Learning with GenAI before the exam:

A pilot study on the use of RAG-based chatbots in a Cognitive Psychology course

Report 3

Manuel Oliveira, Rianne Conijn

Human Technology Interaction Group

Department of Industrial Engineering & Innovation Sciences, TU/e

Project: Scaffolding writing skills using automated essay generation systems

Team: Manuel Oliveira, Rianne Conijn, Carlos Zednik, Bert Sadowski, Gunter Bombaerts

Funding: 4TU.CEE & BOOST!

Eindhoven, 24 October, 2025

Contents

Executive Summary

Context

Generative AI (GenAI) chatbots, especially Retrieval-Augmented Generation (RAG) models that utilize course-specific materials, are being introduced into university courses. Understanding how students naturally engage with these tools and their impact on learning is necessary.

Study Overview

This report presents findings from an observational pilot study in a Bachelor-level Cognitive Psychology course (N=116 students took the exam). Students were given voluntary access to two RAG chatbots (Tilburg.ai and Alexandria.cx) for three weeks before their final exam. The study analyzed chatbot interaction logs, final exam grades, and post-exam survey responses (N=19 completed) to understand usage patterns, student perceptions, and correlations with academic performance. Access was partially disrupted by a university cyberattack.

Main findings:

1. **Usage patterns:** Chatbot use was voluntary and varied significantly among students. Of the 116 students, 35 used Alexandria.cx (data for Tilburg.ai was incomplete). Usage concentrated heavily in the last four days before the exam (85% of interactions), with students primarily asking for explanations of course topics (65% of questions). Advanced features were rarely used.
2. **Student perceptions:** Students who completed the survey reported relatively high AI literacy and generally positive perceptions of the chatbots regarding ease of use, usefulness, and quality of answers, although the sample size for perception data was small.
3. **Learning outcomes:** Students who used the Alexandria.cx chatbot had slightly higher average exam grades than non-users, but this difference was not statistically significant ($p = .068$, Cohen's $d = 0.34$). Exploratory analyses showed that neither the frequency of interaction nor specific types of interaction (like generating practice questions) significantly predicted exam grades among users.

Interpretation

The findings suggest that when RAG chatbots are offered as optional, unscaffolded tools, students tend to use them primarily for last-minute review and clarification. This pattern of superficial engagement did not correlate with significantly better exam performance in this pilot context. The lack of a clear link between usage and grades implies that simply accessing the tool, or accessing it more frequently, is insufficient to improve learning outcomes. Deeper, more "agentic" engagement, potentially facilitated by structured pedagogical approaches, might be necessary.

Limitations

The study was observational with self-selected participants, limiting causal claims. The survey sample size was small, and a cyberattack disrupted access. Findings are preliminary and specific to this course context.

Implications

The mere availability of RAG chatbots does not automatically enhance learning. Educational institutions and instructors should consider implementing structured guidance or specific activities that encourage students to use these tools more proactively and critically throughout the learning process. Teaching AI literacy and promoting student agency in interacting with these tools appear crucial for maximizing their educational potential.

Conclusion

RAG chatbots show potential as learning aids and are perceived positively by students, but effective integration requires careful pedagogical design to encourage deeper engagement beyond superficial, last-minute use.

Abstract

The integration of Generative AI (GenAI) chatbots, particularly Retrieval-Augmented Generation (RAG) models grounded in course materials, is increasing in higher education. This report details an observational pilot study conducted in a university-level Cognitive Psychology course to investigate voluntary student interaction with two RAG chatbots (Tilburg.ai and Alexandria.cx). We examined usage patterns via interaction logs, student perceptions through a post-exam survey (N = 19), and correlations between chatbot use (N = 35 for Alexandria.cx) and final exam grades (N = 116). Results indicated generally positive student perceptions of the chatbots. Usage was highly variable and concentrated in the days immediately preceding the final exam, primarily involving requests for topic explanations. Comparison of final exam grades revealed no statistically significant difference between students who used the Alexandria.cx chatbot and those who did not ($p = .068$). Furthermore, exploratory analyses found no correlation between specific usage metrics (e.g., interaction frequency, asking for practice questions) and exam performance among users. These preliminary findings suggest that the mere availability of RAG chatbots, used voluntarily and without specific pedagogical scaffolding, may not translate directly into improved academic outcomes, highlighting the potential need for structured integration strategies to foster deeper learning engagement.

