Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

RoboTuB: Retrieval-Augmented Tutoring for Adaptive Learning in STEM MOOCs

Avijit PANDEY[a], Gaurav MISHRA[b], Sunny Prakash PRAJAPATI[a] & Syaamantak DAS[a*]
[a]*Centre for Educational Technology, Indian Institute of Technology Bombay, India*
[b]*Department of Mechanical Engineering, Indian Institute of Technology Bombay, India*
*syaamantak.das@iitb.ac.in

Abstract: Educational AI chatbots have shown promise in enhancing learning, yet they often lack adaptive reasoning and pedagogical scaffolding for domain-specific subjects. This paper presents RoboTuB, a Retrieval-Augmented Generation (RAG) powered tutoring chatbot, designed to support team-based learning in STEM education. Unlike traditional question-answering (QA) systems, RoboTuB retrieves structured knowledge from domain-specific corpora, dynamically generating explanations, simulations, and code examples tailored to learners' needs. We evaluate RoboTuB in a two-phase study: (1) a usability assessment with novice learners using System Usability Scale (SUS) and NASA-TLX cognitive load metrics, and (2) a comparative analysis in a MOOC setting. Results from a two-phase evaluation showed that students using RoboTuB demonstrated substantial learning gains and reduced cognitive load compared to Q&A forums. These findings suggest that retrieval-augmented tutoring can provide meaningful support for multi-step problem-solving tasks. We discuss implications for NLP-driven educational chatbots and propose future directions for multimodal AI tutors.

Keywords: question answering, query processing, chatbot, Retrieval Augmented Generation (RAG)

1. Introduction

Large Language Models (LLMs) and retrieval-based architectures have led to significant advances in question-answering (QA) and dialogue systems. However, when applied to educational contexts, especially in STEM domains, they often fall short of delivering sustained, adaptive, and pedagogically grounded learning experiences. Learners require more than factual answers; they need scaffolded support that adapts to their evolving understanding, handles conceptual dependencies, and provides multimodal guidance (e.g., code, simulations, and theoretical context). In this work, we focus on multi-step learning tasks, tasks that require learners to sequentially navigate through multiple interdependent knowledge units. For instance, in robotics education, a student must first understand a physical system's dynamics, simulate it with appropriate models, and then implement a real-time control algorithm. Current QA systems often treat these subtasks in isolation, leading to fragmented or incoherent guidance.

To address this gap, we introduce RoboTuB (Robotics Tutoring Bot), a Retrieval-Augmented Generation (RAG) chatbot tailored to support learners in a MOOC setting focused on robotics and control systems. RoboTuB is designed to function as an AI-powered tutor, offering structured, context-aware explanations that evolve with the learner's progress. It integrates three pedagogically relevant modalities: theory (conceptual foundations), simulation (interactive experimentation), and code (implementation-level examples). For this study, we explore the following research questions:

RQ1: Can a retrieval-augmented LLM-based tutor improve learners' understanding and performance on multi-step tasks in a robotics MOOC?

RQ2: Does using such a system reduce cognitive load compared to traditional QA forums?

RQ3: How do learners perceive the usability and dialogue quality of an AI tutor designed with pedagogical scaffolding in mind?

For these research questions, we evaluated the effectiveness of the RoboTub using a two-phase study design with learners in lab-based and MOOC-based settings.

2. Related Work

Research on intelligent tutoring systems (ITS) and AI-driven educational technologies has grown substantially in recent years. While early ITS models focused on rule-based dialogue or domain-specific pipelines (VanLehn, 2011), recent advances in large language models (LLMs) and retrieval-augmented architectures (Lewis et al., 2020) have made scalable, adaptable AI tutors more feasible. However, translating these architectures into pedagogically effective systems remains a challenge. Prior work has highlighted the limitations of large language models in educational settings, including hallucination, a lack of alignment with curriculum goals, and difficulty sustaining scaffolding across multi-step reasoning (Kasneci et al., 2023). At the same time, emerging studies have begun exploring retrieval-augmented LLMs for dialogue tutoring and knowledge grounding in education (Henkel et al., 2024). Our work builds on this line of research by introducing a layered retrieval architecture (theory, simulation, code) and evaluating its educational impact in authentic learning environments.

