GenAI for teaching and learning: a Human-in-the-loop Approach

Wei Qiu
Centre for the Applications of Teaching & Learning Analytics for Students (ATLAS)
Nanyang Technological University (NTU), Singapore
qiuwei@ntu.edu.sg

Maung Thway
ATLAS NTU, Singapore
maung.thway@ntu.edu.sg

Joel Weijia Lai
ATLAS NTU, Singapore
joellai@ntu.edu.sg

Fun Siong Lim
ATLAS NTU, Singapore
LIM_Fun_Siong@ntu.edu.sg

ABSTRACT: This paper presents a human-in-the-loop development and implementation of a Socratic generative artificial intelligence (GenAI) tutor for undergraduate statistics courses. GenAI has potential to personalize and encourage desired deep learning behaviors in a diverse student population. However, thorough planning and evaluations are essential to ensure responsible use of AI. Our systematic approach started with a GenAI tutor designed with course coordinators and instructors, followed by a trial phase involving student volunteers and instructors. The GenAI tutor was piloted in a real class setting, with data collected on the conversation logs, the experiences of both students and instructors, as well as the resulting outcomes. This approach fosters trust in GenAI and facilitates continuous improvement. The findings contribute to the ongoing discourse surrounding the use of AI in learning environments, with a particular focus on enhancing human capabilities.

Keywords: Human-in-the-loop, generative AI, pedagogically designed chatbot, analytics dashboard, technology-enhanced learning

1 INTRODUCTION

The integration of generative artificial intelligence (GenAI) into education is transforming the way learning support can be designed and delivered. GenAI shows huge potential to offer personalized learning to students at scale. However, ensuring that these agents effectively contribute to the intended learning outcomes requires thoughtful design and continuous human involvement. Before any large-scale deployment, their effectiveness in terms of accuracy of responses, quality of engagement, and learning gains must be rigorously designed and studied.

2 THE CHATBOT DESIGN APPROACH

Figure 1 depicts our approach to implementing GenAI tutors responsibly. Our journey begins with identifying the courses most in need of additional teaching support. Statistics is a core subject for the

large and diverse undergraduate student population at our university. In the statistics course of this paper, 90% of the students have historically requested for tutoring support. There are approximately 600 students in each cohort, making it a prime candidate for experimentation.

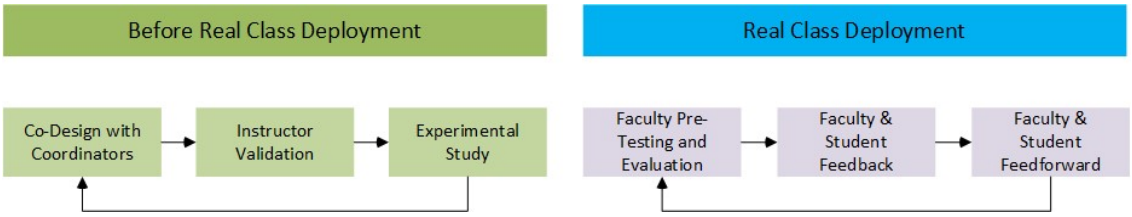

Figure 1: Human-in-the-Loop GenAI Deployment Approach

Co-Design with Coordinators: The coordinators wanted the GenAI tutor to be pedagogically designed as a Socratic GenAI tutor to facilitate inquiry-based, step-by-step problem-solving to deepen students' conceptual understanding of statistic concepts taught in the course. As the university's learning analytics center, the development team also wanted to give the instructors the ability to monitor their students' use of the GenAI tutor in their learning.

As such, the design of the GenAI tutor comprised two main parts: a student-facing GenAI tutor and a faculty-facing dashboard. Developed as a prototype using Streamlit web application framework (https://streamlit.io/), the GenAI tutor aided students in learning statistics through a ChatGPT-like interface. As shown in Figure 2, it was prompt engineered and grounded with curated course materials to enable the desired Socratic pedagogical behavior and learning experience.

