroom deployment, students showed strong pre-exam engagement but declining adoption across cohorts, and 36.8% reported frustration when responses extended beyond a constrained knowledge base (Thesen and Park, 2025). Edison (Miroyan et al., 2025), a GPT-4-based RAG assistant deployed in a large data science course, showed that retrieving from course documents and historical Q&A can support factual and relevant responses to live student questions. The study also demonstrates the value of TA-in-the-loop evaluation, using instructor edits and ratings to assess factuality, relevance, style, and efficiency. EduMod-LLM (Mittal et al., 2026) extends this line of work by treating educational Q&A as a modular pipeline, separating function calling, retrieval, and response generation so that system behavior can be evaluated more transparently.

Dialogue-based tutoring provides another foundation for supporting student reasoning. AutoTutor (Graesser et al., 1999) showed that progressive hints and collaborative answer refinement produced dialogues rated above the "good" threshold by domain experts, with semantic evaluation correlating at 0.49 with expert judgment. More recently, (Li et al., 2026) reported significant gains in self-efficacy ($d = 0.57$) from a Socratic AI platform in healthcare education. LeanTutor (Patel et al., 2026) similarly emphasizes guided feedback by combining LLMs with a theorem prover to check student proofs, identify errors, and provide hints toward a correct proof without giving away the complete answer. These systems show the value of scaffolded feedback for learning tasks that require students to reason through a process, yet they do not incorporate retrieval grounding to keep responses aligned with course-specific materials.

Other systems explore how retrieval and response strategies can be adapted for learning contexts. LPITutor (Liu et al., 2025) supports adaptive difficulty modulation through RAG and prompt engineering. KG-RAG (Dong et al., 2025) combines

