ASCILITE 2024

Navigating the Terrain:

Emerging Frontiers in Learning Spaces, Pedagogies, and Technologies

Using LLMs to support teacher reflections on using questions to deepen learning and promote student engagement

Aman Abidi, Farhan Ali, Choon Lang Gwendoline Quek, Ruilin Elizabeth Koh
National Institute of Education, Nanyang Technological University

While much attention has been focused on improving student outcomes, there is growing interest in supporting teachers' reflection on their teaching practice using Generative Artificial Intelligence (GenAI) technologies. This paper examines the application of Large Language Models (LLMs) such as GPT-4 to automate the analysis of secondary school teachers' teaching practice in Singapore, specifically within the context of one of the teaching areas identified by the Singapore Teaching Practice model: using questions to deepen learning. We aimed to demonstrate the effectiveness of LLMs in analyzing classroom lessons in this teaching area. The methodologies employed in this study included the collection of classroom data and their analysis, both manually and using LLMs. Specifically, this involved transcribing the classroom lessons and analyzing each question using LLMs, with the results compared to a ground-truth dataset created through manual analysis. The findings suggest that LLMs are effective in providing, forming the basis for future teacher reflection and the potential for automated self-reflection tools in Singapore schools.

Keywords: Generative AI, Teaching analytics, Reflective Teaching, Automation, Singapore Education

Introduction

The rapid advancement of generative artificial intelligence (GenAI) has opened new avenues for enhancing teaching and learning processes, particularly through teaching practice tracking and assessment (Wysel, 2023). While much of the focus has been on leveraging AI to improve student outcomes, there is expanding interest in applying GenAI technologies to support teachers in reflecting on their teaching practice. This paper explore proposes the potential use of Large Language Models (LLMs) for automating the analysis of teaching practice in Singapore schools, laying the foundation for future self-reflection applications. By utilizing natural language processing (NLP) techniques and manual labeling, we aim to demonstrate the effectiveness of LLMs in analyzing classroom lessons in the teaching area of using questions to deepen learning and promote student engagement. In the following sections, we will detail the background, methodologies, findings, and discussions on how LLMs can serve as a basis for enhancing teachers' self-reflection in the future.

Background

The integration of generative artificial intelligence (GenAI) into education has the potential to impact teaching and learning. Advances in natural language processing (NLP) have led to the development of large language models (LLMs) like GPT-4, which can analyze text in a human-like manner. In Singapore, the precise educational standards in the Singapore Teaching Practice (STP) model (MOE, 2023) require innovative approaches to teaching practice tracking, which LLMs can provide. Reflective teaching has evolved but remains qualitative and subjective, leading to biases and limitations in self-assessment (Hairon, 2020). Psychological biases such as hindsight bias and confirmation bias can affect reflection accuracy (Mahon & O'Neill, 2020). Additionally, unconscious memories and habits influence teaching practices (Miller & Shifflet, 2016). LLMs offer data-driven insights, mitigating biases and enhancing reflective practices by providing objective analyses (Wysel, 2023; Ndukwe & Daniel, 2020). Immediate feedback from LLMs helps teachers quickly identify areas for improvement. They can also track changes over time, supporting sustained reflective practices (Arteaga et

al., 2015). Moreover, LLMs facilitate collaborative reflection by providing common data points for peer discussions, fostering a community of practice focused on continuous improvement (Ku et al., 2018).

Methods

Figure 1 details the structured workflow for the study on using LLM to automate the self-reflection of teaching practice in using questions to deepen learning and promote student engagement. The process was divided into three main phases: data collection, data processing, and data analysis, ensuring a comprehensive approach to investigating LLM's potential for self-reflection of the teaching practice.

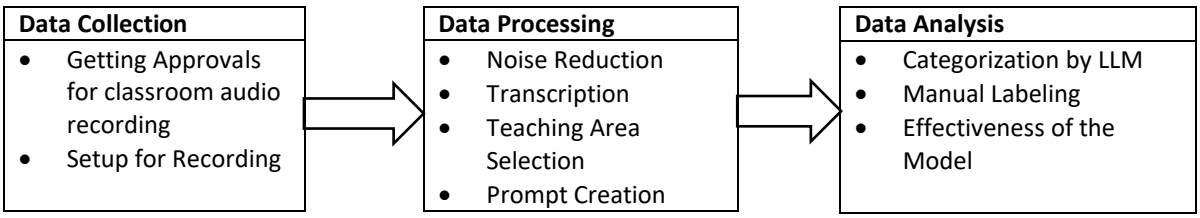

Figure 1. Workflow for the study

Data collection

Classroom data for a Chemistry teacher at a Secondary School was collected with appropriate ethical approvals. The lesson was audio-recorded using high-quality microphones, serving as the primary data. Currently, the study is in its early stages, focusing on one teacher and one lesson for analysis.

