11th International Conference on Higher Education Advances (HEAd'25)
Universitat Politècnica de València, València, 2025
DOI: https://doi.org/10.4995/HEAd25.2025.20027

Case Study of Improving Educational Chatbots with Customized Information Retrieval

Linda Maria Antonia Becker[1], Nazim Ali[1], Thomas Spelten[2], Semih Severengiz[1]
[1]Sustainable Technologies Laboratory, Applied University of Bochum, Germany, [2]School of Medicine, Keele University, UK.

How to cite: Becker, L. M. A.; Ali, N.; Spelten, T.; Severengiz, S. (2025). Case Study of Improving Educational Chatbots with Customized Information Retrieval . In: 11th International Conference on Higher Education Advances (HEAd'25). Valencia, 17-20 June 2025. https://doi.org/10.4995/HEAd25.2025.20027

Abstract

Generic chatbots can function as virtual tutors, assessing responses and providing feedback. However, trained on generalized data limits them to act as subject-specialist tutors, resulting in weak or inaccurate responses and feedback, particularly for complex queries. We aimed to develop and evaluate whether a specialised information retrieval system can enhance a generic chatbot's capabilities. The customized chatbot was integrated into a teaching session and evaluated by the students. Our approach enabled the chatbot to access relevant information and provide tailored feedback, resulting in 75% of students finding the teaching session engaging and helpful. Students also expressed a preference for the customized chatbot over a generic one. However, occasional inaccuracies in responses to highly complex queries occurred and were attributed to incomplete optimization of the retrieval system. To address this, we recommend developing a comprehensive knowledge database and fully optimising the retrieval system to ensure consistently accurate and contextually relevant responses.

Keywords: *AI; chatbots; customization; specialised; feedback; RAG*

1. Introduction

Studies on the use of AI chatbots are growing at an ever-increasing rate. For instance, a search on Google Scholar returned 102,000 results for 2024 alone, representing a significant increase of over 110% from the 48,400 results for 2023. These numbers highlight the growing interest in using chatbots, such as ChatGPT, Gemini, and Copilot, commercially available from OpenAI, Google and Microsoft respectively, for a wide range of applications including teaching and learning practices. While many educational institutions encourage the use of chatbots to enhance learning, concerns related to inaccurate responses and data privacy are persistently flagged by studies (Ansari *et al.,* 2024; Labadze *et al.,* 2023). These concerns emphasize the critical need

for solutions that prioritize accuracy in chatbot responses and ensure robust data security measures to foster student trust and engagement.

Despite the widespread adoption of chatbots, there is a critical gap in research on practical strategies to address these issues, as well as limited understanding of students' perspectives on these concerns and how they impact their engagement with these tools for learning. In this study, we address these issues by investigating whether we can enhance the capability of a generic chatbot to access our own specific knowledge base to provide contextually accurate responses.

2. Approaches to enhance chatbot capabilities

Chatbot development typically begins with the use of a foundation model (FM). Foundation models are pre-trained on a vast body of generalized data, equipping them with a broad knowledge base and the ability to generate coherent responses. To further enhance the capabilities of these models, Reinforcement Learning (RL) is often employed. In this process, human evaluators review the chatbot's responses, providing feedback on critical factors such as accuracy, relevance, and fairness. The chatbot incorporates this feedback into its learning process, enabling it to adjust its parameters and prioritize outputs that better align with the desired criteria. This iterative feedback loop allows the chatbot to continuously evolve and has been reported to improve its ability to deliver accurate and user-aligned responses (Izadi and Forouzanfar 2024). However, retraining foundation models through RL for domain-specific or organizational applications is computationally and financially intensive. An alternative approach is the Retrieval-Augmented Generation (RAG) model, which bypasses the need for extensive retraining. Instead, RAG combines a generic LLM with a retrieval system, enabling it to access a specifically developed database containing domain-specific information (Gao *et al.,* 2023). We reason that the technique could be applied in teaching and learning to empower chatbots to handle complex tasks and provide contextually relevant responses on specific subjects. Several recent studies have explored the application of RAG in different contexts. Saha et al. (2024) showed that standard RAG configurations outperform traditional LLM outputs in terms of factual precision and contextual alignment in question-answering tasks. Modran et al. (2024) developed a RAG-based tutoring system and highlighted the anticipated benefits of the architecture in supporting domain-specific learning. However, as their work remains in preliminary stages, insights from the learner’s perspective are yet to be published. Recognising the importance of student-centred evaluation, this study presents a practical implementation of a RAG-enhanced chatbot in a live educational setting. The system was embedded with curriculum-specific content and its effectiveness assessed through student feedback and comparative analysis.