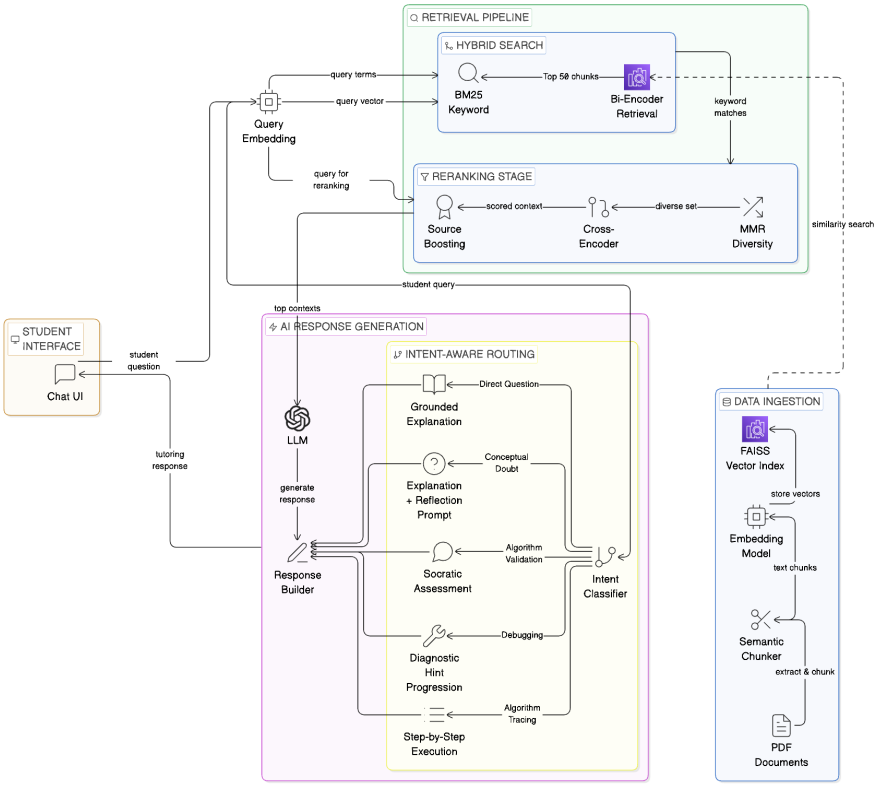

Figure 1: KITE architecture

semantic retrieval with an expert-validated knowledge graph and reports a 35% improvement in learning outcomes ($d = 0.86$) in a study of 76 students, though its reliance on manual expert validation limits scalability. AutoTA (Dahal et al., 2025) provides a related approach to intent-aware educational assistance by classifying student queries and routing them to specialized response strategies. These systems show the value of adapting retrieval, domain structure, and response behavior to different learning needs. KITE builds on this direction by pairing multi-stage retrieval with intent-aware tutoring strategies for algorithmic reasoning tasks.

3 System Design

KITE is a retrieval-augmented tutoring system designed to support course-grounded dialogue for algorithmic reasoning and problem-solving tasks. As shown in Figure 1, the system includes five components: document preprocessing, embedding generation, multi-stage retrieval, intent-aware response generation, and session management.

3.1 Phase 1: Document Ingestion and Preprocessing

KITE begins by extracting text from course PDFs using *PyMuPDF*. Extraction is performed page by page so that the original document structure remains traceable during retrieval. To reduce noise before indexing, the system applies a **frequency-based** cleaning procedure that removes repeated headers, footers, page numbers, and other formatting artifacts. Specifically, it examines the first and last two lines of each page, identifies repeated patterns that occur across pages, removes those patterns along with page numbers and special characters, and normalizes whitespace.

The cleaned text is then segmented into semantically coherent chunks for retrieval. We use **section-aware chunking** with a target size of 500 characters, or about 125 tokens, and a 100-character overlap. Headers are identified and retained to preserve local structure, while overlap carries forward the final two sentences of the preceding chunk.

3.2 Phase 2: Embedding Generation

Each chunk is encoded using OpenAI's `text-embedding-3-large` model, producing 3072-dimensional embeddings. These vectors are L2-normalized so cosine similarity reflects semantic direction and are stored in a FAISS index (Johnson et al., 2019) for efficient local retrieval.

3.3 Phase 3: Multi-Stage Retrieval Pipeline

KITE uses a multi-stage pipeline designed to balance high recall and precision in retrieving course content. Retrieval begins with a dense bi-encoder search that returns the top 50 candidate chunks for a given student query. The query and document chunks are encoded independently, and similarity is computed using cosine similarity, allowing the system to capture semantically related content.

The candidate set is then refined through hybrid retrieval. Dense similarity contributes 70% of the retrieval score, while sparse BM25 keyword matching contributes 30%. This combination captures both semantic similarity and exact lexical overlap, which is useful when students use course-specific terminology, notation, or algorithm names.

To reduce redundancy among retrieved passages, KITE applies Maximal Marginal Relevance (MMR) with λ set to 0.7:

$$MMR = \lambda \times \text{Relevance} + (1 - \lambda) \times \text{Diversity}$$

The retrieved candidates are reranked using a `cross-encoder/ms-marco-MiniLM-L-6-v2` reranking model implemented through Sentence Transformers, where the query and document are jointly encoded to produce more precise relevance scores. Finally, KITE applies source-based boosting so that chunks from official course materials receive higher priority. Chunks with reranking scores above 0.6 receive an additional boost of 0.3. The final context passed to the generator consists of the top eight retrieved chunks.

3.4 Phase 4: Intent Classification and Pedagogical Response Generation

KITE does not use a single response strategy for all student questions. Instead, it first classifies each query by pedagogical intent and then generates a response that matches the instructional purpose of the interaction. This allows the system to distinguish among questions, debugging requests, and other forms of help-seeking.

3.4.1 Intent Classification

Each incoming query is classified into one of five pedagogical intents using a keyword and pattern-matching classifier, as shown in Figure 1.