2.1 Educational NLP and Scaffolding

Scaffolding refers to instructional strategies that provide learners with just-in-time support and fade assistance as learners gain competence (Wood, Bruner & Ross, 1976). In the context of AI tutors, this implies tailoring explanations, offering intermediate steps, and adjusting complexity dynamically. Prior works like (Winkler & Söllner, 2018; Stahl et al., 2024) emphasizes the importance of aligning chatbot responses with scaffolding principles to enhance conceptual understanding and promote cognitive engagement. Despite these advances, few systems offer retrieval-based scaffolding across multiple modalities such as theory, simulation, and code.

2.2 Retrieval-Augmented Generation for Educational Tasks

Retrieval-Augmented Generation (RAG) has shown promise in QA and summarization tasks by enhancing factuality and grounding generative outputs in external corpora (Lewis et al., 2020; Karpukhin et al., 2020). In education, RAG can reduce hallucinations and provide students with more domain-consistent explanations (Béchard & Ayala, 2024). However, most prior implementations focus on single-turn responses or factoid-style QA, which limit their pedagogical utility in supporting multi-step reasoning or personalized tutoring.

2.3 Dialogue Quality and Evaluation in Educational Chatbots

Evaluating AI tutors requires more than standard NLP metrics (e.g., BLEU or ROUGE). Usability frameworks such as SUS (Brooke, 1996) and workload metrics like NASA-TLX (Hart & Staveland, 1988) are essential to assess learners' cognitive and affective experiences. Additionally, the PARADISE framework (Walker et al., 1997) enables analysis of dialogue success based on task completion, efficiency, and user satisfaction. In their work, (Istrate, & Velea, 2024) argue that aligning chatbot evaluation with educational learning outcomes is crucial for real-world applicability, especially in MOOCs and self-paced learning environments.

2.4 Generative AI in STEM Education

The use of generative models in STEM instruction is on the rise, yet most systems either rely on static prompts or offer minimal interactivity. Khanmigo (Khan Academy, 2023) and Google's Socratic App provide lightweight tutoring experiences, but lack deep integration with structured pedagogical frameworks. Our work builds upon this gap by using a structured retrieval pipeline and LLM prompting strategy to simulate domain-specific scaffolding. RoboTuB distinguishes itself through its focus on multi-modal support and dialogic adaptability tailored to robotics education.

3. Methodology

3.1 System Architecture and Retrieval Pipeline

RoboTuB follows a modular RAG architecture. The core system consists of:

- Corpus Segmentation: The knowledge base is divided into three thematic layers: (1) theory (concepts, equations), (2) simulation (code snippets with explanations), and (3) programming (Python/C++ implementations). Each document is chunked into 100–250 token passages and embedded using MiniLM-L6 sentence-transformers.
- Indexing and Retrieval: Chunks are indexed using FAISS (Facebook AI Similarity Search) (Johnson and Jégou, 2019) for approximate nearest neighbor retrieval. For each query, top-5 passages from each layer are retrieved and passed through a reranker using cosine similarity with the dialogue context embedding.
- Context Construction: Retrieved texts are concatenated using a delimiter scheme and prepended to the system prompt. Summarized multi-turn history (using GPT-3.5) ensures conversational coherence.

3.2 Prompt Engineering and Generation

We use GPT-3.5 in a zero-shot setting with domain-adaptive prompts that dynamically insert context-relevant scaffolding instructions. Prompt templates were iteratively refined with the following goals:

- Encourage stepwise explanations (“Explain like a tutor to a novice”).
- Format responses using sections: Conceptual Overview, Simulation Walkthrough, Code Illustration.
- React to prior dialogue context (“Based on what we discussed before…”).

