From Transcripts to Themes: A Trustworthy Workflow for Qualitative Analysis Using Large Language Models*

Aneesha Bakharia[1,*], Antonette Shibani[2], Lisa-Angelique Lim[2], Trish McCluskey[3] and Simon Buckingham Shum[2]

[1] The University of Queensland, Australia
[2] University of Technology Sydney, Australia
[3] Deakin University, Australia

Abstract

We present a novel workflow that leverages Large Language Models (LLMs) to advance qualitative analysis within Learning Analytics, addressing the limitations of existing approaches that fall short in providing theme labels, hierarchical categorization, and supporting evidence, creating a gap in effective sensemaking of learner-generated data. Our approach uses LLMs for inductive analysis from open text, enabling the extraction and description of themes with supporting quotes and hierarchical categories. This trustworthy workflow allows for researcher review and input at every stage, ensuring traceability and verification, key requirements for qualitative analysis. Applied to a focus group dataset on student perspectives on generative AI in higher education, our method demonstrates that LLMs are able to effectively extract quotes and provide labeled interpretable themes compared to traditional topic modeling algorithms. Our proposed workflow provides comprehensive insights into learner behaviors and experiences and offers educators an additional lens to understand and categorize student-generated data according to deeper learning constructs, which can facilitate richer and more actionable insights for Learning Analytics.

Keywords

LLM, Generative AI, Qualitative analysis, Inductive coding, Education, Learning Analytics

1. Introduction

Qualitative, or sometimes ethnographic, interpretative research has emerged as a valuable approach in educational studies, providing deep understanding and rich insights into the complex nature of learning environments and experiences. Qualitative research, although having many definitions, is ”a form of social inquiry that tends to adopt a flexible and data-driven research design, to use relatively unstructured data, to emphasize the essential role of subjectivity in the research process, to study a small number of naturally occurring cases in detail, and to use verbal rather than statistical forms of analysis” [1]. This is in contrast to quantitative research, where data can be quantified, generally arising from large, representative samples from a target population and analysed through statistical procedures [2]. While Artificial Intelligence (AI) can spot patterns and handle numerical data, often more efficiently than humans for quantitative analysis, its capabilities for qualitative analysis, traditionally considered deeply human-oriented and values-based, remain under-explored.

In Learning Analytics (LA), prior research has employed qualitative analysis methods to study student perspectives [3, 4] and educator perspectives [5] in case studies of LA tools and interventions. This involves the analysis of in-depth interviews, focus groups, and/or field observations, allowing researchers to explore participant’s perspectives in a naturalistic setting. While quantitative analysis of data is thought to be more objective and systematic, the value of qualitative analysis of data lies in its ability to capture the richness of human experiences, emotions, and social contexts. The

* *Joint Proceedings of LAK 2025 Workshops, co-located with the 15th International Conference on Learning Analytics and Knowledge(LAK 2025), Dublin, Ireland, March 03–07, 2025.*
* Corresponding author
a.bakharia1@uq.edu.au (A. Bakharia); antonette.shibani@uts.edu.au (A. Shibani); lisa-angelique.lim@uts.edu.au (L. Lim); trish.mccluskey@deakin.edu.au (T. McCluskey); simon.buckinghamshum@uts.edu.au (S. Buckingham Shum)

approach is particularly relevant to the field of education, as it helps to understand the nuances of teaching and learning experiences and is often used in combination with other approaches.

The rich, descriptive data also poses challenges in data analysis, with several automated approaches developed to support qualitative analysis. Parts of the thematic analysis process are to extract key patterns (themes) from interview transcripts and open-ended responses were supported by unsupervised approaches such as topic modeling, clustering, and visualisations [6, 7], and supervised techniques such as predictive modelling helped researchers semi-automate 'coding' that helps break down large quantities of data into manageable chunks for interpretation and categorization [8, 9]. However, traditional approaches were limited in reasoning and only performed a surface-level analysis compared to what a human analyst would produce. With Generative Artificial Intelligence (GenAI) applications using Large Language Models (LLMs) now offering a broad range of capabilities that can be applied to various tasks, recent works explore how LLMs can be used to support qualitative analysis in more effective ways.

