EVALUATING RAG-BASED CHATBOT PERFORMANCE IN STEM AND SOCIAL SCIENCES: A METRIC-DRIVEN COMPARISON

Huiyu Zhang*[a], Ester Goh[a], Kalyankumar Subramaniyan[b], Kok Hian Lee[b] and Nurzahiah Jumat[c]

[a] School of Informatics & IT
[b] IT Services
[c] School of Humanities & Social Sciences
Temasek Polytechnic, Singapore

zhang_huiyu@tp.edu.sg*

Students are becoming more familiar with ChatGPT, a generative AI chatbot that provides quick access to information, facilitates Q&A interactions, and offers task-related feedback. However, ChatGPT's tendency to generate hallucinatory responses presents learning challenges. To address this, the TP AI Assistant, a Retrieval-Augmented Generation (RAG)-based chatbot, was developed to provide a structured and reliable learning environment. Unlike ChatGPT, it delivers accurate, context-specific information aligned with learning objectives, ensuring relevance to coursework.

Deployed over an academic semester starting in October 2024, the TP AI Assistant supported approximately 611 Year 1 STEM students and 350 Year 1 Humanities and Social Science students at Temasek Polytechnic, Singapore. This study examines chatbot engagement metrics to analyze usage patterns across the two diverse disciplines. Data collected from chatbot logs was transformed into key metrics, including response rate, confusion rate, containment rate, conversation length, and duration. These metrics assessed engagement levels, the chatbot's role in personalized learning, and its reliability in supporting the varied learning needs across disciplines. Findings revealed that regardless of discipline, chatbot engagement depended on several common factors, including the students' familiarity with the chatbot, the perceived usefulness of the responses, reliability and user experience provided by the chatbot. Interaction patterns suggested that students engaging in more generative tasks, such as programming applications, may have reached conversation dead ends more frequently as their queries extended beyond the RAG's knowledge base. In contrast, students who appeared to seek factual, conceptual, and procedural support tended to have more sustained interactions. This suggests that while the RAG-based chatbot effectively provided targeted information at scale, it had limitations in handling tasks requiring broader adaptability.

This study underscores the importance of instructional design in shaping effective AI-assisted learning. Regardless of the subject area, a well-structured instructional design ensures that AI tools align with diverse learning needs, cognitive processes, and pedagogical strategies. It also advances the understanding of RAG-based chatbots in supporting discipline-specific learning needs, such as the need for multimodal RAG models and additional student support mechanisms, such as prompt engineering guidance and enhanced chatbot responses, particularly for STEM subjects like engineering, IT, and applied sciences, where visual representations are crucial for understanding complex concepts.

Keywords: *personalized learning, Retrieval-augmented generation (RAG), self-regulated learning, STEM, Social Science*

Introduction

The Covid-19 pandemic lockdowns highlighted the diverse needs of students, reinforcing the importance of personalized learning. Students not only faced academic challenges but also socio-emotional difficulties, such as maintaining motivation, building autonomy, and staying connected. Learning needs varied widely, including differences in pacing, preferred formats, and feedback preferences, emphasizing the demand for tailored educational approaches. Personalized learning, as defined by Gunawardena (2024), involves adapting instruction to individual strengths, needs, and interests while allowing choice, voice, and flexibility in achieving learning outcomes. Advances in AI now offer new ways to meet these needs, fostering greater equity and inclusivity.

One promising application of AI in education is through dynamic question-and-answer interactions, essential for clarifying concepts and deepening understanding. Research has shown that student-generated questions promote critical thinking and inquiry-based learning, aligning with Vygotsky's social constructivism (1978), which emphasizes the importance of interactive learning. AI chatbots can facilitate such interactions at scale, enabling self-paced, reflective conversations while providing instant, tailored responses. This approach supports the inquiry-driven nature of learning in virtual and blended classrooms, promoting both engagement and academic growth.