Example prompt prefix: “*You are a robotics tutor assisting a student working on PID control. Given the following materials and prior questions, provide a clear explanation including theory, a simulation example, and runnable code.*”

3.3 Evaluation Design and Controls

- **Phase 1** involved 19 graduate students using RoboTuB in a lab setting. SUS and NASA-TLX were administered after 30-minute tutoring sessions.
- **Phase 2** used data from two editions of a MOOC on control systems. Edition 1 (n = 92) used a traditional Q&A forum; Edition 2 (n = 91) used RoboTuB.
- Both editions followed an identical curriculum, grading, and teaching support. Prior academic background, declared programming experience, and control exposure were collected to ensure demographic balance.
- Data from platform logs, post-test scores, and chatbot interactions were anonymized and analyzed using Welch’s t-test.

It is important to note certain methodological limitations. First, while RoboTuB leverages a retrieval-augmented pipeline, our evaluation did not include an LLM-only baseline or a systematic ablation of retrieval layers and prompt templates. We chose to focus on the contrast between RoboTuB and the discussion forum baseline, as this reflects learners’ existing support environment in the MOOC. Future work will isolate the contribution of individual system components through controlled ablation studies. Second, while the system is adaptable to different learner queries, personalization in the present study was limited to prompt-template scaffolds (e.g., encouraging step-by-step reasoning). More advanced personalization techniques, such as learner modeling, adaptive difficulty, or long-term tracking, were outside the scope of this study, but remain promising directions.

4. Results & Analysis

4.1 RQ1: Performance on Multi-Step Tasks

To address RQ1, we analyzed learners' task accuracy and post-test performance across two MOOC cohorts. Edition 1 (n = 92) used a traditional discussion forum, while Edition 2 (n = 91) interacted with RoboTuB. Post-test scores revealed a 15.8% average increase in problem-solving accuracy for the RoboTuB group compared to the forum group. While this indicates a substantial improvement, we did not conduct fine-grained statistical subgroup analysis, and therefore interpret the result as an observed trend rather than a conclusive effect.

4.2 RQ2: Cognitive Load Comparison

To evaluate RQ2, we administered the NASA Task Load Index (NASA-TLX) after 30-minute tutoring sessions with 19 graduate students during Phase 1. Results indicated reductions in both mental demand (Mean = 34.8) and frustration (Mean = 31.5) compared to baseline scores from Q&A-based learning. Learners reported that RoboTuB's structured explanations and multimodal output helped reduce cognitive burden, particularly when transitioning between theory and code. Learners with lower perceived workload generally performed better in follow-up tasks, though this relationship was not statistically tested.

4.3 RQ3: Usability and Dialogue Quality

For RQ3, we assessed learner perceptions using the System Usability Scale (SUS) and qualitative feedback. RoboTuB received a mean SUS score of 76.58, indicating strong usability. Participants appreciated the multi-step explanations and adaptability of responses.
Manual inspection of conversations suggested that RoboTuB responses were more often complete and relevant compared to peer replies in the discussion forum. Learner feedback during SUS administration and open-ended reflections suggested satisfaction with the chatbot's clarity and instructional helpfulness. Based on an inspection of interaction logs, most learner problems appeared to be resolved within a few conversational turns (typically 2–4), although turn counts were not systematically recorded.

5. Discussion & Future Work

The findings from our evaluation suggest that RoboTuB can meaningfully enhance the learning experience for students engaged in complex STEM topics. By combining retrieval-based grounding with pedagogically aligned generation, the system improved usability, cognitive load, and learning outcomes. However, these contributions must be interpreted with consideration of several design constraints.

5.1 Pedagogical Implications

RoboTuB's effectiveness stems from its ability to simulate instructional scaffolding, drawing on retrieval to tailor explanations and context across multiple modalities. The observed improvements in performance and satisfaction, especially among novice learners, highlight the value of adaptive support in MOOCs. These results support recent literature advocating for dialogic, context-aware AI tutors that move beyond factoid-level QA. These observed improvements should be interpreted with caution, as subgroup analyses were not conducted.