3. Methodology

3.1. Development of a local chatbot with RAG capabilities and its workflow

Building on previous work demonstrating that chatbots can deliver feedback aligned with predefined assessment criteria, this study introduced a teaching session where students received tailored feedback from a RAG-enhanced chatbot. For the chatbot a small LLM called discolm-mfto-7b-german-v0.1 was chosen due to it needing reduced computing power while providing robust performance in German. (discolm-mfto-german, 2024). The chatbot was integrated with RAG, allowing it to retrieve task-specific data such as student tasks, sample solutions, and relevant datapoints. Supplemented by a tailored prompt describing assessment criteria, this setup enabled the chatbot to deliver a more reliable, targeted feedback aligned to specific exercises (Woo et al., 2024). The chatbot followed a multi-step workflow: first, it received the input, then it queried the retrieval system for relevant context and finally generated feedback by combining this retrieved information with the instructional prompt.

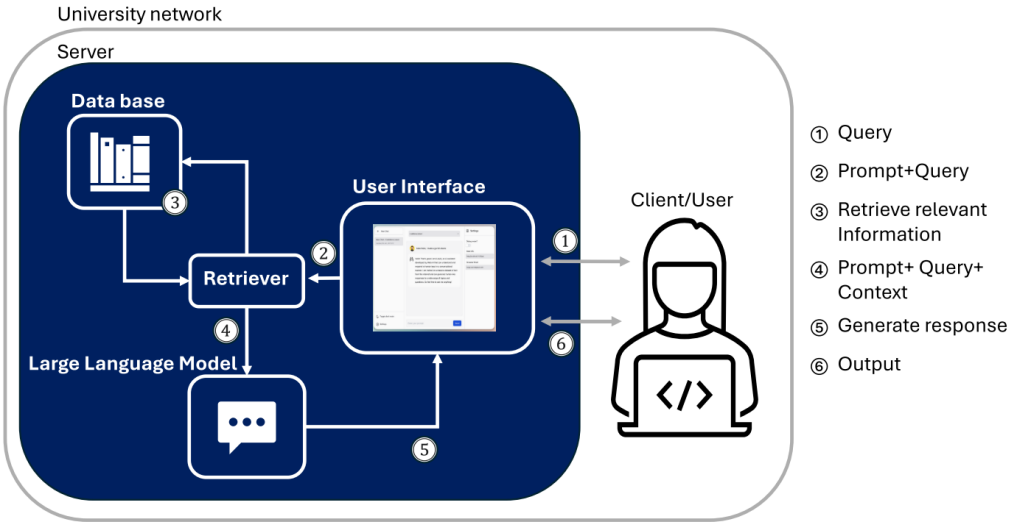

Figure 1. The implementation of the RAG model. A locally situated chatbot was combined with a retrieval system enabling it to access a highly selective database to provide domain specific responses.

To enhance accuracy and contextual relevance, the temperature was set to 0 to reduce variability, and the Top-K parameter was lowered to 5 to limit retrieved text sections (Chang et al., 2023; Yu, 2022). Further details on the setup, hardware, and software can be found on the Bochum University's Sustainable Technologies Laboratory 's documentation website.