Developing effective AI chatbots has become a priority. AI chatbots generally fall into two types: retrieval-based and generation-based. Retrieval-based chatbots respond based on a set of predetermined answers, ensuring relevance but sometimes lacking depth, especially when addressing nuanced or discipline-specific questions. In contrast, generation-based chatbots create responses from a pre-trained knowledge base, offering flexibility but occasionally producing less accurate replies. Balancing these strengths and limitations, the Retrieval-Augmented Generation (RAG) architecture (Lewis et al., 2020) combines both approaches, retrieving accurate information while generating context-aware responses. Tapping into its possibilities, Temasek Polytechnic (TP), one of the five public polytechnics in Singapore, has also implemented a RAG-powered chatbot, the TP AI Assistant.

However, while the RAG model improves chatbot performance by merging retrieval and generation methods, its effectiveness can vary significantly across academic disciplines. Based on the findings by Shanahan and Shanahan (2008), preliminary evidence suggests that experts in disciplines such as mathematics, chemistry, and history engage with texts in markedly different ways. These differences point to the need for distinct comprehension strategies tailored to each discipline and highlight that different academic fields demand unique approaches to thinking, reasoning, and communication. In STEM fields, students must possess strong problem-solving and critical-thinking skills, enabling them to plan, analyze, and develop processes and projects that address real-world problems (Gao et al., 2020; Han et al., 2021). Similarly, the TPACK framework (Mishra & Koehler, 2006) highlights the importance of integrating content knowledge with pedagogical approaches specific to each subject. More recent studies have also emphasized the need for discipline-specific AI applications in education.

VanLehn's (2005) work on Intelligent Tutoring Systems highlights that AI systems in STEM disciplines need to provide structured support to help students develop higher-order thinking skills essential for deep comprehension and reasoning. On the other hand, Olatunbosun et al. (2024) draw attention to the challenges posed by the technical complexity of STEM content itself, noting that AI and Machine Learning applications must be able to interpret and respond to highly specialized, domain-specific knowledge in order to support effective learning. In contrast, Mayfield and Black (2020) argue that in writing tasks, AI support should emphasize learner autonomy and preserve the writer's unique voice, advocating for guidance that nurtures reflection and development rather than directive scaffolding. Qu et al. (2024) also pointed out the differences in how students from varying disciplines interpret and interact with GenAI chatbot responses. These findings highlight the need to conduct comparative analysis using unbiased chatbot interaction logs to better understand how RAG chatbots perform across different academic contexts and identify areas for improvement in discipline-specific applications.

The TP AI Assistant: Technology Description

The core of the TP AI Assistant system revolves around the RAG architecture, which integrates retrieval-based and generation-based AI capabilities to enhance the chatbot's performance. It was developed by leveraging various Azure services to create a robust, secure, and efficient chatbot system. The process begins with subject leaders uploading training materials to a cloud-based collaboration platform under Microsoft. To ensure centralized access, indexing, and retrieval while minimizing manual effort and errors, an Azure Logic App then automatically copies these documents to a secure Storage Account. This setup not only streamlines document management but also supports efficient data processing.

Once the documents are securely stored, the Logic App triggers an indexer within AI Search, making the uploaded content retrievable. Indexing involves analyzing the text, extracting key phrases, and creating structured representations that the system can quickly search through. This indexing process is essential for efficiently retrieving relevant information later on. The RAG process works as follows:

Indexing: After the documents are uploaded and processed, the AI Search component creates an index of the data, making it searchable. The indexer converts raw text into structured data, allowing for fast lookups.

Retrieval: When a user poses a question via the ChatGPT-like web application, the system first searches the indexed data to find the most relevant content. This step ensures that the response is grounded in the existing knowledge base, maintaining accuracy and relevance.

Augmenting the Prompt: The retrieved content is then used to formulate a context-enhanced prompt for the Azure OpenAI model. By incorporating the most relevant information into the prompt, the model generates responses that are not only contextually accurate but also enriched with precise data.