Data processing

The data processing phase involved the following steps.

1. **Noise reduction.** Background noise was removed from the audio recordings to ensure clarity.
2. **Transcription.** The cleaned audio files were transcribed using Whisper on Google Colab, a highly accurate speech recognition model. Transcription accuracy was evaluated using word error rate (WER), showing high reliability for further analysis. Table 1 presents the transcription accuracy of a selected teacher's lesson.

Table 1:
Transcription accuracy for a selected teacher's lesson

Metric	WER	Number of words in reference	Number of words in hypothesis	Number of matched words
Value	0.0404	7739	7769	149

3. **Teaching area selection.** The STP model outlines four core teaching processes: classroom culture, lesson preparation, lesson enactment, and assessment. From its 24 areas, "using questions to deepen learning and promote student engagement" was selected for examination due to its focus on higher-order thinking and reflective learning. Research shows that such questioning techniques significantly enhance student engagement and critical thinking, aligning with Cohen et al. (2018), who emphasize interactive methods in fostering student participation and improving learning outcomes.
4. **Prompt creation.** To effectively analyze teaching practice using LLMs, specific rules for prompts were developed based on the selected teaching area. A prompt is a carefully crafted question or statement designed to elicit a specific response from the AI, guiding it to generate relevant outputs and ensuring the analysis remains focused and contextually appropriate. Two rules were established: first, the prompts identified if the teacher question required students to think deeply and critically, using words like "why" and "how." These high-level questions were tailored to the Singaporean educational

setting, for example: "You are a Singaporean educational researcher analyzing classroom teacher talk for evidence of using high-level questions to deepen learning and promote student engagement. High-level questions include words like 'why,' 'how,' 'what if,' 'why not,' 'is it,' 'whether,' 'what else,' 'think,' 'imagine,' and similar words. Questions involving multiple choices or asking students to justify their choices are also considered high-level." Second, prompts also determined if teacher questions encouraged students to elaborate on their answers with detailed explanations relevant to the lesson.

Data analysis

The data analysis phase involved the following steps.

1. **Categorization by LLMs.** We utilized the LLM, specifically gpt-4o, in our research, which was asked to categorize each question using the following query: 'Does the following question fall under the category of using questions to deepen learning and promote student engagement? Answer only 1 if it is in the category and 0 otherwise: [question]' This step aimed to automate the categorization of questions, providing a scalable solution to the challenges posed by traditional reflective practices.
2. **Manual labeling**. Each question in the transcription was manually labeled by an educational expert as *1* (belonging to the category of using questions to deepen learning and promote student engagement) or *0* (otherwise). Manual labeling was necessary to create a ground-truth dataset that could be used to evaluate the AI-generated categorization. The labeled dataset served as a benchmark, allowing researchers to assess the AI's performance. This process aligns with Cohen et al. (2018), approach for ensuring validity through systematic methods and careful data categorization.
3. **Effectiveness of the model.** The effectiveness of the LLMs-generated categorization was evaluated using the performance metrics of precision, sensitivity, F1 score, and accuracy. These metrics were derived from the confusion matrix, which shows how well the LLM performed by comparing correct and incorrect predictions. It helped us see where the LLM made right or wrong predictions by comparing its output to the ground-truth dataset created by manual labeling.

Findings

Figure 2 shows the distribution of questions that fall under the category of using questions to deepen learning and promote student engagement (green) and those that do not (red) across a time interval. We divide the teachers' lessons into 5-minute segments and count the number of utterances in each segment. As a result, Figure 2 contains 14 segments, indicating that the total duration of the lesson is 1 hour and 10 minutes.