Introduction

Chatbots powered by generative artificial intelligence (GenAI) are increasingly integrated into educational settings, offering students instant, personalized support. With the rise of large language models (LLMs), their capabilities have expanded significantly, enabling new forms of interaction with course content. However, concerns remain regarding academic integrity, potential overreliance, fairness, and the actual impact on student learning outcomes (Kasneci et al., 2023; Memarian & Doleck, 2023). Recent work also highlights the importance of how AI tools are integrated; structured, scaffolded use may enhance student agency and confidence, whereas unstructured use might increase anxiety (Smirnova, 2025).

A specific application that has been increasingly gaining traction in applications of GenAI technology to diverse use cases is the Retrieval-Augmented Generation (RAG) chatbot. Put simply, RAG is a technique that enhances GenAI models by allowing them to first retrieve information from specific, relevant data sources before providing an answer (Lewis et al., 2020). This approach can be used to ground a chatbot's responses in a curated set of documents, such as course-specific materials, to ultimately provide more accurate and contextually relevant support, at least when compared to GenAI-powered chatbots relying on general-purpose LLMs. Despite the potential of RAG-based chatbots, it remains unclear how students voluntarily engage with these tools when offered as a supplementary resource.

In this report, we describe the results of an observational pilot study investigating how students interact with two RAG chatbots (viz. Tilburg.ai and Alexandria.cx) in a university-level Cognitive Psychology course. Our aim was to understand patterns of use, student perceptions of the tools, and whether voluntary use correlates with learning outcomes, measured via final exam grades. This study provides preliminary insights into student engagement with course-specific AI tools in a naturalistic setting.

Background

The integration of AI-powered chatbots into higher education has evolved, progressing from simple question-answering systems to more sophisticated generative models. Standard LLMs generate responses based on vast, static training data (Meyer et al., 2023). RAG systems address the limitation of contextual specificity by dynamically retrieving information from a curated knowledge base (Jeong, 2023; Maryamah et al., 2024). This allows RAG chatbots to function as course-specific learning aids, aiming for greater accuracy and relevance compared to general LLMs (Parekh et al., 2025).

Research examining the educational impact of RAG chatbots suggests they can enhance student learning experiences. Studies indicate that course-specific RAG chatbots can provide personalized guidance and contextually relevant information, thereby positively affecting student engagement and exam preparedness (Thway et al., 2024). RAG systems can offer tailored assistance that might help improve student understanding (Modran et al., 2024). For instance, some students have reported satisfaction with RAG chatbots for providing targeted explanations aligned with course content (Lang & Gürpinar, 2025). Furthermore, practical applications show potential for RAG chatbots to support students in complex tasks, such as data analysis, by offering on-demand assistance (Zhang, 2025).

Beyond direct student support, RAG chatbots are also being explored for institutional roles, such as streamlining administrative processes (Dharshan S et al., 2025; Nisanth, 2025) or providing cost-effective educational support (e.g., well-optimized open-source RAG

implementations; Kizi & Suh, 2025). However, faculty perspectives highlight remaining challenges, including technical limitations and ethical concerns, such as the potential for AI hallucinations (i.e., producing incorrect information), which require careful consideration before widespread deployment (e.g., Dakshit, 2024). Broader meta-analyses on educational chatbots also clarify the existence of implementation difficulties alongside potential benefits (Labadze et al., 2023).

Although the potential benefits of RAG chatbots are becoming evident, the effectiveness of their integration depends ultimately on *how* students engage with them. Recent work emphasizes the importance of student agency (i.e., capacity to regulate and take ownership of learning) in distinguishing between superficial use and deeper cognitive engagement with AI tools (e.g., Oliveira et al., 2025; Yang et al., 2024). Pedagogical approaches also matter as unstructured access to AI might differ from a more structured, scaffolded use of the technology. Smirnova (2025), for example, found that scaffolding AI interactions enhanced student agency and confidence in an academic writing context, whereas unstructured use of AI was sometimes linked to higher student anxiety.