Generation: Finally, the OpenAI model uses the augmented prompt to generate a well-formed response, which is delivered back to the user through the web app.

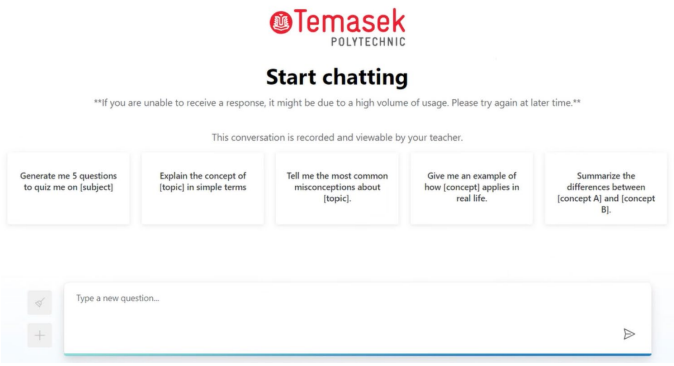

Fig. 1. User interface of the TP AI Assistant.

To facilitate user access, a ChatGPT-like web application (see Fig. 1) was designed and hosted on Azure App Service. Users must authenticate through Microsoft Entra ID to ensure secure and authorized

access to the platform. This authentication layer guarantees that only verified users can interact with the assistant. All interactions, including both user questions and assistant responses, are logged and stored securely in Cosmos DB. This database not only keeps a record of conversations for analysis but also supports continuous improvement by allowing developers to assess chatbot performance.

The entire architecture is primarily built within the Singapore Government on Commercial Cloud (GCC) environment, where most Azure services are hosted, except for Azure OpenAI Service, which resides outside Singapore. To maintain robust security and confidentiality, all data transfers and storage operations are secured using industry-standard encryption protocols, adhering strictly to regulatory compliance and data protection policies.

The Study

The study took place at Temasek Polytechnic (TP), focusing on pre-employment training (PET) students. A total of 611 Year 1 students from six different diplomas within the School of Informatics & IT (IIT), representing the STEM disciplines, and 350 Year 1 students from three different diplomas within the School of Humanities and Social Sciences (HSS), representing the social science disciplines. The gender distribution for IIT was approximately 70% male and 30% female, whereas for HSS, it was about 15% male and 85% female. The median age of participants was 17, reflecting the typical demographic profile of Singaporean youth, comprising a diverse mix of ethnicities such as Chinese, Malay, Indian, and Eurasian.

During the academic semester starting in October 2024, students were granted access to the TP AI Assistant. The respective subject teams conducted an induction to familiarize participants with the features and functions of the chatbot.

Table 1. Purpose Of the TP AI Assistant by Discipline

	STEM	Social Science
Subject	App Development	Effective Communication
Objective	Enhance database design for a project	Answer FAQs related to subject administration
Training materials for RAG	Theoretical concepts of normalization and database anomalies & practical SQL operations	Lecture slides, assessment specification document and student handbook
Remarks	Students experienced the 1st iteration of the TP AI Assistant in the April 2024 academic semester in foundational programming subject.	In the same subject, students were also given access to a separate chatbot, built on a different platform, which provided feedback on their report writing.

Data and Techniques

The activity logs from the TP AI Assistant were extracted from Cosmos DB, covering the period from 21 October 2024 to 26 January 2025. This timeframe was selected as it encompassed the conclusion of the chatbot-supported assessment for IIT students and for both subjects, the subsequent two weeks were reserved for consultations and in-class assessments, during which no formal lessons were conducted. All student identifiers were removed and replaced with numeric participant IDs to ensure anonymity. The logs comprised four main elements: 'timestamp' (indicating the precise date and time of each interaction), 'ID' (an anonymized participant number), 'content' (the text exchanged during chatbot sessions), and 'role' (to indicate whether the message came from the student or the chatbot).