3.2. Design of the teaching session integrated with the RAG powered chatbot

To evaluate the performance of the custom-built chatbot as a virtual tutor, it was integrated into a 90-minute trial teaching session on Ecodesign, delivered in December 2024 at Bochum University to undergraduate students enrolled in the Sustainable Development bachelor's degree. The design of the session involved students completing a practical exercise drawn from the Ecodesign lecture. The exercise required them to calculate which life cycle phase: production, transport, or use had the highest global warming potential (GWP), using data provided on an e-scooter and reference values for the GWP of various materials and processes. Based on their calculations, students were then asked to make practical design recommendations to reduce environmental impact in line with Ecodesign principles. Run as a cohort-wide activity, providing individualized support and feedback to each group's recommendations during the session has previously been impractical due to time constraints and the scale of the class. The integration of a local chatbot tutor powered by RAG was proposed as a solution to address these challenges as students would be able to complete the tasks while receiving real-time, constructive feedback tailored to their specific input.

3.3. Evaluation instrument and analysis

We developed a structured questionnaire designed to systematically collect both quantitative and qualitative data. The instrument was informed by established principles of educational evaluation (Kelley *et al,* 2003) and reviewed by academic staff to ensure validity and reliability. A mix of Likert-scale, ranking, and open-ended questions allowed for a well-rounded and triangulated analysis of user experience. The questionnaire covered key areas relevant to chatbot effectiveness, including performance metrics and student preferences as shown in Table 1.

Table 1. Survey insights: Areas of interest and examples of metrics which were evaluated

Area of interest	**Examples**
Ranking of requirements	accuracy, speed, data security
Evaluation of the trail session	helpfulness, personal impressions
Comparison with other feedback methods	commercial offers, sample solutions

The study was designed and conducted in accordance with the university's Educational Research standards. All students were fully informed of the study's purpose and assured that participation was voluntary, thereby enabling informed consent. To protect student confidentiality, no personal data was collected during the evaluation. Although all students (n = 30) were invited to complete the questionnaire, 12 responses were received. While this number does not support statistical significance, a 40% response rate exceeds the widely accepted

minimum threshold of 30% for online surveys in educational settings (Nulty, 2008) and was therefore considered acceptable for analysing the results and identifying trends.

4. Results and discussion

4.1. Identifying students' requirements for engagement with a learning chatbot

One of the first findings of this study focused on identifying students' expectations and priorities regarding AI-driven educational tools. Research shows that students may hold ethical concerns or skepticism about the reliability of AI-generated feedback, which can impact their engagement (Schei et al., 2024). To inform the design of a locally situated chatbot, students were asked to rate the importance of three key functionalities: accuracy, response speed, and data security.

Findings showed that accuracy was the highest priority, with all students rating it as "very important." Speed was also valued, with 75% rating it as "important" or "very important." In contrast, opinions on data security were split evenly. These findings are consistent with broader research (Schei et al., 2024).

4.2. Evaluation of chatbot functionality and impact on student learning

The chatbot, customized with RAG-enabled parameters, was integrated into the Ecodesign session to assess students' solutions and provide feedback. This evaluation followed an Exploratory Testing approach as defined within the international software testing standard (ISTQB, n.d.). Its accuracy was evaluated by comparing its responses to the worked model solutions, and it was found to correctly identify and report errors in students' submissions. As part of the test session it was observed that the chatbot hallucinated content—particularly in its explanations of errors when responding to incorrect student submissions—and that its mathematical calculations were unreliable. While our setup ensured data security, as all information entered into the chatbot was stored within the local server, its relatively slow processing speed in comparison to commercially available chatbot like ChatGPT was noted by students which may have impacted their overall user experience. Table 2 provides a summary of the chatbot's performance across these key functionalities mapped against their priority as perceived by students.

We then sought to understand students' perceptions of the chatbot's feedback compared to other forms of support, including sample solutions, tutorial sessions, and commercially available chatbots. While we acknowledge the limitations of retrospective evaluations, the responses clearly indicated a preference for using the chatbot over receiving no feedback, suggesting it added value to the learning process. However, traditional methods such as written feedback with opportunities for clarification and tutor-led tutorials remained the most preferred. This suggests

that while chatbot feedback is beneficial, it is best positioned as a complement rather than a replacement for human-guided support (Figure 2).