Despite growing interest and the reported positive preliminary findings, a gap still remains in our understanding of how students voluntarily use course-specific RAG chatbots for exam preparation in a naturalistic, unscaffolded setting, and how this usage relates to objective academic performance (i.e., grades). The current pilot study addresses this gap by observing voluntary usage patterns of two different RAG chatbots and examining correlations with final exam grades.

Methodology

Design

An observational pilot study with a within-subject component was used to evaluate the use and impact of two chatbots. Students in a Bachelor course on Cognitive Psychology were given access to both tools three weeks before the final exam. All interactions with the chatbots were logged, and students' final exam grades were collected. After the exam, students were asked to complete a survey assessing general AI literacy and their perceptions regarding the chatbots.

Participants

A total of 140 students enrolled in the Cognitive Psychology course. Of these, 116 students completed the final exam. Usage data showed that 35 students used the Alexandria.cx chatbot. Data regarding the exact number of Tilburg.ai users or those using both chatbots were incomplete in the available logs. A total of 21 students began the post-exam survey, and 19 students completed it. Participation in the survey was voluntary.

Procedure & Materials

Three weeks (21 days) prior to the final course exam, all students were granted access to the Tilburg.ai and Alexandria.cx chatbots through the Canvas Learning Management System (LMS). Both chatbots employed RAG, drawing information from course materials provided via Canvas. Access was disrupted by a university-wide cyberattack, and consequently, the Canvas LMS and the Tilburg.ai chatbot were inaccessible from day 3 through day 10 of the access period. The Alexandria.cx chatbot remained accessible. Following the exam, all

students were invited to complete the survey. A screenshot of the Tilburg.ai interface is shown in Figure 1, and a screenshot of the Alexandria.cx interface is displayed in Figure 2.