- **Direct Question**: factual queries seeking definitions or explanations (e.g., "What is A*?").
- **Conceptual Questions**: deeper *why* or *how* questions probing understanding (e.g., "Why does BFS guarantee shortest path?").
- **Algorithm Validation**: queries where a student submits their own implementation or trace for assessment.
- **Debugging**: queries involving a specific error or incorrect output.
- **Algorithm Tracing**: requests to step through the execution of an algorithm on a concrete problem instance (e.g., "Trace A* on this graph starting from node S").

The classified intent determines which response generation strategy is invoked. In addition to these five query intents, KITE includes a dedicated answer evaluation mode for cases in which a student submits a written answer for assessment. This mode bypasses intent classification and routes directly to the feedback generation pipeline.

3.4.2 LLM Generation and Intent-Aware Response Strategy

All response generation in KITE is handled by GPT-5. Outputs are grounded in a structured prompt that injects the top eight retrieved chunks into a `[CONTEXT]` block. The model is instructed to prioritize course materials and avoid introducing information that is unsupported by the retrieved context, helping keep responses aligned with the course.

For direct questions and conceptual doubts, KITE produces explanations grounded in the retrieved material. Responses are written in a tutor-like tone that emphasizes reasoning instead of brief answer delivery. For conceptual doubts, the response also includes a follow-up question intended to prompt reflection.

For algorithm validation tasks, KITE adopts a Socratic assessment strategy instead of directly identifying errors. Responses include a brief evaluation of the student's approach, acknowledgement of correct components, and guiding questions that target specific issues. This design supports learning without explicitly revealing the final solution, encouraging students to continue working through the problem independently.

For debugging assistance, KITE generates diagnostic prompts that guide students toward identifying errors through self-examination. Each response follows a structured hint progression and includes a learning point that connects the observed bug to the underlying conceptual principle, reinforcing understanding beyond the immediate correction.

For algorithm tracing queries, KITE retrieves the relevant procedural steps and rules from course materials and applies them step by step to the student's specific problem instance. Each step explicitly maintains and updates algorithmic state variables such as OPEN lists, CLOSED sets, and selected nodes, following the tie-breaking rules and constraints defined in the query. The response concludes with the final path and cost when applicable.

3.5 Phase 5: Session Management

KITE maintains session state across multi-turn interactions to preserve continuity within a conversation. For each session, the system stores the original query, detected intent, prior responses, and any hints provided. When a student submits a follow-up query, KITE uses this stored context to determine how the interaction should continue: related direct and conceptual questions are treated as follow-ups, while validation, debugging, and tracing requests remain within their intent-specific response strategy when they concern the same problem or algorithm. For these turns, KITE constructs a brief context summary from the prior interaction and appends it to the retrieval prompt to reduce repetition and support progressive guidance.

4 Methodology

To evaluate KITE, we use three forms of assessment. First, we examine non-procedural responses using RAGAs-based metrics for grounding, relevance, and answer quality against instructor-authored reference answers. We then use a simulated student pipeline to assess whether KITE's feedback helps produce improved responses on procedural and tracing questions. Finally, experts evaluate the pedagogical quality of KITE's feedback and the resulting answer revisions.

4.1 Evaluation Dataset

We constructed an evaluation dataset of 109 questions drawn from the lecture slides and textbook used in a university *Introduction to AI* course, with each question paired with an instructor-verified reference answer. The dataset included 42 algorithmic questions, 51 procedural questions, and 16 direct-retrieval questions. We applied RAGAs to the 58 non-procedural questions, consisting of the algorithmic and direct-retrieval subsets, because these responses can be evaluated against reference answers for grounding, relevance, and answer quality. Questions requiring procedural reasoning or algorithm tracing were evaluated separately through the simulated student pipeline and expert review described in Section 4.3.