Table 2. Summary of the evaluation of the chatbot tutor based on the student requirements

Functionality	Relevance	Result	Further details of result
Accuracy	very important	partly achieved	+ Feedback finds relevant mistakes within the students' solution. - Feedback contains hallucinated content. - Mathematical operations within the feedback are not reliable.
Speed	important	not achieved	- Response time of more than 5 seconds, sometimes up to several minutes - Unable to process several requests in parallel
Data Security	partly important	achieved	+ Entered data is stored on the local server and can only be viewed by the application administrator.

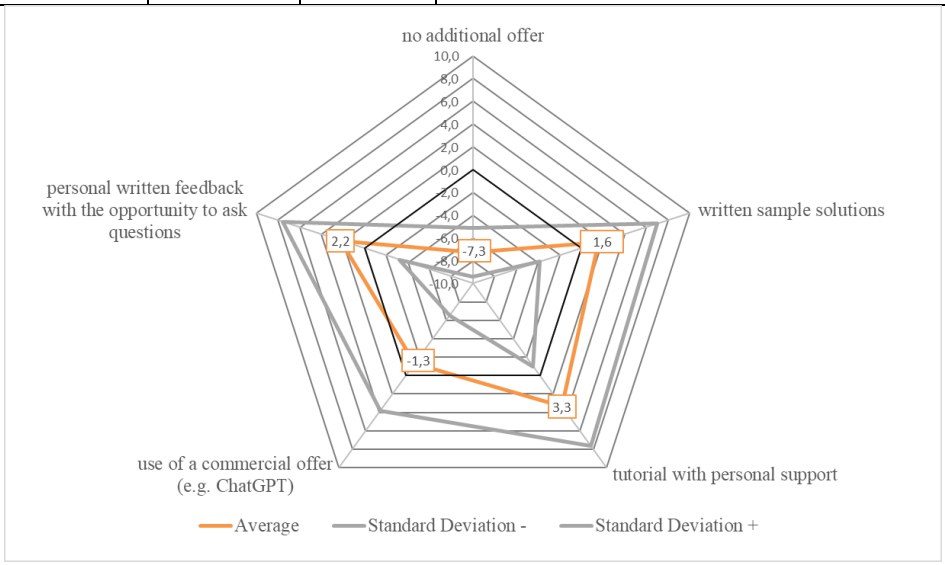

Figure 2. Students' perception of the bespoke chatbot in comparison to other feedback methods. Figure shows how the students ranked their preference on average on a scale from -10 (preferred chatbot) to 10 (preferred alternative method).

5. Summary and Outlook

This study developed and evaluated a German-language chatbot tutor using open-source frameworks Ollama and OpenWebUI, tailored to help students identify weaknesses in their learning and provide domain-specific feedback in an Ecodesign course. By integrating RAG, the chatbot effectively helped students identify errors in their submitted work.

Pedagogically, we believe there are several implications of this approach on student learning. Firstly, this approach fosters self-regulated learning, which can enhance student engagement. By enabling students to assess their own understanding, they gain greater ownership over their learning leading to increased academic performance. Research indicates that self-regulated learning helps students gain confidence and develop critical thinking skills which contribute to improved academic outcomes (Broadbent *et al.,* 2021). This approach also helps to make learning an active process. The provision of immediate, context-specific feedback supports learners in actively constructing their understanding, aligning with principles of constructivist learning theory. The approach of providing instant assessment and feedback aims to help students identify knowledge gaps and adapt their strategies accordingly. This form of metacognitive regulation is associated with increased motivation and deeper learning engagement (Nicol & Macfarlane-Dick, 2006). However, we acknowledge that our current implementation did not fully meet these objectives. The chatbot's delayed response time and hallucinated content limited the overall effectiveness of the feedback. We attribute this to two main factors: insufficient curation of domain-specific content and the limited computational optimization of our deployment environment.

