Chan, A. et al. (Eds.) (2025). Proceedings of the 1st International Conference on Learning Evidence and Analytics. Asia-Pacific Society for Computers in Education

Unlocking student voices: using large language models to analyse open-ended survey responses at scale

Albert CHAN[a*], Ada TSE[a], Johnny YUEN[a] & Chun Sang CHAN[a]
[a]*Educational Development Centre, The Hong Kong Polytechnic University, HKSAR CHINA*
*ak.chan@polyu.edu.hk

Abstract: This paper overviews a project using large language models (LLMs) to analyse open-ended responses from a student learning survey. Traditional qualitative analysis is time-consuming, but LLMs can quickly process large volumes of text while preserving detail. The project applies LLMs to classify diverse student comments, aiding understanding of student experiences and improving reporting speed. This reduces manual effort and supports timely, evidence-based decisions. The paper also explores practical insights and challenges encountered, including prompt optimisation and the validation of AI-generated classifications to support meaningful insights for targeted follow-up actions.

Keywords: LLM, Open-ended survey, Educational data analysis, Text classification

1. Introduction

Analysing open-ended student feedback is vital for understanding learning experiences and guiding institutional improvements, but manual coding is time-consuming and prone to subjective bias (Stadel et al., 2025). Advances in AI, especially large language models (LLMs), provide new methods for qualitative analysis in education. However, LLM outputs require validation against human judgment to mitigate algorithmic bias and hallucinations, ensuring reliability (Almatrafi et al., 2024). Ethical issues such as data privacy and transparency also remain critical as AI adoption grows (Shi et al., 2024). This paper examines using LLMs to categorise institution-wide student survey responses, highlighting challenges and practical insights to inform evidence-based educational practices.

2. Practical Implementation

The analysis focused on open-ended responses collected from the institutional student learning experience survey in 2024/25, targeting first-year undergraduate students. The dataset comprised 385 comments from students addressing major problems or barriers they encountered in their studies, and 505 comments providing suggestions on how the student learning experience could be improved (Table 1). To classify these responses while ensuring data confidentiality, the project employed the QWQ-32B model from Ollama, an open-source large language model developed by Alibaba. This reasoning model, with 32 billion parameters, offers performance comparable to larger state-of-the-art models like DeepSeek-R1 but uses fewer resources, making it ideal for local deployment. The model was run within a Python 3.9.10 environment using Ollama's local inference framework, with quantisation techniques applied to optimise memory usage to approximately 19.85GB without

compromising accuracy. The implementation of the QWQ-32B model was carried out on a desktop computer equipped with an NVIDIA RTX A5000 graphics card and 64GB of RAM. By processing all data locally, the approach safeguarded student privacy while leveraging the model's advanced reasoning capabilities to categorise and interpret the qualitative feedback effectively.

Survey Target	Survey Question	Number of comments collected
First Year Ug student	Please indicate major problems/barriers, if any, that you have encountered in your study at PolyU.	385
First Year Ug student	Please suggest how the student learning experience at PolyU can be improved.	505

Table 1. Open-ended questions asked in the survey and the numbers of comment collected.

To enable effective classification of survey comments, the input labels are organised in a row-wise format using Python. This arrangement allows the LLM to accurately reference the labels and identify the relevant categories for each comment. The process involved assigning one or more labels to each response based on predefined categories, which were then incorporated into the prompts given to the LLM for analysis. See Table 2.

Label format:
*Row 0: **Description of category** description of label category \n **category** label category*

Table 2. Label format for data preparation process.

Prompt was designed to enhance the accuracy of the LLM by including descriptions of each category within the label set. This approach enabled the model to analyse and classify each comment according to the most relevant category. The structured prompt was formulated as in Table 3. Embedding the category descriptions directly into the prompt guided the LLM to take into account the specific features of each category during classification. This approach improved the model's ability to accurately assign comments to the correct categories by combining a structured input format with detailed category information.

```
prompt = f"""
Analyze this Excel data and answer my question:
{file_content}

You are an assistant tasked with classifying comments based on student learning
experience.
According to description of category in {file_content} and analyze which category
{file_content} that provided text is mentioned. It could be more than one category.

Comment: {comment}

Please respond ONLY with "Category:"

"""
```

Table 3. The prompt design.