Key Chatbot Metrics

Key chatbot usage and performance metrics were used to guide the comparative analysis, as detailed in Table 2. Among these metrics, the frequency of the default fallback message "The requested information is not available in the retrieved data. Please try another query or topic." was also examined as an indicator of user engagement.

Table 2. Key Chatbot Metrics

Metric	Description
Total interactions	Total number of queries made by users
Total sessions	A session starts when a user sends a message and ends after inactivity or when the chat is closed. Total such conversations made by users
Conversation length	Average number of queries per session. *Total interactions ÷ Total sessions*
Duration	Time spent per chatbot session.
Adoption rate	Percentage of cohort who used the chatbot
Retention rate	Percentage of users who have used on repeated occasions over a given period
Single turn rate	Percentage of sessions with only one query
Multi-turn rate	Percentage of sessions with more than one query in context
Containment rate	Percentage of sessions successfully resolved without the fallback message
Confusion rate	*Total fallback responses ÷ total interactions *100%*

Results

Building on the earlier table that outlined the key metrics, the activity logs provided the actual values for the IIT and HSS participants as depicted in Table 3.

Table 3. Usage Patterns by Discipline

Metric	STEM	Social Science
Total interactions	342	275
Total sessions	114	99
Conversation length	3	2.78
Duration (sec)	242,468	106,560
Adoption rate (%)	14.6	16.0
Retention rate (%)	10.1	30.4
Single turn rate (%)	33.3	41.4
Multi-turn rate (%)	66.7	58.6
Containment rate (%)	67.5	48.5
Confusion rate (%)	52.3	56.4

STEM students appeared to demonstrate higher engagement with the chatbot, reflected by a total of 342 interactions and 114 sessions; compared to 275 interactions and 99 sessions by Social Science students. STEM students also exhibited longer usage durations (242,468 sec) than Social Science students (106,560 sec), and their conversations were slightly longer, averaging 3 interactions per session compared to 2.78 in Social Science which is approximately 23 times higher. Both disciplines displayed higher multi-turn interactions than single-turn interactions.

The containment rate was higher for STEM students (67.5%; higher is better) than for Social Science students (48.5%). Meanwhile, confusion rates were relatively similar: 52.3% for STEM and 56.4% for Social Science. Social Science students showed a slightly higher adoption rate (16.0%) compared to STEM students (14.6%). Retention was also higher among Social Science students (30.4%) than STEM students (10.1%).

Discussion

The differences in engagement, retention, and interaction patterns across disciplines point to underlying factors that may have shaped how students from different academic contexts engaged with the TP AI Assistant.

Engagement Intensity: It was also observed that the STEM students exhibited longer, more exploratory interactions. This suggests that they were more inclined to engage in extended dialogues with the chatbot, possibly driven by the technical complexity of their queries and the TP AI Assistant could help to address- to implement a technical solution, which often required multiple exchanges to reach a resolution. In contrast, conversations by Social Science students tended to be shorter, with a lower average number of turns per session. This pattern can partly be attributed to the chatbot's original design, which focused on answering administrative FAQs that address non-academic or logistical aspects related to the subject, including class schedules, assessment formats, and submission procedures, leading to quicker, task-oriented exchanges. Additionally, analysis of chatbot responses revealed that some Social Science students engaged the chatbot to rewrite parts of their assignments, receiving immediate outputs that reinforced brief interactions and instant gratification.

Meanwhile, STEM students exhibited longer, more sustained dialogues, as the TP AI Assistant implemented gatekeeping mechanisms that refused to provide direct coding solutions. This approach encouraged deeper, multi-step problem-solving conversations, aligning with the more complex and iterative nature of technical queries.

Students often required more than one exchange to fully address their queries, reflecting the need for clarification, elaboration, or step-by-step support during the interaction. In RAG-based chatbots, where users pose open-ended questions rather than selecting from a fixed menu, deeper support for self-regulation is often needed. Once again, we observed STEM students required more multi-turn interactions, as the complexity of technical queries and the nature of support provided by the chatbot demanded iterative problem-solving rather than immediate answers.