The aim of our study is to investigate how LLMs can aid qualitative research with in-built verification mechanisms that offer transparent, trustworthy, and supervisable outputs. In our work, we explore if LLMs can be used in thematic analysis with no pre-defined codes to look for (a bottom-up, inductive approach), where the LLM must pick key themes that emerge from the data. We define minimum requirements that an LLM-supported qualitative analysis workflow must satisfy in order to be deemed reliable and trustworthy to a researcher performing the analysis. We present an LLM-based workflow that derives an initial set of themes from given text data, while maintaining traceability of original sources. The workflow satisfies the minimum requirements we set, and is transferable to other LA contexts performing qualitative analysis of open-ended text. Our prototype tool, which results from the workflow, enables the researcher to extend LLM outputs.

2. Using Large Language Models for Qualitative Analysis

An increasing number of studies are showcasing how GenAI, specifically LLMs, can be integrated into the qualitative analysis process to aid researchers, while also discussing its potential pitfalls and ethical issues. Within LA, LLMs offer solutions to the challenges in processing unstructured text data and analysing these for insights and meaningful indicators of student learning - this is a key phase in the learning analytics cycle [10]. The majority of LLMs' use for qualitative coding/ content analysis has focused on deductive analysis, where the human analyst sets out a gold standard by coding data themselves first (with reliability tested across multiple human raters). For example, Xiao et al [11] describe how they employed an LLM-based approach combining GPT-3 with expert-developed codebooks, testing two design dimensions: codebook-centred vs example-centred; and different example centred designs (i.e., zero-shot vs one-shot vs few-shots). They found that codebook-centred designs outperformed example-centered designs, and that the provision of examples was an important factor in the model's performance. Recognising the importance of providing examples, Hou et al [12] examined the use of GPT for deductive coding of social annotations, and found that fine-tuning the LLM model with more than 100 examples was instrumental in increasing reliability, especially for some of the codes being examined. However, while such outcomes are promising for the use of LLMs for deductive analysis, possibly different approaches might be needed for inductive analyses, which are more 'bottom up' and exploratory [11]. When it comes to inductive analysis, the role of the (human) researcher is critical in drawing on relevant experience as well as creativity, to interrogate textual data and bring out both semantic and latent meanings; however, GenAI tools mainly rely on the input data and therefore there may need more considerations to be able to tease out these meanings [13].

Recent work has explored the design of prompts to leverage LLMs for the kind of inductive analysis described above. One example is illustrated by Rao et al.'s QuaLLM framework [14], comprising a four-step prompting process. It begins with a generation stage, akin to open coding, carried out by the LLM. The next classification stage involves humans in the loop, whereby the researcher identifies four to five primary themes based on existing evidence, together with the

outputs of the generation stage, which they consider similar to thematic analysis. The third and fourth steps - aggregation prompt and prevalence prompt - are conducted entirely by the LLM, and are designed to gain quantitative insights. Their framework builds in evaluation checks within each prompt step; for example, in the generation prompt, the output is evaluated for completeness by asking the LLM to ensure that all relevant concerns are present. Pham et al. [15] approached the challenge somewhat differently by developing their TopicGPT analysis tool leveraging LLM to augment the capabilities of topic modeling approaches such as Latent Dirichlet Allocation (LDA). The tool works in three stages: firstly, the model generates new topics identified in an input dataset, with humans in the loop to refine topics. The second stage is the assignment stage which comprises two sub-steps. The LLM is provided with the refined topic list created by, and of interest to, the researchers, with a few examples. The next sub-step in this stage involves the researcher prompting the model to assign topics as appropriate, to the input data. This provides "a valid and interpretable association between the generated topics and the documents in our datasets" (p.3). The model then provides the output in the form of an assigned topic and the topic description, together with a supporting quote from the document. In the third, self-correction, stage, a built-in parser flags hallucinations or wrongly-assigned topics, which is then fed back to the LLM with a prompt to rectify the assignment. From these examples, it can be seen that quality outputs from LLM for inductive qualitative analysis can be prompted by building in the following steps into a prompt framework: 1) having humans in the loop to "vet" the themes; 2) multiple iterations; 3) including metrics for evaluation; and 4) chain-of-thought prompting for the LLMs.

Drawing from prior work but extending to a fully automated workflow from data to themes, we define the following requirements for LLM-generated inductive coding that need to be satisfied for trustworthy LLM-based tools, which we have yet to see implemented in the current state-of-the-art:

- **Requirement 1: To maintain the integrity of coded textual extracts:** (i) verify against the source data that quotes are verbatim and not hallucinated, and (ii) verify that they are meaningfully classified under the assigned code.
- **Requirement 2: To maintain the transparency of the coding:** (i) explain the rationale for each code, and (ii) trace every code, whatever level of abstraction, back to its source data.