4.2 RAGAs Evaluation

We evaluate KITE's non-procedural responses using the RAGAs framework (Es et al., 2024; Roychowdhury et al., 2024), reporting six metrics:

- **Faithfulness**: Measures whether statements in the generated response are supported by the retrieved context, computed as the proportion of answer claims judged to be grounded in the retrieved chunks.
- **Answer Relevance**: Measures how well the generated response addresses the original question, computed from the cosine similarity between the user's question and questions generated from the response.
- **Context Relevance**: Measures how much of the retrieved context is relevant to answering the question, computed as the proportion of retrieved statements judged to be useful.
- **Answer Similarity**: Measures semantic similarity between the generated response and the instructor-authored reference answer using sentence embeddings.
- **Factual Correctness**: Measures factual agreement between the generated response and the reference answer using an F1 score over claims classified as true positives, false positives, and false negatives.
- **Answer Correctness**: Measures overall correctness of the generated response relative to the reference answer as a weighted combination of factual correctness (0.75) and answer similarity (0.25).

All metrics use `gpt-4o-mini` as the judge model and `text-embedding-3-small` for similarity metrics. Retrieval uses `top_k=5` from an initial candidate pool of 50.

4.3 Simulated Student Evaluation and Expert Evaluation

For procedural and algorithm-tracing questions, KITE provides Socratic feedback and guidance, making standard automatic scoring less appropriate. To evaluate how well this feedback supports learning-oriented revision, we use a two-stage simulated student pipeline followed by expert review.

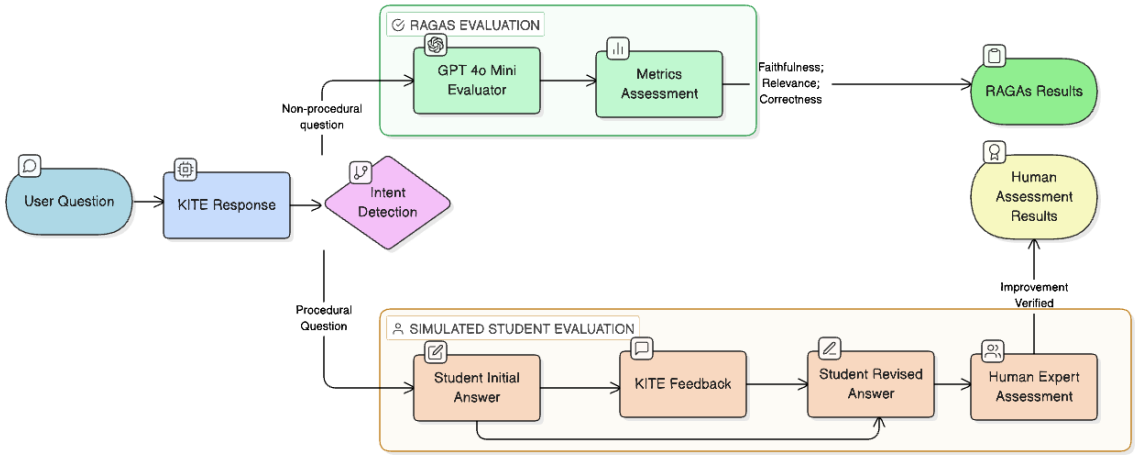

Figure 2: Evaluation pipeline

Simulated Student Pipeline. Building on prior work that uses simulated student-tutor interactions to evaluate pedagogical support (Dinucu-Jianu et al., 2025), we use Meta-Llama-3.1-70B-Instruct as a proxy student in a structured interaction with KITE:

1. **Round 1**: The student model answers each question without assistance, establishing an unaided baseline.
2. **KITE Feedback**: KITE evaluates the student's answer and provides feedback intended to guide revision.
3. **Round 2**: The student model revises its answer using KITE's feedback.

Expert Evaluation. Three experts reviewed each interaction set, consisting of the Round 1 answer, KITE's feedback, and the Round 2 answer. They judged whether the revised response demonstrated improved correctness and reasoning, and evaluated the quality of KITE's feedback using a structured rubric adapted from prior work (Pauzi et al., 2025).