Query Resolution Efficiency: The higher containment rate among STEM students may stem from the more structured and specific nature of their queries in the IT domain, which aligned well with the training materials provided and the RAG system's retrieval capabilities. The lower containment rate in Social Science students reflects the challenges the RAG chatbot faced when handling subjective, contextually layered queries common in Social Sciences disciplines. The slightly higher confusion rate in Social Science also suggests that the system may struggle more when responding to open-ended or interpretive questions. In addition to this inherent complexity, another possible factor is that the same group of Social Science students was also introduced to a second chatbot during the same period, which was intended for providing feedback. This overlap may have led to confusion over which chatbot to use for which purpose, resulting in mismatched queries to the TP AI Assistant and subsequently affecting the retrieval accuracy.

Adoption and Sustainability: Social Science students appeared to show a slightly higher adoption rate than the STEM students suggest they were more open to trying the system, likely driven by the need for administrative clarity such as the defined boundaries on the Dos and Don'ts for assessments, before they could continue with their tasks. Furthermore, the higher retention rate among Social Science students than the STEM students indicates that they were more willing to revisit the TP AI Assistant, even when queries were not fully resolved. This trend is notable given the lower containment rate and slightly higher confusion rate were observed in these Social Science students, suggesting that they students valued the

assistant's support, despite occasional retrieval mismatches.

Conversely, although STEM students demonstrated higher engagement intensity, including longer sessions, more multi-turn conversations, and a higher containment rate, they exhibited lower retention. This paradox highlights that even when queries were successfully addressed in individual sessions, STEM users were less inclined to return. The relatively lower confusion rate in STEM indicates that while the assistant was able to handle the queries well in general, the interaction remained largely transactional, which could be a characteristic of STEM discipline, as discussed by Fairhurst et al. (2023). The combination of high containment but low retention among STEM users suggests that once technical issues were resolved, there was little perceived need for further engagement, possibly compounded by fatigue from their academic workload, as reflected in the students' subject evaluation survey. This survey, regularly conducted at the end of each semester in TP, gathers feedback on educational quality across subjects. These findings emphasize the need for discipline-specific refinements: while Social Science students may benefit from improved information synthesis to reduce confusion, STEM users may require strategies that promote relational use, encouraging continued engagement beyond immediate task resolution, as echoed by Kunze and Rutherford (2022).

Other Factors Influencing Performance: Building upon the observed engagement patterns, resolution outcomes, and adoption trends derived from unbiased metrics, and considering the broader learning context, several additional factors were identified that may have influenced students' interactions with the TP AI Assistant. Students' familiarity with the chatbot appeared to affect their ability to engage effectively, particularly in cases where overlapping chatbot services introduced confusion. Perceptions of the chatbot's usefulness and reliability, reflected through containment and confusion rates, also played a role in shaping sustained use. Furthermore, the observed interaction patterns may suggest that the nature of students' tasks influenced the quality of their engagement. For instance, students involved in more complex, generative tasks—such as programming or application design—could have been more likely to encounter conversation dead ends, as reflected in longer sessions with lower containment for some users. Meanwhile, students seeking factual, conceptual, or procedural support appeared to maintain more sustained interactions, which were more readily addressed within the RAG system's retrieval scope. However, these interpretations remain speculative and would require direct analysis of the response texts for confirmation. These patterns suggest that while the RAG-based chatbot effectively provided targeted information at scale, it faced limitations when supporting broader, more adaptive learning tasks, underscoring the need for continuous enhancement of retrieval adaptability and user support strategies.

Conclusion

This study enhances our understanding of how students from various academic disciplines interact with a RAG-based chatbot designed for personalized learning, promoting self-directed acquisition of knowledge and skills. By leveraging Azure services, the proposed RAG approach demonstrates scalability and reliability, addressing the need for an effective learning tool that minimizes time commitment for both students and lecturers. Its ability to adapt to diverse subject areas, support discipline-specific learning needs, and generate data for learning analytics, while maintaining consistent performance, positions it as a valuable asset for AI-driven educational practices.