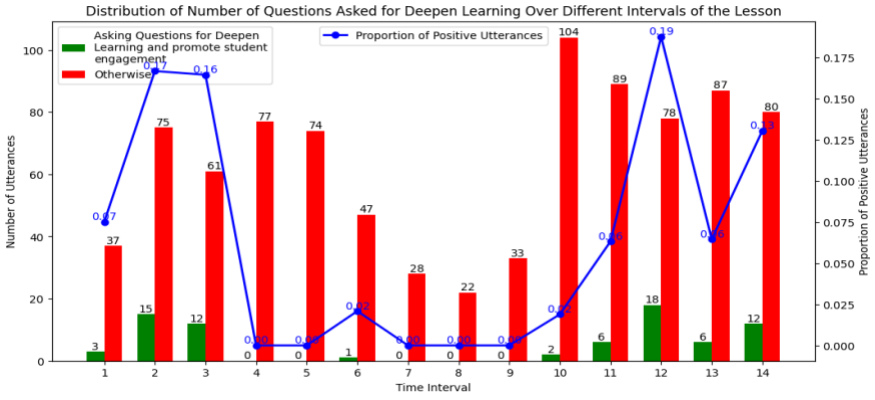

Figure 2. Distribution of questions

Observations and Suggestions. Based on the graph (Figure 2), we can draw the following observations and suggestions for enhancing teacher's instructional practices. At the beginning of the lesson (Intervals 1-3), there is a higher frequency of questions aimed at deepening student learning, indicating stronger engagement. Teachers often use open-ended, thought-provoking questions during this phase to activate prior knowledge and

stimulate curiosity, fostering a reflective learning environment. During the middle part of the lesson (Intervals 4-9), there is a noticeable drop in higher-order questions, suggesting a decline in student engagement. This period typically involves delivering core material, where teachers focus on maintaining momentum. However, incorporating more higher-order and analytical questions in these intervals could help sustain engagement and facilitate deeper learning, ensuring students remain actively involved. Teachers are encouraged to make a deliberate effort to keep the lesson dynamic and interactive during this stage. Towards the end of the lesson (Intervals 10-14), we observe an increase in questioning aimed at deepening learning, reflecting re-engagement with the topic. This phase is crucial for consolidating learning, where teachers can incorporate reflective and summative questions to reinforce key concepts and ensure students have fully grasped the material. By concluding with such questions, the teacher encourages students to reflect on the lesson and solidify their understanding of the topic. In Table 2, the teaching practice metrics in classroom teaching is presented below:

Table 2
Definitions of predictive teaching practice metrics numbers

		Predicted classification for if the Utterance lies under Using Questions to Deepen Learning and Promote Student Engagement	
		Yes	No
Manual classification for if the Utterance lies under Using Questions to Deepen Learning and promote student engagement	Yes	True Positive (TP)	False Negative (FN)
	No	False Positive (FP)	True Negative (TN)
$Accuracy = \frac{TP+TN}{\mathrm{TP}+FP+TN+FN}, Precision = \frac{TP}{\mathrm{TP}+FP}, Sensitivity = \frac{TP}{\mathrm{TP}+FN}$, an$F1\ Score = 2 \times \frac{Sensitivity \times Precision}{Sensitivity+Precision}$			

Further, we present Table 3 representing the confusion matrix. Table 4 summarises the teaching practice metrics. The metrics demonstrate the effectiveness of LLM in categorizing the questions. These metrics are standard and widely accepted across the AI community. The AI model's high precision and accuracy confirm its reliability in identifying advanced questions that enhance learning and engage students. Its sensitivity and F1 scores demonstrate a balanced detection capability.

Table 3
Confusion matrix

	Predicted positive	Predicted negative
Actual positive	*TP*: 52	*FN*: 25
Actual negative	*FP*: 9	*TN*: 846

Table 4
Effectiveness of the LLM-generated categorization

Metric	Precision	Sensitivity	F1 score	Accuracy
Value	0.852	0.675	0.754	0.964

Discussion

The effectiveness of LLMs in categorizing the questions highlights the potential of using the tool in automating the self-reflection of teaching practice in Singapore schools. This is especially so as no machine learning pre-training and testing were done and only the 'of-the-shelf' gpt-4o LLM was used, saving processing time overall and achieving positive categorization results. Manual labeling took approximately three hours for one lesson, while the LLM completed the task in less than an hour, demonstrating significant time efficiency. This promising initial step demonstrates how leveraging GenAI can contribute to professional development. With these automated analyses, data-driven feedback can be provided to teachers to stimulate the reflective process, enabling them to identify areas for improvement and refine their instructional strategies. However, implementing the use of LLMs is not without challenges. One of the key challenges pertains to data privacy and security. Teachers must be assured that their data is being used ethically and securely, with appropriate measures in place to protect their privacy. Another challenge is associated with the potential resistance to change. Introducing new technologies into established educational institutions may be met with skepticism or

resistance from educators who are accustomed to traditional methods. It is essential to involve teachers in the development and implementation process, making sure that they find LLMs easy to use (McCoy & Shih, 2016). There is thus a need for professional development programs to help teachers develop the necessary competencies to utilize LLMs effectively.