The rubric includes three dimensions. *Mistake Remediation* assesses whether the tutor correctly identifies the student's error and explicitly acknowledges it in the response. *Scaffolding and Guidance* assesses whether the tutor provides appropriate support without revealing the answer and offers clear next-step direction. *Coherence and Tone* assesses whether the dialogue reads naturally and maintains an encouraging and supportive tone. Each criterion is scored as Yes/No, with NA used when a criterion is not applicable.

5 Results

5.1 RAGAs Evaluation

Table 1 reports the six RAGAs metrics evaluated on the 58 non-procedural questions, consisting of 42 algorithmic questions and 16 direct-retrieval questions.

Metric	Mean	Std. Dev.
Faithfulness	0.8486	0.2103
Answer Relevance	0.7558	0.2032
Context Relevance	0.9352	0.1905
Answer Similarity	0.7586	0.0923
Factual Correctness	0.4483	0.2477
Answer Correctness	0.6363	0.1810

Table 1: RAGAs evaluation summary ($n = 58$).

KITE performs strongly on retrieval and grounding metrics. Faithfulness (0.85) indicates that most answer statements are supported by the retrieved context, while context relevance (0.94) shows that the retrieved passages are highly pertinent to the question. Answer relevance (0.76) and answer similarity (0.76) further suggest that KITE's responses remain on-topic and semantically aligned with instructor-authored reference answers.

Factual correctness (0.45) is lower than the other RAGAs measures. As discussed in Section 7, this metric is sensitive to claim-level overlap with a single reference answer and may understate the quality of responses that are accurate but phrased differently or provide additional valid detail. For this reason, answer similarity is used as the primary indicator of response quality in this setting. Its low

variance ($\sigma = 0.09$) also suggests relatively consistent performance across the evaluated questions.

5.2 Simulated Student and Expert Evaluation

Table 2 summarizes the expert rubric scores for 44 simulated student–KITE interaction triples. Inter-rater agreement between the two expert annotators was high, with Cohen's $\kappa = 0.88$ and a raw agreement rate of 98.15%, indicating strong consistency in rubric judgments.

Metric	% Yes	% No	% N/A
Mistake Remediation (Identifying)	63.63	6.82	29.55
Mistake Remediation (Acknowledging)	63.63	6.82	29.55
Scaffolding	93.18	6.82	—
Guidance	93.18	6.82	—
Coherence (Naturalness)	93.18	6.82	—
Tone (Encouraging)	93.18	6.82	—

Table 2: Expert evaluation rubric results ($n = 44$).

KITE receives consistently high ratings for scaffolding, guidance, coherence, and tone, with 93.18% Yes judgments on each dimension. These results indicate that its feedback is generally well-structured, actionable, and supportive throughout the interaction. Mistake remediation receives 63.63% Yes judgments, but 29.55% of cases are marked N/A because the simulated student's initial response was already correct and no error identification was required. When remediation is applicable, the results indicate that KITE identifies and acknowledges student errors appropriately.

Answer Improvement. Table 3 reports how students responses changed from Round 1 to Round 2 after receiving KITE's feedback. The transition labels reflect expert judgments of whether responses were Incorrect, Partially Correct, or Correct with respect to the course materials. Among the 27 interactions that were not already correct, 24 showed improvement after KITE's feedback (88.89%).

The most common transition was from Partially Correct to Correct, occurring in 14 cases (31.82%). This suggests that KITE is particularly effective at helping students resolve remaining reasoning gaps in responses that are already moving in the right direction. In six additional cases (13.63%), the response remained Partially Correct but still improved in quality, indicating that KITE's feedback can support meaningful revision even when the student model does not reach a correct answer.

Transition	Count	%
Incorrect → Correct	1	2.27
Incorrect → Partially Correct	3	6.82
Already Correct	17	38.64
Partially Correct → Correct	14	31.82
Partially Correct → Partially Correct with Improvement	6	13.63
N/A	3	6.82

Table 3: Answer improvement breakdown ($n = 44$).