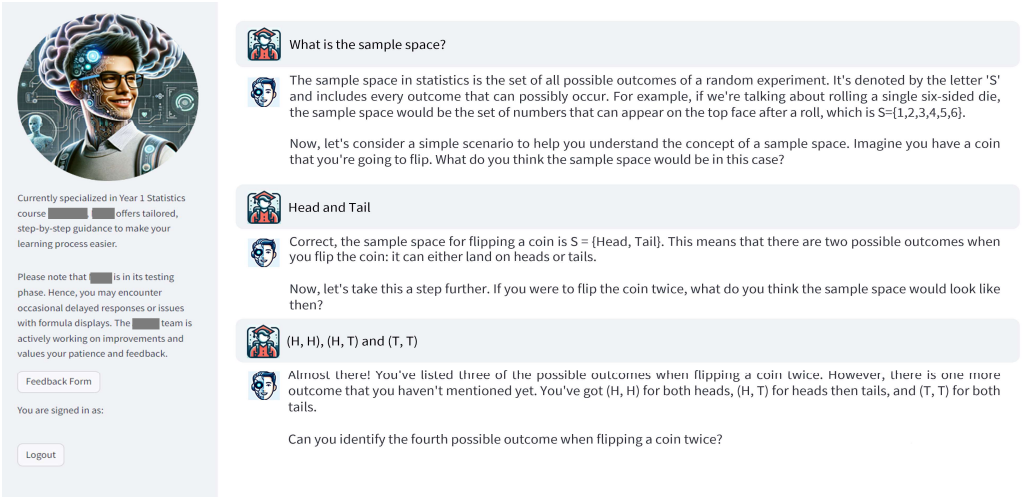

Figure 2: GenAI tutor which is 1) grounded with course materials, and 2) prompt-engineered to deconstruct questions and guide students.

Instructor Validation: Before experimenting the GenAI tutor with students, course instructors were recruited to assess the accuracy and the quality of the GenAI tutor's response (Qiu W. , et al., 2024). These subject matter experts were asked to evaluate the chatbot's answers against two predefined criteria identified by the course coordinators – accuracy of the response and quality of the engagement. The evaluation revealed an accuracy rate exceeding 80%, with a high level of agreement among evaluators, reflected in a Randolph kappa score of 0.704. At the same time, we noted from the instructor validation and feedback that the Socratic engagement needed to be further refined with prompt engineering.

Experimental Study with Students: Before piloting the GenAI tutor in a real class, we conducted an experimental study to understand its impact on student learning. Following approval from the

Institutional Review Board, we invited students who were either weak in statistics or had never taken statistics courses before to participate in an experimental study. Students were asked for their consent at the beginning of the study and were informed that they could withdraw from the study at any given time. Moreover, participants who provided feedback were made aware that their responses would be recorded for research purposes and were assured that their personal information would remain confidential throughout the process.

The three-week study randomly assigned 45 student volunteers into a control and experimental group. The control group interacted with the baseline chatbot using GPT-4-Turbo. The experimental group used our Socratic GenAI tutor with the same GPT-4 model. Both chatbots are grounded with the same curated materials. The key difference was that the baseline chatbot was not prompt engineered, and students were unaware of their assignment. Pre and post-tests were administered on 12 topics learned and post-study feedback was collected along with conversation log data. The experimental group showed higher learning gains compared to the control group (Qiu W. , et al., 2024), especially when they questioned the GenAI tutor's response and asked for application examples to test their knowledge (Lai, et al., 2024). The Socratic approach encouraged students to engage more often and on more complex topics. Finally, student feedback indicated a preference for our GenAI tutor, given its explanations and guidance, but suggested that its response time can be improved. Overall, the findings gave us the confidence to deploy the GenAI tutor in a real class with some refinements to ensure faster performance.

At the same time, the student-chatbot interactions were collected, analyzed, and presented in a faculty- dashboard. In essence, we wanted the dashboard to be both a learning support tool and a resource for improving teaching strategies based on real-time analytics. We surveyed the coordinators and other faculty (n = 15) on the measures that they want presented on the dashboard. The popular choices included the frequency students use the GenAI tutor, the topics asked, the cognitive level of student questions, and student feedback on the GenAI tutor's answers. These feedback were incorporated into the design of the faculty dashboard as shown in Figure 3.