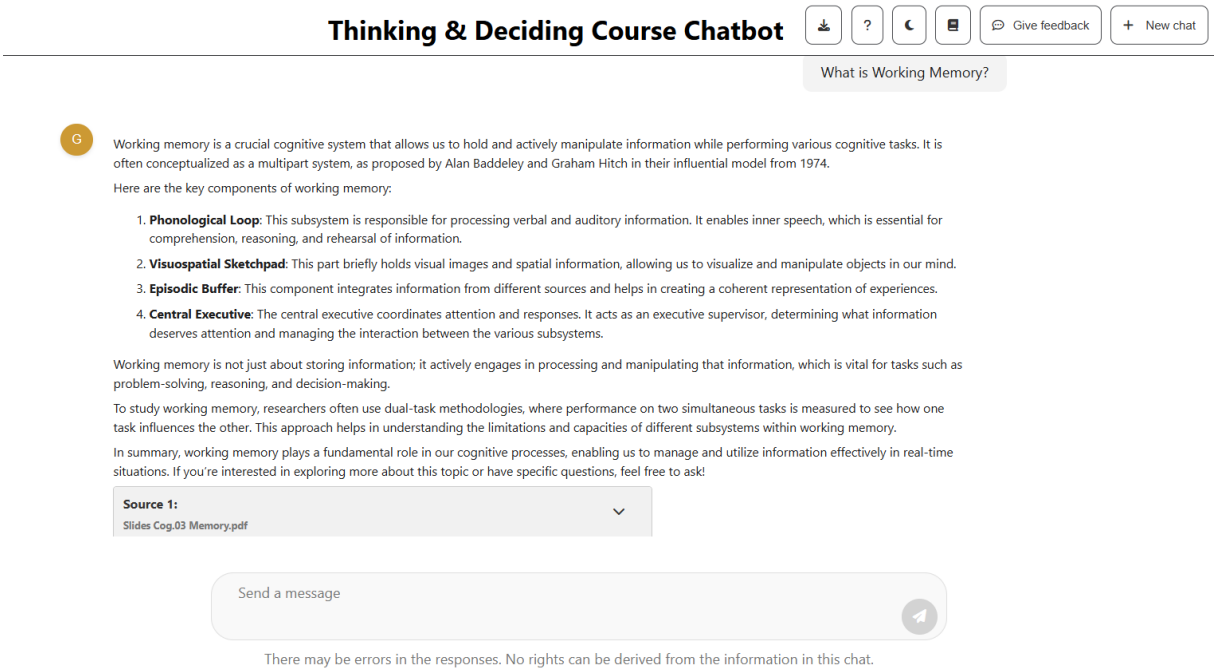

Figure 1. Screenshot of the interface of Tilburg.ai chatbot

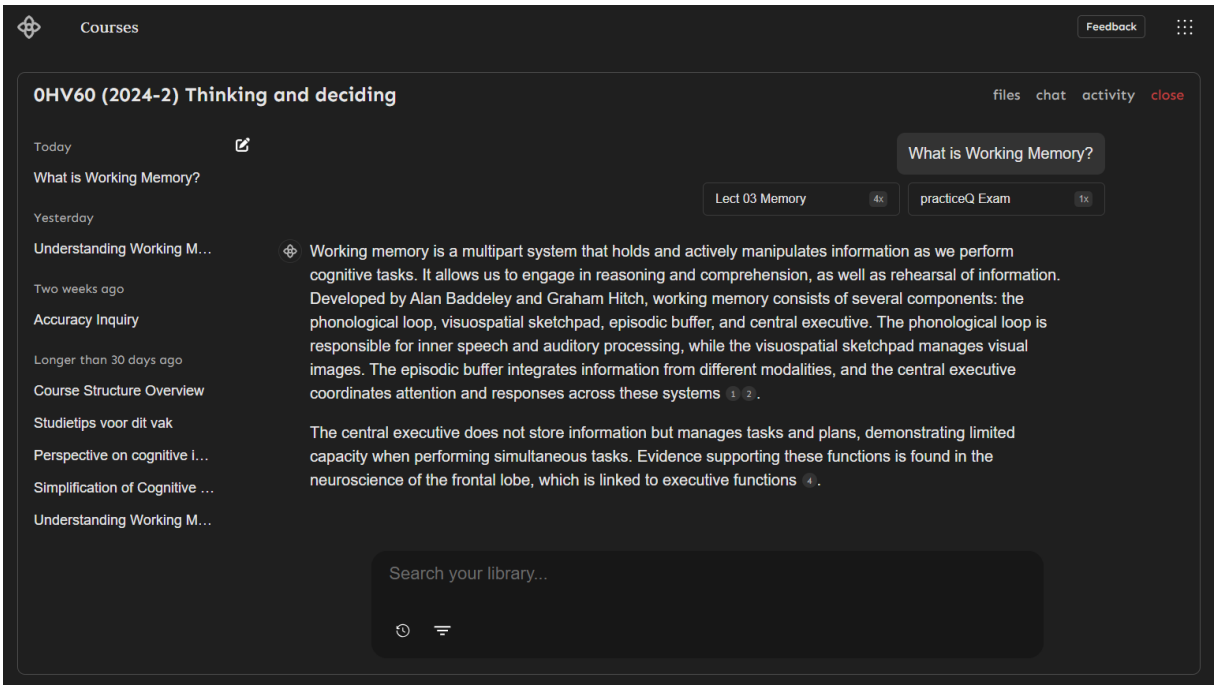

Figure 2. Screenshot of the interface of Alexandria.cx chatbot

Measures

Chatbot Interactions

Interaction data included timestamped questions asked by students and the corresponding chatbot answers. A user identifier linked chatbot data to exam grades and survey responses. Questions were coded by the course teacher into thematic categories (e.g., explaining topics, creating practice questions).

Answer Quality

The questions asked to the chatbot were coded by the course teacher into themes including: general greetings, gaining insights into capabilities, explain topic, create practice questions, course organization questions, summarize files, rephrasing questions to improve output, and other (empty and non-English questions).

Final Exam Grade

The final exam consisted of both multiple-choice and open questions. Grades ranged from 1 to 10, with 5.5 representing a passing score.

Survey

AI literacy was measured using the apply-AI subscale of AI-literacy scale (Ng et al., 2021). The scale consisted of six items (e.g., “I can operate generative AI applications in everyday life.”). The reliability of the scale was high (Cronbach’s $\alpha = 0.94$, $N = 21$). The perceptions of the chatbots included ease of use, usability, and perceived quality, which were measured for both chatbots separately, and only to the students who interacted with the corresponding chatbot(s). Ease of use was adapted from the scale by Davis (1989). The scale consisted of six items (e.g., “Learning to operate the chatbot would be easy for me.”). The reliability of the scale was high (Cronbach’s $\alpha = 0.93$, $N = 15$). Usability was also adapted from the scale by Davis (1989). The scale consisted of six items (e.g., “Using the chatbot enabled me to accomplish learning in this course more quickly.”). The reliability of the scale was high (Cronbach’s $\alpha = 0.98$, $N = 15$). Finally, perceived quality was adapted from the perceived recommender quality scale (Knijnenburg et al., 2012). The scale consisted of six items (e.g., “I liked the answers by the chatbot.”). The reliability of the scale was high (Cronbach’s $\alpha = 0.91$, $N = 15$).