In terms of areas for future research, the study highlights the need for further investigation into the long-term impact of LLMs-driven reflective practices on teacher practice and ultimately student outcomes. Longitudinal studies could provide valuable insights into how continuous use of LLMs influences teaching efficacy and professional growth over time. Additionally, expanding the scope of the study to include a broader range of subjects, educational levels, and teaching areas would help to generalize the findings, and provide a more comprehensive understanding of the benefits and challenges associated with the use of LLMs in education.

Conclusion

This study demonstrated the effectiveness of LLMs in analyzing classroom lessons in the teaching area of using questions to deepen learning and promote student engagement. It presents a promising avenue for automating the self-reflection of teaching practice in Singapore schools. The data-driven feedback received can encourage reflective practices and foster professional development. Addressing the challenges associated with data privacy, data security, and resistance to change would be crucial to ensuring that LLMs is used in a way that supports and empowers educators. Future work should expand the study to include other subjects, educational levels, and teaching areas, and utilize longitudinal studies to further validate LLM's effectiveness.

References

Arteaga, P., Batanero, C., Contreras, J. M., & Cañadas, G. R. (2015). Statistical graphs complexity and reading levels: A study with prospective teachers. *Statistique et Enseignement, 6*(1), 3–23. https://csbigs.fr/index.php/StatEns/article/view/430

Cohen, L., Manion, L., & Morrison, K. (2018). Research Methods in Education (8th ed.). London: Routledge. https://doi.org/10.4324/9781315456539

Hairon, S. (2020). Back to the future: Professional learning communities in Singapore. *Asia Pacific Journal of Education, 40*(4), 501–515. https://doi.org/10.1080/02188791.2020.1838880

Hatton, N., & Smith D. (1995). Reflection in teacher education: Towards definition and implementation. *Teaching and Teacher Education, 11*(1), 33–49. https://doi.org/10.1016/0742-051X(94)00012-U

Ku, O., Liang, J.-K., Chang, S.-B., & Wu, M. (2018). Sokrates Teaching Analytics System (STAS): An automatic teaching behavior analysis system for facilitating teacher professional development. In J. C. Yang, M. Chang, L.-H. Wong, & M. M. T. Rodrigo (Eds.), *ICCE 2018: 26th International Conference on Computers in Education Main Conference Proceedings* (pp. 696–705). Asia-Pacific Society for Computers in Education.

Mahon, P., & O'Neill, M. (2020). Through the looking glass: The rabbit hole of reflective practice. *British Journal of Nursing, 29*(13), 777–783. https://doi.org/10.12968/bjon.2020.29.13.777

McCoy, C., & Shih, P. (2016). Teachers as producers of data analytics: A case study of a teacher-focused educational data science program. *Journal of Learning Analytics, 3*(3), 193–214. https://doi.org/10.18608/jla.2016.33.10

Miller, K., & Shifflet, R. (2016). How memories of school inform preservice teachers' feared and desired selves as teachers. *Teaching and teacher education, 53*, 20–29. https://doi.org/10.1016/j.tate.2015.10.002

Ministry of Education (MOE). (2023, July 27). *Our teachers*. https://www.moe.gov.sg/education-in-sg/our-teachers

Ndukwe, I. G., & Daniel, B. K. (2020). Teaching analytics, value and tools for teacher data literacy: A systematic and tripartite approach. *International Journal of Educational Technology in Higher Education, 17*(22). https://doi.org/10.1186/s41239-020-00201-6

Wysel, M. (2023). Frenemies - Unleashing the power of ChatGPT in assessments. In T. Cochrane, V. Narayan, C. Brown, K. MacCallum, E. Bone, C. Deneen, R. Vanderburg, & B. Hurren (Eds.), *Proceedings of ASCILITE 2023:*

ASCILITE 2024

Navigating the Terrain:

Emerging Frontiers in Learning Spaces, Pedagogies, and Technologies

People, partnerships and pedagogies (pp. 614–618). Australasian Society for Computers in Learning in Tertiary Education. https://doi.org/10.14742/apubs.2023.653

Abidi, A., Ali, F., Quek, C.L.G., & Koh, R.E. (2024). Using LLMs to support teacher reflections on using questions to deepen learning and promote student engagement. In Cochrane, T., Narayan, V., Bone, E., Deneen, C., Saligari, M., Tregloan, K., Vanderburg, R. (Eds.), *Navigating the Terrain: Emerging frontiers in learning spaces, pedagogies, and technologies*. Proceedings ASCILITE 2024. Melbourne (pp. 447-452). https://doi.org/10.14742/apubs.2024.1195