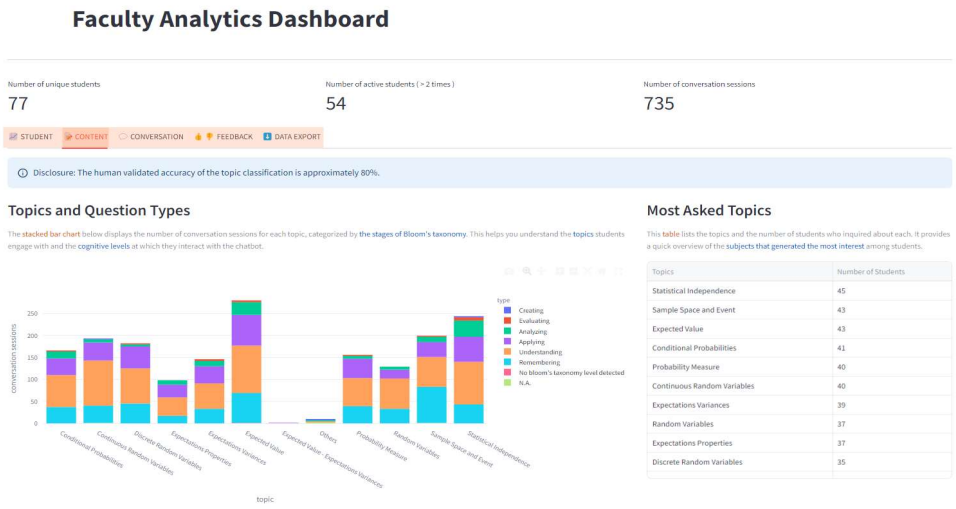

Figure 3: The Content section view of the faculty analytics dashboard.

3 THE EXPERIMENT IN PRACTICE

In preparation for a large-scale rollout, we collaborated extensively with course coordinators and instructors to conduct a pilot study within real classroom environments for a few weeks. This pilot was an essential component of our human-in-the-loop approach, ensuring that the GenAI tutor was

not only effective in a controlled environment but was also aligned with both faculty expectations and student learning needs in a real class.

Faculty Pre-Testing and Evaluation: The course coordinator and instructors were again invited to thoroughly test the GenAI tutor's performance since a new model, GPT-4o, was released. They assessed its responses across a variety of content areas, ensuring that the GenAI tutor's outputs were accurate, contextually appropriate, and pedagogically sound. They also gave feedback on the faculty dashboard, which led to the addition of a new data export function and several quality-of-life improvements.

Faculty/Student Feedback and Feedforward: Throughout the pilot, the course coordinators and instructors worked with the development team to address major technical and performance issues. At the same time, they monitored and provided elaboration on responses from the Socratic GenAI tutor that students downvoted. This feedback loop was essential for not only addressing hallucinations but also ensuring the completeness of the responses, with the ultimate goal of improving the quality of the responses generated by the chatbot.

At the time of writing, the GenAI tutor is still being piloted. We intend to collect student and faculty experience as well as performance data to improve its usefulness for statistical learning before rolling out for the entire semester and other similar courses. We anticipate that feedforward will be an ongoing iterative process as the technology advances and as learning needs evolve.

4 CONCLUSION

Our experience in implementing a pedagogically designed GenAI tutor demonstrated both the potential and importance of involving humans when using AI to support learning. The encouraging findings from the various stages of implementation may have shown the power of AI, but this is only possible with human involvement every step of the way. We hope that our approach offers ideas for other institutions to discuss how best to implement GenAI for education responsibly.

ACKNOWLEDGEMENTS

We want to express gratitude to our ATLAS team members Ms. Nurabidah Binte Jamil, Ms. Chit Lin Su, Mr. Samuel Ng Soo Hwee, Mr. Lei Zhang, and Mr. Daniel Hernandez for their invaluable contributions to the development and study of the GenAI tutor.

REFERENCES

Lai, J., Qiu, W., Thway, M., Zhang, L., Jamil, N., Chit, L., . . . Lim, F. (2024). Leveraging Process-Action Epistemic Network Analysis to Illuminate Student Self-Regulated Learning with a Socratic Chatbot. *EdArXiv Preprints*.

Okonkwo, C. A.-I. (2021). Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*. doi:https://doi.org/10.1016/j.caeai.2021.100033

Qiu, W., Chit, L., Jamil, N., Ng, S., Chen , C.-M., & Lim, F. (2024). "I Am Here To Guide You": A Detailed Examination of Late 2023 Gen-AI Tutors Capabilities in Stepwise Tutoring in an Undergraduate Statistics Course. *18th International Technology, Education and Development Conference*.

Qiu, W., Chit, L., Jamil, N., Thway, M., Ng, S., Zhang, L., . . . Lai, J. (2024). A Systematic Approach to Evaluate the Use of Chatbots in Educational Contexts: Learning Gains, Engagements and Perceptions. *EdArXiv Preprints.*