6 Discussion

This study examined whether a course-grounded, intent-aware tutoring system could provide reliable retrieval-based support and pedagogically useful feedback for problem-solving tasks. The results are encouraging with regard to both aims. For RQ1, KITE's faithfulness (0.85) and context relevance (0.94) indicate that its responses are closely grounded in retrieved course material, while answer similarity (0.76) shows consistent alignment with instructor-authored reference answers. For RQ2, among the 27 simulated-student interactions with KITE, in which the initial student response was not already correct, experts judged 24 revised answers (88.89%) as improved after receiving KITE's feedback. This suggests that KITE's feedback provided guidance the student model could use to correct or strengthen its reasoning in a follow-up response. Experts rated 93.18% of KITE's feedback positively for scaffolding and guidance, further indicating that the feedback was instructionally purposeful and well-structured.

Retrieval and response quality. The RAGAs evaluation indicated that KITE performed well on measures tied to retrieval and grounding, with context relevance of 0.94 and faithfulness of 0.85. These results show that KITE's multi-stage retrieval pipeline surfaced course-specific material relevant to the questions and that its responses remained closely grounded in that retrieved context. At the same time, the lower factual correctness score (0.45) warrants careful interpretation, particularly relative to answer similarity (0.76). Prior work has noted that RAGAs-style claim matching and score stability can vary across response formulations and evaluation settings (Roychowdhury et al., 2024; Antal and Buza, 2025). In light of prior work, the 0.31-point gap observed in our results may reflect limitations of factual correctness as a reference-based metric for evaluating KITE's pedagogically framed responses, although our evaluation does not isolate the source of that discrepancy.

Pedagogical effectiveness and design implications. Following the RAGAs evaluation, the simulated student and expert evaluations examined whether KITE's feedback supported stronger revised answers, and was judged to be pedagogically appropriate. Specifically, the 88.89% improvement rate in the simulated student pipeline, together with the strong expert rubric scores for scaffolding, guidance, coherence, and tone, support that KITE's feedback can support stronger follow-up answers on procedural and tracing questions. This emphasis on feedback quality is consistent with prior survey work on LLM applications in programming education, which argues that the educational value of these systems depends on aligning model capabilities with pedagogical goals, including the use of scaffolding and feedback strategies (Pitts et al., 2025a). In KITE, this alignment is reflected in pairing retrieval-grounded generation with feedback strategies designed for different forms of student support, such as direct explanations, or algorithmic-tracing guidance. While these findings are encouraging, the current evaluation design limits their generalizability, as discussed in Section 6.1.

6.1 Limitations and Future Work

This study has limitations that motivate future work. First, our evaluation of factual correctness is constrained by limitations of the RAGAs framework. KITE produces pedagogically framed explanations that may paraphrase or elaborate on course material, whereas RAGAs decomposes each response into atomic claims and uses NLI-style entailment to assess agreement with a single instructor-authored answer. With this, responses that are semantically aligned with the expected answer but differ in phrasing, detail, or framing may receive lower factual correctness scores. The 0.31-point gap between factual correctness (0.45) and answer similarity (0.76) in Table 1 is consistent with this concern, although our evaluation does not isolate the source of that discrepancy. Answer similarity, which is based on semantic embeddings, remains substantially higher and shows low variance ($\sigma = 0.09$), indicating relatively stable semantic alignment across questions. We therefore treat answer similarity as the primary quality indicator and note that using multiple human-written answers could reduce this limitation in future evaluations. Related concerns have been noted in prior work: Roychowdhury et al. (Roychowdhury et al., 2024) discuss limitations in how RAGAs decomposes and assigns statements during metric computation, while Antal and Buza (Antal and Buza, 2025) show that RAGAs-based evaluation outcomes vary across question types and retrieval conditions. Future evaluations could use richer answer sets and metrics.