To address these limitations and enhance reliability, we propose the following improvements:

1. Improved Accuracy: Enhancing RAG capabilities through curated domain-specific materials and refining both the retriever and the language model.
2. Performance Optimization: Upgrading to a GPU with at least 28GB and implementing parallel model instances or efficient request handling to support multiple users simultaneously.
3. Comprehensive Evaluation: Conducting additional tests, including functional and stress testing, and collaborative evaluations with students to ensure usability and performance.

In summary, the proposed improvements in performance, accuracy, and evaluation will help to refine the development of GenAI-driven educational tools.

References

All-MiniLM-L6-v2 (Version all-minilm-l6-v2-f32:latest). (2024). [Software]. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Ansari, A.N., Ahmad, S. & Bhutta, S.M. (2024). Mapping the global evidence around the use of ChatGPT in higher education. A systematic scoping review. *Educ Inf Technol* 29, 11281–11321. https://doi.org/10.1007/s10639-023-12223-4

Broadbent, J., Sharman, S., Panadero, E., Fuller-Tyszkiewicz, M. (2021). How does self-regulated learning influence formative assessment and summative grade? Comparing online and blended learners, The Internet and Higher Education, 50. https://doi.org/10.1016/j.iheduc.2021.100805

Chang, C.-C., Reitter, D., Aksitov, R., & Sung, Y.-H. (2023). KL-Divergence Guided Temperature Sampling (arXiv:2306.01286). https://doi.org/10.48550/arXiv.2306.01286

Discolm-mfto-german. (2024). [Software]. https://ollama.com/cas/discolm-mfto-german

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J. and Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:23*

International Software Testing Qualifications Board (ISTQB). (n.d.). *Glossary*. Retrieved April 22, 2025, from https://istqb-glossary.page/exploratory-testing/

Izadi, S., & Forouzanfar, M. (2024). Error Correction and Adaptation in Conversational AI: A Review of Techniques and Applications in Chatbots. *AI*, 5(2), 803-841. https://doi.org/10.3390/ai5020041

Kelley, K., Clark, B., Brown, V. and Sitzia, J. (2003) 'Good practice in the conduct and reporting of survey research.' International Journal for Quality in Health Care. 15(3), 261-266. Available at: https://doi.org/10.1093/intqhc/mzg031

Labadze, L., Grigolia, M. & Machaidze, L. (2023). Role of AI chatbots in education: systematic literature review. *Int J Educ Technol High Educ* 20, 56. https://doi.org/10.1186/s41239-023-00426-1

Modran, H. A., Ursuțiu, D., Samoilă, C., & Gherman-Dolhăscu, E.-C. (2024). Developing a GPT Chatbot Model for Students Programming Education. In M. E. Auer, R. Langmann, D. May, & K. Roos (Eds.), Smart Technologies for a Sustainable Future (pp. 72–82). Springer. https://doi.org/10.1007/978-3-031-61905-2_8

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: what can be done? *Assessment & Evaluation in Higher Education*, *33*(3), 301–314

Saha, B., Saha, U., Zubair Malik, M. (2025). QuIM-RAG: Advancing Retrieval-Augmented Generation with Inverted Question Matching for Enhanced QA Performance. *arXiv e-prints*. doi:10.48550/arXiv.2501.02702

Schei, O. M., Møgelvang, A., & Ludvigsen, K. (2024). Perceptions and Use of AI Chatbots among Students in Higher Education: A Scoping Review of Empirical Studies. *Education Sciences*,*14*(8), 922. https://doi.org/10.3390/educsci14080922

Woo, J., Yang, A., Olsen, R., Hasan, S., Nawabi, D., Nwachukwu, B., Williams, R and . Ramkumar, P. (2024). Custom Large Language Models Improve Accuracy: Comparing Retrieval Augmented Generation and Artificial Intelligence Agents to Noncustom Models for Evidence-Based Medicine. *Arthroscopy: The Journal of Arthroscopic & Related Surgery.* https://doi.org/10.1016

Yu, W. (2022). *Retrieval-augmented Generation across Heterogeneous Knowledge.* Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, 52–58. https://doi.org/10.18653/v1/2022.naacl-srw.7