Four open questions were used to get some more insights into the use and perceptions of the chatbot: “Why / with what goal did you use the chatbot?”; “For what goals / which questions did you feel the chatbot worked well?”; “For what goals / which questions did you feel the Alexandria.cx chatbot did not work well?”; “Do you have any suggestions for improving the chatbot?”

Results

General Use of Chatbots

This analysis focused on the 35 users of Alexandria.cx due to incomplete logs for the other tool. The number of conversations per student varied considerably (M = 10.0, SD = 10.7). Conversations were typically short (M = 3.5 questions, SD = 2.4). Usage peaked in the final four days preceding the exam, accounting for 85% of total interactions. Thematic analysis of the 1274 questions showed that most (65%) were requests for explaining specific course topics, followed by requests to create practice questions (13%).

Perceptions of Chatbots

Survey respondents (N = 19 completed) indicated relatively high general AI literacy (M = 3.6 on a 5-point scale, SD = 1.0). Most reported using generative AI tools regularly. Students who used Alexandria.cx (n = 12 providing ratings) generally held positive perceptions regarding its ease of use, usefulness, and the quality of its answers. The perceptions were slightly more negative for the Alexandria.cx chatbot (see Figure 3), however only 3 students provided insights on their perceptions of the Tilburg.ai chatbot, which does not allow for any reliable conclusions to be drawn from the current data.

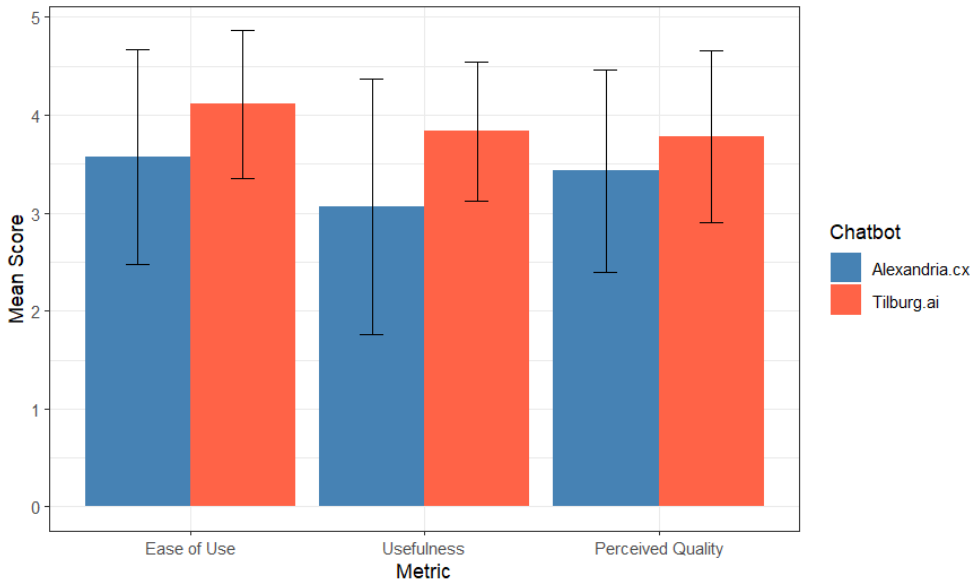

Figure 3. Perceptions of the two chatbots Alexandria.cx (*n* = 12), Tilburg.ai (*n* = 3).

Effects on Learning

An independent samples t-test compared the final exam grades of students who used the Alexandria.cx chatbot (n = 35) with those who did not (n = 81). Students who used the chatbot achieved slightly higher scores (M = 5.68, SD = 1.23) compared to non-users (M = 5.17, SD = 1.59). This difference did not reach statistical significance, t(82) = -1.85, p = .068, d = 0.34, 95% CI [-0.07, 0.75]. Exploratory linear regression models found that usage metrics (e.g., total interactions, frequency of practice questions) did not significantly predict final exam grades.