3. Key Insights

Experiment was done to allow the LLM to perform unsupervised learning, letting it define category names. While results were similar to predefined categories, using predefined categories speeds up processing and better suits administrative needs by aligning responses with specific offices for efficient follow-up.

The QWQ model effectively classifies comments using customised keywords, enabling researchers to create specialised labels with unique terms or abbreviations relevant to the university context. For example, comments containing keywords like "WIE" (Work-integrated Education) or "Exchange" are accurately classified by the LLM under the category related to internships or placements. This category encompasses terms such as Internship, Placement,

WIE, Exchange, and Study tour, demonstrating the model's ability to associate diverse but related keywords with the appropriate label.

Validation against human coding showed 60-70% agreement. Accuracy was higher for comments with single labels than multiple labels, highlighting the need to refine methods for multi-label classification. Observations of comments with multiple labels also reveal that the model performs best when the content corresponding to each label is clearly separated. Using commas or conjunctions to divide different parts of a comment helps in achieving this. An excerpt of student comments and their corresponding labels assigned by the LLM are shown in Table 4.

In addition, when the LLM assigns an incorrect label to a comment, researchers can review the model's internal reasoning - often referred to as its "thought process", "self-reflection", "internal dialogue" or sometimes "reasoning thinking part", to understand the decision-making process. By examining this reasoning, researcher can adjust the label inputs or modify the prompt, thereby fine-tuning the LLM to improve its answers. It is recommended that researcher utilise a reasoning model in conjunction with the output from the "reasoning thinking part" to gain deeper insights into how the LLM interpreted the prompt and generated its classification.

Comment	Label classified by LLM
量化學生的學業壓力，增加興趣學會以及學生活動，提供中英雙語電郵，減少測驗，優化複習教材	Workload/ Time management, Social life, Language issue in teaching environment, Assessment, Teaching quality
多建宿舍和教室	Teaching and Learning Facility, Other facility

Table 4. Examples of comment categorised by LLM.

4. Conclusion

Using large language models like QWQ-32B to classify open-ended survey comments provides an efficient way to manage qualitative data. Embedding category descriptions improves precision. While challenges with multi-label comments and misclassifications remain, the model's internal reasoning aids prompt refinement and performance enhancement. Combining this with a reasoning model further supports tuning.

Fine-tuning the model can be costly, requiring careful budget planning. Nonetheless, this approach streamlines the time-consuming coding process. It offers a promising solution for educational stakeholders to extract meaningful insights from large qualitative datasets while maintaining data confidentiality through local processing.

References

Almatrafi, O., Johri, A., & Lee, H. (2024). A systematic review of AI literacy conceptualization, constructs, and implementation and assessment efforts (2019-2023). *Computers and Education Open*, 100173. https://www.sciencedirect.com/science/article/pii/S2666557324000144

Shi, Lehong, and Ikseon Choi, 'A Systematic Review on Artificial Intelligence in Supporting Teaching Practice: Application Types, Pedagogical Roles, and Technological Characteristics', in Xiaoming Zhai, and Joseph Krajcik (eds), Uses of Artificial Intelligence in STEM Education (Oxford, 2024; online edn, Oxford Academic, 21 Nov. 2024), https://doi.org/10.1093/oso/9780198882077.003.0015, accessed 17 Apr. 2025.

Stadel, M., Langener, A. M., Hoemann, K., & Bringmann, L. F. (2025). Assessing daily life activities with experience sampling methodology (ESM): Scoring predefined categories or qualitative analysis of open-ended responses?. *Methods in Psychology*, *12*, 100177. https://www.sciencedirect.com/science/article/pii/S2590260125000037