Second, the simulated student pipeline relies on a single LLM, Meta-Llama-3.1-70B-Instruct, as a proxy for student behavior. As a result, improvement between Round 1 and Round 2 should be interpreted as evidence that KITE's feedback makes a stronger answer more recoverable, not as direct evidence of genuine student learning. Real learners may differ substantially in both the magnitude and pattern of improvement. Although this design is useful for early-stage evaluation, it cannot substitute for classroom evidence. A necessary next step is deployment with real students, including pre- and post-interaction assessments, analysis of revision behavior over time, and closer examination of how learners engage with feedback.

Third, the expert evaluation covers a relatively limited set of interaction cases, which constrains the generalizability of the findings. Although inter-rater agreement was strong ($\kappa = 0.88$), judgments of answer improvement and pedagogical quality still involve subjectivity, and the sample size limits precision. Future work should expand the annotation set and use finer-grained scoring schemes to better capture variation in feedback quality and answer improvement across question and error types.

7 Conclusion

We presented KITE, a RAG-based intelligent tutoring system that combines a five-stage retrieval pipeline with intent-aware pedagogical response generation. KITE adapts its responses to the type of student query, providing grounded explanations for factual questions and Socratic scaffolding for procedural and reasoning tasks. To evaluate these response modes, we introduced a two-part evaluation framework. RAGAs metrics assess retrieval quality, while a simulated student pipeline examines whether KITE's feedback supports improved responses on procedural and tracing questions. Expert review using a structured rubric further evaluates the pedagogical quality of KITE's feedback and verifies improvement in students' revised answers. This work contributes an intent-aware tutoring architecture and an evaluation approach for RAG systems with mixed response strategies.

Acknowledgements

This research was supported by the U.S. National Science Foundation (NSF) under Grant #2426837. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect views of the NSF. This work was additionally supported by Research Council of Finland grants #356114 and #367787.

References

Margit Antal and Krisztian Buza. 2025. Evaluating open-source llms in rag systems: a benchmark on diploma theses abstracts using ragas: M. antal, k. buza. *Acta Universitatis Sapientiae, Informatica*, 17(1):5.

Allan Collins, John Seely Brown, and Susan E. Newman. 1989. Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In Lauren B. Resnick, editor, *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser*, pages 453–494. Lawrence Erlbaum Associates.

Rajashree Dahal, Greg Murray, Robin Chataut, Mohamed Hefeida, Anurag Srivastava, and Prashnna Gyawali. 2025. Autota: A dynamic intent-based virtual teaching assistant for students using open source llms. *IEEE Access*.

David Dinucu-Jianu, Jakub Macina, Nico Daheim, Ido Hakimi, Iryna Gurevych, and Mrinmaya Sachan. 2025. From problem-solving to teaching problem-solving: Aligning llms with pedagogy using reinforcement learning.

Chenxi Dong, Yimin Yuan, Kan Chen, Shupei Cheng, and Chujie Wen. 2025. How to build an adaptive ai tutor for any course using knowledge graph-enhanced retrieval-augmented generation (kg-rag). In *2025 14th International Conference on Educational and Information Technology (ICEIT)*, pages 152–157. IEEE.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations*, pages 150–158.

Harry Barton Essel, Dimitrios Vlachopoulos, Albert Benjamin Essuman, and John Opuni Amankwa. 2024. Chatgpt effects on cognitive skills of undergraduate students: Receiving instant responses from ai-based conversational large language models (llms). *Computers and Education: Artificial Intelligence*, 6:100198.

Arthur C Graesser, Katja Wiemer-Hastings, Peter Wiemer-Hastings, Roger Kreuz, Tutoring Research Group, and 1 others. 1999. Autotutor: A simulation of a human tutor. *Cognitive Systems Research*, 1(1):35–51.

Hadi Hasan, Ali Ismail, Ammar Mohanna, and Ali Chehab. 2025. Kag: A scalable knowledge-augmented generation system for educational content management. In *2025 3rd International Conference on Foundation and Large Language Models (FLLM)*, pages 503–508. IEEE.

