Developing LLM-Powered Trustworthy Agents for Personalized Learning Support

Benyamin Tabarsi

North Carolina State University
Raleigh, NC, 27695, USA
btaghiz@ncsu.edu

Abstract

Large Language Models (LLMs) have shown promise in educational applications, but challenges such as hallucinations, lack of contextual relevance, and limited personalization impede their practical adoption. To address these issues, my research introduces MerryQuery, an LLM-powered educational agent that integrates Retrieval-Augmented Generation (RAG), rule-based content control, and Reinforcement Learning from Human Feedback (RLHF). The system features a dynamic learning profile module for adaptive personalization and a multi-step verification framework that cross-checks responses against external sources to enhance trustworthiness. A functional prototype of MerryQuery is being piloted in a real-world classroom. Preliminary results demonstrate improved response reliability and student understanding.

Introduction

LLMs and AI-driven educational tools have the potential to transform learning by providing personalized support, automating repetitive tasks, and offering real-time feedback. However, integrating AI into educational settings is not without challenges. Concerns about misinformation, contextual inaccuracies, biases, and AI misalignment with educational goals have raised skepticism among educators and students. To create effective AI-powered systems, it is essential to prioritize trustworthiness in every aspect of their design and deployment. Moreover, personalization is critical, as each learner has unique needs and learning paths.

In educational contexts, these challenges are amplified as students and teachers rely on AI not just for answers but for guidance and support in developing complex skills. A trustworthy AI system must, therefore, go beyond generating correct information; it must ensure that its outputs are contextually relevant, aligned with pedagogical objectives, and free from biases. Moreover, personalization should be dynamic, continuously adjusting to each student's unique learning profile by utilizing interaction data and refining recommendations and learning strategies accordingly in real-time. Achieving these objectives requires advanced AI methodologies that combine techniques such as RAG (Lewis et al. 2020), multimodal learning, and adaptive response mechanisms.

To do so, my dissertation aims to answer three research questions.

1. **How can LLM-powered agents effectively deliver personalized learning experiences that adapt to individual students' needs and evolving educational goals?** I develop methods that leverage diverse student data to generate dynamic, individualized learning strategies.
2. **How should the trustworthiness of LLM-generated responses be enhanced, ensuring accuracy and alignment with learning objectives while mitigating hallucinations and biases?** I build robust validation and context-aware filtering mechanisms.
3. **How should adaptive AI methods, such as reinforcement learning, be utilized to optimize an agent's behavior to align better with student and teacher expectations?** I use continuous feedback loops to refine agent outputs and improve relevance over time.

Related Work

Previous research efforts and AI tools have shown promising results in education. Liffiton et al.'s *CodeHelp* offered on-demand coding assistance without full solutions, which was positively received by students (Liffiton et al. 2023). Liu et al. implemented a suite of RAG-based AI tools in CS50 for code explanation, code style improvement, and a chatbot for answering course-related questions, which saw frequent use and positive feedback from students (Liu et al. 2024). Hicke et al.'s *AI-TA* improved response quality by 30% using a combination of RAG, supervised fine-tuning (SFT), and Direct Preference Optimization (DPO) (Hicke et al. 2023).

MerryQuery builds on these successes but goes further by integrating techniques such as RAG for context-awareness, rule-based content control for filtering and verification, and RLHF for continuous adaptation, aiming to develop a personalized, adaptive, and trustworthy educational assistant.

Proposed Research

A high-level pipeline of MerryQuery is presented in Figure 1. MerryQuery's innovation is multifold. First, **multimodal data processing** and **dynamic vector database** allow the system to contextualize responses based on diverse student data (e.g., assignments, historical interactions) and course materials. Additionally, a **Memory Mechanism for**

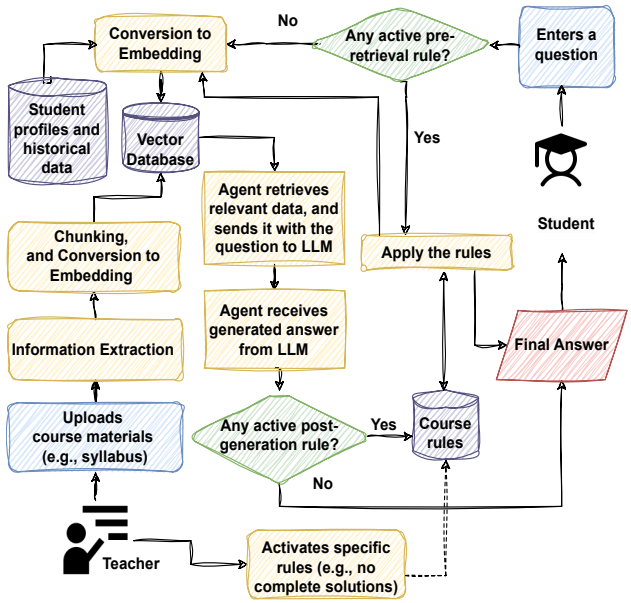

Figure 1: Pipeline of LLM-powered trustworthy agents.