Discussion

This pilot study aimed to provide preliminary insights into student usage patterns, perceptions, and the learning impact of RAG chatbots. We found that while students who used the tools perceived them positively, usage was voluntary, varied greatly, and was concentrated in the days just before the exam. This pattern of use, primarily for topic clarification, did not show a statistically significant correlation with final exam grades.

The lack of a clear learning effect, combined with the observed last-minute usage pattern, aligns with the theoretical frameworks presented in the background. The unscaffolded, voluntary nature of the intervention may have encouraged a "surface learning approach" (Yang et al., 2024) rather than the deeper, "agentic" engagement that pedagogical scaffolding might promote (Smirnova, 2025). Students appeared to use the tool reactively for clarification rather than proactively as a study partner. While usage was associated with positive perceptions (Thway et al., 2024), this pilot suggests that positive perceptions and mere availability do not automatically translate to improved, measurable learning outcomes.

Limitations and Future Work

These findings must be interpreted with caution. First, the study is subject to self-selection bias; students who chose to use the chatbot may differ from non-users in motivation or prior knowledge. Second, the pilot nature of this study, with a small number of survey respondents (N=19) and users (N=35), means the data are not sufficient to derive solid conclusions. The cyberattack also presented a significant, uncontrolled disruption. A broader challenge for this line of research is the lag in institutional infrastructure. Facilitating the creation and deployment of RAG chatbots by teachers for research is not yet streamlined. This creates a delay between pedagogical need and technical capability. This institutional pace conflicts with the pace of the AI industry, which rapidly deploys new tools that students may prefer to use. This complicates research, as students may opt for external, commercial tools instead of university-provided (and potentially older) technology being tested. Future work could address these issues with larger, controlled studies. Such studies might integrate educational principles directly into chatbot system prompts to provide the "scaffolding" (e.g., Smirnova, 2025) that was absent here, guiding students toward more agentic learning practices.

Implications for Educational Practice

The preliminary findings from this pilot study can, at this point, already offer several considerations for educators. First, the results suggest that simply providing access to a course-specific RAG chatbot, even one perceived positively by students, is not a guarantee of improved learning outcomes. Instructors should not assume that students will spontaneously use these tools in pedagogically optimal ways. Our current data is suggesting that the common student may engage with these tools at a very superficial level by default (e.g., last-minute clarification). To foster the deeper, agentic engagement associated with positive learning (Smirnova, 2025; Yang et al., 2024), instructors might design specific, structured activities. For example, rather than leaving use entirely open, an educator could require students to use the chatbot to generate practice questions early in a module, or to use the chatbot to find flaws in an argument, or to critique a chatbot-generated summary of a complex topic. This approach shifts the student's role from passive consumer to active,

critical evaluator. Finally, educators should remain mindful of the technological "lag" discussed in the limitations. If institutional tools are perceived as less capable than rapidly evolving commercial alternatives, students may ignore them. This reality suggests that a robust educational strategy should focus on teaching general AI literacy and critical engagement skills that are transferable across platforms, rather than focusing pedagogy only on a specific, institution-provided tool that may quickly become outdated.

Concluding Remarks

This pilot study provides initial evidence that while students perceive course-specific RAG chatbots as usable and helpful, their mere availability in an unscaffolded, voluntary context does not guarantee enhanced exam performance. The findings suggest that effective pedagogical integration, possibly through structured guidance, is necessary to move students from superficial clarification to deeper, agentic engagement with these tools.

References

Dakshit, S. (2024). *Faculty perspectives on the potential of RAG in Computer Science higher education* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2408.01462

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 319–340.