Cindy E Hmelo-Silver and Howard S Barrows. 2006. Goals and strategies of a problem-based learning facilitator. *Interdisciplinary journal of problem-based learning*, 1(1):4.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE transactions on big data*, 7(3):535–547.

Kenneth R Koedinger and Vincent Aleven. 2007. Exploring the assistance dilemma in experiments with cognitive tutors. *Educational psychology review*, 19(3):239–264.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Yan Li, Janelle Yorke, Jiaying Li, Mengting He, Yushen Dai, Yan Zhao, Jing Qin, and Xiangen Hu. 2026. An innovative socratic method-based artificial intelligence platform for healthcare education: A quasi-experimental study. *Nurse education in practice*, page 104770.

Zongxi Li, Zijian Wang, Weiming Wang, Kevin Hung, Haoran Xie, and Fu Lee Wang. 2025. Retrieval-augmented generation for educational application: A systematic survey. *Computers and Education: Artificial Intelligence*, 8:100417.

Zhensheng Liu, Prateek Agrawal, Saurabh Singhal, Vishu Madaan, Mohit Kumar, and Pawan Kumar Verma. 2025. Lpitutor: an llm based personalized intelligent tutoring system using rag and prompt engineering. *PeerJ Computer Science*, 11:e2991.

Mihran Miroyan, Chancharik Mitra, Rishi Jain, Gireeja Ranade, and Narges Norouzi. 2025. Analyzing pedagogical quality and efficiency of llm responses with ta feedback to live student questions. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, pages 770–776.

Meenakshi Mittal, Rishi Khare, Mihran Miroyan, Chancharik Mitra, and Narges Norouzi. 2026. Edumod-llm: A modular approach for designing flexible and transparent educational assistants. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 40652–40660.

Alexander Tobias Neumann, Yue Yin, Sulayman Sowe, Stefan Decker, and Matthias Jarke. 2024. An llm-driven chatbot in higher education for databases and information systems. *IEEE Transactions on Education*, 68(1):103–116.

Manooshree Patel, Rayna Bhattacharyya, Thomas Lu, Arnav Mehta, Niels Voss, Narges Norouzi, and Gireeja Ranade. 2026. Leantutor: Towards a verified ai mathematical proof tutor. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 40670–40678.

Zaki Pauzi, Michael Dodman, and Manolis Mavrikis. 2025. Automating pedagogical evaluation of llm-based conversational agents. In *Ceur Workshop Proceedings*, volume 4006. CEUR.

Griffin Pitts, Anurata Prabha Hridi, and Arun Balajiee Lekshmi Narayanan. 2025a. A survey of llm-based applications in programming education: Balancing automation and human oversight. In *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+ NLP)*, pages 255–262.

Griffin Pitts, Viktoria Medvedeva Marcus, and Sanaz Motamedi. 2025b. Student perspectives on the benefits and risks of ai in education. In *2025 ASEE Annual Conference & Exposition*.

Griffin Pitts and Sanaz Motamedi. 2026. What drives students' use of ai chatbots? technology acceptance in conversational ai. *arXiv preprint arXiv:2602.20547*.

Griffin Pitts, Neha Rani, and Weedguet Mildort. 2026. Trust and reliance on ai in education: Ai literacy and need for cognition as moderators. *arXiv preprint arXiv:2604.01114*.

Griffin Pitts, Neha Rani, Weedguet Mildort, and Eva-Marie Cook. 2025c. Students' reliance on ai in higher education: identifying contributing factors. In *International Conference on Human-Computer Interaction*, pages 86–97. Springer.

Sujoy Roychowdhury, Sumit Soman, HG Ranjani, Neeraj Gunda, Vansh Chhabra, and SAI KRISHNA BALA. 2024. Evaluation of rag metrics for question answering in the telecom domain. In *ICML 2024 Workshop on Foundation Models in the Wild*.

Thomas Thesen and Soo Hwan Park. 2025. A generative ai teaching assistant for personalized learning in medical education. *NPJ Digital Medicine*, 8(1):627.