Coherent Conversations ensures that the agent maintains context over interactions, making it a more interactive and personalized learning companion. This enables the agent to offer tailored feedback and learning strategies that evolve with the student's progress.

To establish **trustworthiness**, I incorporate a **Hallucination Prevention Mechanism** using RAG to ground its responses in verified data, combined with **rule-based controls** that allow instructors to customize response generation. It ensures that the agent's outputs are accurate and also aligned with course policies. Instructors can set pre-retrieval and post-generation rules, making the system highly flexible for various educational contexts.

Lastly, I apply **Reinforcement Learning from Human Feedback (RLHF)** and **Guided Learning Dialogue Design** to enhance adaptability. RLHF allows the system to refine its behavior based on real-time feedback from students and teachers, creating a continuous improvement loop. The **Guided Learning Dialogue** focuses on encouraging critical thinking by providing scaffolded hints and problem-solving strategies rather than direct answers. This enables an **adaptive agent** capable of adjusting its guidance level based on student engagement and progress.

Current Progress and Timeline

The development of MerryQuery, which I have led, has recently reached a major milestone with the release of a working prototype, and a detailed user guidance[1]. This prototype has already undergone a limited pilot study in the **CSC113 - Introduction to Computing (MATLAB)** course at NC State as part of a usability study aimed at gathering real student feedback and assessing the system's feasibility for broader deployment. Early results from both the pilot study and internal testing revealed that RAG implementation functions properly even with large datasets, handling complex PDFs accurately, and the memory mechanisms performing as expected, although certain areas benefit from refinements.

One of the major innovations in this prototype is the integration of **prompt engineering techniques** to enforce teacher-defined rules regarding the depth of answers. The implemented approach successfully prevented the disclosure of complete solutions without the need for our previously developed post-generation agent to verify compliance with teachers' policies.

Two studies are planned to follow the implementation of the aforementioned features. A controlled experiment at the end of November will be conducted to evaluate MerryQuery against a baseline LLM like ChatGPT, focusing on metrics such as student engagement, accuracy, and perceived trustworthiness. Additionally, by February 2025, an ablation study will be completed to assess the effectiveness of each implemented feature using course forum data with verified answers as ground truth.

Future Work and Plans

Future work will enhance personalization, trustworthiness, and adaptive learning capabilities, and evaluate their effectiveness. For personalization, I will develop a **dynamic learning profile system** and **context-aware multi-modal embeddings** to refine responses. Effectiveness will be measured through **longitudinal studies** tracking changes in individual learning outcomes, engagement, and user satisfaction using A/B testing against static recommendation baselines.

Further, I incorporate established educational theories, such as scaffolding, into MerryQuery. My previous work on Parsons's problems—where students solved jumbled code tasks—resulted in improved learning outcomes, including increased task completion speed and higher accuracy scores (Tabarsi et al. 2024). Exploring such strategies as alternatives to direct solutions will be one of my key focuses. These enhancements will be evaluated through **controlled classroom experiments** using metrics such as accuracy and trust scores from users.

References

Hicke, Y.; Agarwal, A.; Ma, Q.; and Denny, P. 2023. ChaTA: Towards an Intelligent Question-Answer Teaching Assistant using Open-Source LLMs. *arXiv:2311.02775*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS'20*, 33: 9459–9474.

Liffiton, M.; Sheese, B. E.; Savelka, J.; and Denny, P. 2023. Codehelp: Using large language models with guardrails for scalable support in programming classes. In *Koli'23*, 1–11.

Liu, R.; Zenke, C.; Liu, C.; Holmes, A.; Thornton, P.; and Malan, D. J. 2024. Teaching CS50 with AI: leveraging generative artificial intelligence in computer science education. In *SIGCSE'24*, 750–756.

Tabarsi, B.; Reichert, H.; Lytle, N.; Catete, V.; and Barnes, T. 2024. Scaffolding Novices: Analyzing When and How Parsons Problems Impact Novice Programming in an Integrated Science Assignment. In *ICER'24*, 42–54.

[1]https://exploremq.benyamintabarsi.com/