Dharshan S, Puneeth Ks, Sanjay G J, Vivek V, & Dr Anitha Db. (2025). A review on RAG-based student assistant chatbot using LangChain. *EPRA International Journal of Research & Development (IJRD)*, 112. https://doi.org/10.36713/epra23698

Jeong, C. (2023). A Study on the Implementation of Generative AI Services Using an Enterprise Data-Based LLM Application Architecture. *Advances in Artificial Intelligence and Machine Learning*, *03*(04), 1588–1618. https://doi.org/10.54364/AAIML.2023.1191

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kizi, M. K. Z., & Suh, Y. (2025). Design and performance evaluation of LLM-based RAG pipelines for chatbot services in international student admissions. *Electronics*, *14*(15), 3095. https://doi.org/10.3390/electronics14153095

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, *22*(4–5), 441–504. https://doi.org/10.1007/s11257-011-9118-4

Labadze, L., Grigolia, M., & Machaidze, L. (2023). Role of AI chatbots in education: Systematic literature review. *International Journal of Educational Technology in Higher Education*, *20*(1), Article 1. https://doi.org/10.1186/s41239-023-00426-1

Lang, G., & Gürpinar, T. (2025). AI-powered learning support: A study of retrieval-augmented generation (RAG) chatbot effectiveness in an online course. *Information Systems Education Journal*, *23*(2), 4–13.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., & Rocktäschel, T. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

Maryamah, M., Irfani, M. M., Tri Raharjo, E. B., Rahmi, N. A., Ghani, M., & Raharjana, I. K. (2024). Chatbots in Academia: A Retrieval-Augmented Generation Approach for Improved Efficient Information Access. *2024 16th International Conference on Knowledge and Smart Technology (KST)*, 259–264. https://doi.org/10.1109/KST61284.2024.10499652

Memarian, B., & Doleck, T. (2023). ChatGPT in education: Methods, potentials, and limitations. *Computers in Human Behavior: Artificial Humans*, *1*(2), 100022. https://doi.org/10.1016/j.chbah.2023.100022

Meyer, J. G., Urbanowicz, R. J., Martin, P. C. N., O'Connor, K., Li, R., Peng, P.-C., Bright, T. J., Tatonetti, N., Won, K. J., Gonzalez-Hernandez, G., & Moore, J. H. (2023). ChatGPT and large language models in academia: Opportunities and challenges. *BioData Mining*, *16*(1), Article 1. https://doi.org/10.1186/s13040-023-00339-9

Modran, H., Bogdan, I. C., Ursuțiu, D., Samoila, C., & Modran, P. L. (2024). LLM intelligent agent tutoring in higher education courses using a RAG approach. *Preprints*. https://doi.org/10.20944/preprints202407.0519.v1

Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., & Qiao, M. S. (2021). AI literacy: Definition, teaching, evaluation and ethical issues. *Proceedings of the Association for Information Science and Technology*, *58*(1), 504–509. https://doi.org/10.1002/pra2.487

Nisanth, P. (2025). RAG-based AI chatbot for student and institutional assistance. *International Journal for Research in Applied Science and Engineering Technology*, *13*(9), 858–863. https://doi.org/10.22214/ijraset.2025.73970

Oliveira, M. J. B., Zednik, C., Bombaerts, G., Sadowski, B., & Conijn, R. (2025). *Assessing students' DRIVE: An evidence-based framework to evaluate learning through students' interactions with generative AI*. PsyArXiv. https://doi.org/10.31234/osf.io/cne9j_v3

Parekh, K. V., Saxena, N., & Ansari, M. A. (2025). A comparative study of retrieval-augmented generation (RAG) chatbots. *2025 International Conference Automatics, Robotics and Artificial Intelligence (ICARAI)*, 1–6. https://doi.org/10.1109/ICARAI67046.2025.11137956

Smirnova, L. (2025). Developing students' agency and voice by using generative AI in an online EAP module. *Innovation in Language Learning and Teaching*, 1–11. https://doi.org/10.1080/17501229.2025.2538781

Thway, M., Recatala-Gomez, J., Lim, F. S., Hippalgaonkar, K., & Ng, L. W. T. (2024). *Battling botpoop using GenAI for higher education: A study of a retrieval augmented generation chatbots impact on learning* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2406.07796

Yang, Y., Luo, J., Yang, M., Yang, R., & Chen, J. (2024). From surface to deep learning approaches with Generative AI in higher education: An analytical framework of

student agency. *Studies in Higher Education*, *49*(5), 817–830. https://doi.org/10.1080/03075079.2024.2327003

Zhang, Z. (2025). Integrating generative AI in higher education: Practical applications and institutional guidelines. *Education Journal*, *14*(3), 88–102. https://doi.org/10.11648/j.edu.20251403.12