The RAG-based TP AI Assistant was evaluated across different academic contexts, revealing distinct discipline-specific usage patterns. Generally, STEM discipline was characterized by higher containment, longer session duration, and lower retention, while Social Science discipline exhibited shorter sessions, higher retention, and simpler conversations. Conversation length and session duration were significantly higher in STEM, aligning with more complex, multi-step problem-solving, whereas Social Science users engaged in briefer, often exploratory interactions. Despite the higher containment rate in STEM, the slightly higher confusion rate observed in Social Science highlights the challenges RAG systems face when addressing the subjective, interpretive nature of social sciences queries. These findings emphasize that instructional design must not only address cognitive load and task complexity but also consider how different disciplines approach knowledge acquisition and engagement with AI tools.

Moving forward, the results suggest several areas for enhancing RAG-based educational chatbots. In STEM domains, the need for multimodal RAG models becomes evident, as visual representations and richer content delivery can better support the comprehension of complex technical concepts. Additionally, providing students with prompt engineering guidance and refining chatbot responses could strengthen self-directed learning, especially for technically demanding fields like engineering, IT, and applied sciences. For Social Science disciplines, improvements in context handling and retrieval adaptation could enhance user experience, supporting sustained engagement even when definitive answers are not always available. Overall, this study underscores that effective AI-assisted learning requires careful instructional design, ensuring that AI tools are pedagogically aligned and adaptable to diverse disciplinary contexts.

References

Fairhurst, N., Koul, R., & Sheffield, R. (2023). Students' perceptions of their STEM learning environment. *Learn. Env. Research*, 26(3), 977–998.

Gao, X., Li, P., Shen, J., & Sun, H. (2020). Reviewing assessment of student learning in interdisciplinary STEM education. *Int. J. STEM Educ.* 7(24).

Gunawardena, M., Bishop, P., & Aviruppola, K. (2024). *Personalized learning: The simple, the complicated, the complex and the chaotic*. Teach. Educ., 139, 104429.

Han, J., Kelley, T., & Knowles, J. G. (2021). Factors influencing student STEM learning: self-efficacy and outcome expectancy, 21st century skills, and career awareness. *J. STEM Educ. Res.* 4, 117–137.

Kunze, A. & Rutherford., T. (2022). Students' Discipline Specific Perceptions of Learning Practices. *Intl. J. Teach. Learn. Higher Edu.,* 33(2), 53-167.

Lewis, P., E. Perez, E, Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, N., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & D. Kiela. (2020). *Retrieval-augmented generation for knowledge-intensive NLP tasks.* Ann. Neural Inf. Pro cess. Syst. (NeurIPS), 33, 9459–9474.

Mayfield, E. & Black, A.W. (2020). Should You Fine-Tune BERT for Automated Essay Scoring? *Workshop on Innovative Use of NLP for Building Educational Applications*.

Mishra, P. & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for integrating technology in teacher knowledge. *Teach. Col. Rec., 108*(6), 1017–1054.

Olatunbosun, J. O. & Nwankwo, C. A. (2024). Integrating AI and machine learning in STEM education: Challenges and opportunities. *Comput. Sci. IT Res. J.*, 5(8), 7–8.

Qu, Y., Tan, M.X.Y. & Wang, J. (2024). Disciplinary differences in undergraduate students' engagement with generative artificial intelligence. *Smart Learn. Environ.* 11, 51.

Shanahan, T. & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: Rethinking content-area literacy. *Harvard Edu. Rev., 78*(1), 40–59.

VanLehn, K. (2005). The behavior of tutoring systems. *Intl. J. AI Ed*., 16(3), 227–265.

Vygotsky, L. (1978). *Mind in Society*. Massachusetts: Harvard Univ. Press.