5.2 Limitations

The system was evaluated within a specific MOOC focused on robotics and control systems, and the transferability to other subjects remains unverified. The study took place during the availability of GPT 3.5 in late 2023 to early 2024, without the opportunity to utilize advanced LLM models like GPT 4/Claude. Although no fine-tuning was applied, prompt engineering proved to be complex and may necessitate task-specific efforts to achieve generalization. Additionally, some responses tended to be overly verbose or misaligned with learners' intent, highlighting a need for real-time adaptation of responses.

5.3 Future Research Directions

Future work will explore integrating RoboTuB with adaptive learner models to enable more fine-grained personalization, as well as conducting controlled ablation studies to evaluate retrieval and prompting strategies in isolation. Richer qualitative analysis of learner interactions, and replication across multiple MOOCs, will further validate the system's generalizability. In conclusion, this study demonstrates that a retrieval-augmented tutoring system can enhance learners' performance, reduce perceived cognitive load, and provide a usable and pedagogically meaningful alternative to traditional Q&A forums in STEM MOOCs.

6. Conclusion

This paper introduced RoboTuB, a retrieval-augmented tutoring system that supports multi-step learning in a robotics MOOC. Through two-phase evaluation, we showed that RoboTuB improves problem-solving accuracy, reduces perceived workload, and enhances learner satisfaction. Our contributions include a modular RAG pipeline for multimodal tutoring, a validated prompt engineering framework, and evidence-based evaluation. While limitations exist around generalizability and verbosity, the system lays groundwork for future AI tutors that are pedagogically aware, contextually grounded, and student-adaptive. By refining how LLMs engage with domain knowledge, RoboTuB demonstrates the potential of NLP to deliver personalized, structured educational support in open-access environments like MOOCs.

References

Brooke, J. (1996). SUS: A quick and dirty usability scale. In P. W. Jordan, B. Thomas, & B. A. Weerdmeester (Eds.), Usability Evaluation in Industry (pp. 189–194). CRC Press.

Béchard, P., & Ayala, O. M. (2024). Reducing hallucination in structured outputs via Retrieval-Augmented Generation. arXiv preprint arXiv:2404.08189..

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Advances in Psychology, 52, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Henkel, O., Levonian, Z., Li, C., & Postle, M. (2024). Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference. In Proceedings of the 17th International Conference on Educational Data Mining (pp. 315-320).

Istrate, O., & Velea, S. (2024). Guidelines for online and blended learning: Design, delivery, assessment, evaluation of study programmes. Premises of academic curriculum digitalisation. September, 1-113.

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3), 535-547.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and individual differences, 103, 102274.

Karpukhin, V., Oguz, B., Min, S., Lewis, P. S., Wu, L., Edunov, S., ... & Yih, W. T. (2020, November). Dense Passage Retrieval for Open-Domain Question Answering. In EMNLP (1) (pp. 6769-6781).

Khan Academy. (2023). Khanmigo: AI-powered tutoring for learners and educators. Retrieved from https://www.khanacademy.org/khanmigo

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Küttler, H., ... & Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems (NeurIPS), 33, 9459–9474.

Stahl, M., Biermann, L., Nehring, A., & Wachsmuth, H. (2024). Exploring LLM prompting strategies for joint essay scoring and feedback generation. arXiv preprint arXiv:2404.15845.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

Walker, M. A., Litman, D. J., Kamm, C., & Abella, A. (1997). PARADISE: A framework for evaluating spoken dialogue agents. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL), 271–280. https://doi.org/10.3115/976909.979652

Winkler, R., & Söllner, M. (2018, July). Unleashing the potential of chatbots in education: A state-of-the-art analysis. In Academy of management proceedings (Vol. 2018, No. 1, p. 15903). Briarcliff Manor, NY 10510: Academy of Management.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem-solving. Journal of Child Psychology and Psychiatry, 17(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x