3. Methodology

This section presents the methodology employed in developing and evaluating our proposed workflow for integrating Large Language Models (LLMs) into qualitative inductive analysis, and allowing for custom deductive analysis if required. The dataset we used comprises transcripts from twenty (20) focus groups conducted across four universities. These focus groups were part of a more extensive study exploring learner experiences with GenAI in higher education, with ethics approval for using LLM models for analysis. Participants came from diverse backgrounds, and studied courses at various levels in the four universities. The transcripts provided rich qualitative data suitable for thematic analysis with additional metadata, such as the university and focus group number. While analysis of this data provides interesting insights into student perspectives on Generative AI in higher education, the results of the qualitative study are covered elsewhere [16]. The purpose of the current study is to examine the potential of LLM-based workflows to support the qualitative research process for LA researchers. While it presents some outputs in the form of visualisations and tool screenshots, note that this paper does not discuss the insights from the data itself in detail.

Development of the workflow followed an iterative and collaborative design-based research (DBR) approach [17], involving continuous refinement starting from initial prototyping in the LLM prompt playground, to a Jupyter notebook environment, and finally to a more refined, interactive web tool prototype. This approach is suitable for our study as it allows for the practical investigation of LLM capabilities in thematic analysis and supports the co-construction of knowledge among researchers. A group of researchers (with data science, education, and HCI backgrounds) collaborated weekly over Zoom meetings to explore the use of LLMs for inductive and deductive

thematic analysis. During these sessions, they tested prompts using two large language models: GPT-4-32K (via Azure)[1] and Claude Sonnet 3.5 (via AWS)[2]. The models were accessed via APIs enabled through secure institutional access as the data set contained sensitive student data - no public APIs or LLM tools were used for data processing.

The workflow aims to enhance the extraction, validation, and visualization of themes from learner generated textual data while keeping the researcher at the center of extracting insights from the data. It addresses the limitations of traditional methods and considers the limitations of LLMs while emphasizing verification through researcher input, evidence gathering, and sensemaking. The workflow is iterative and modular, allowing adjustments based on data characteristics and analysis objectives. To verify and justify key decisions in designing the workflow (in addition to inbuilt verification mechanisms), the following evaluations are conducted:

- LLM Performance Comparison: Assessing the performance of GPT-4, Claude Sonnet 3.5 and open-source models in quote extraction.
- Theme Interpretation: Comparing the theme output with traditional topic modeling algorithms like None Negative Matrix Factorization to highlight improvements in theme relevance and interpretability.

4. The LLM-Aided Thematic Analysis Workflow

This section outlines the designed workflow for integrating LLMs into qualitative inductive and deductive thematic analysis (Figure 1). The workflow is built using Python, comprising of a Jupyter notebook[3] and a Flask application[4]. Figure 1 shows workflow components and the flow of outputs between each step. Associated LLM prompts are provided in the Appendix.

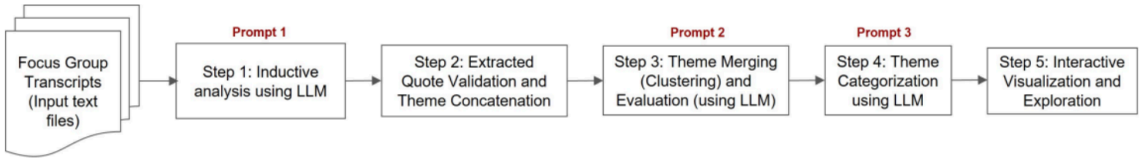

Figure 1: Workflow of steps in the automated pipeline for qualitative analysis using LLM

4.1. Step 1: Inductive Analysis using LLM

Within this step, an LLM is prompted to perform initial theme extraction, labelling, rationale generation, and verbatim quote extraction from qualitative data sources. Thematic analysis is performed on individual datasets (i.e., text from each focus group transcript in this case) with additional metadata linked to the dataset. The process involves prompting the LLM to identify significant statements or quotes that represent key ideas, group similar statements to form preliminary themes, and label each theme with a concise and descriptive title, rationale, and list of defining keywords.

Due to the context size limitations of LLMs, the analysis is conducted on smaller subsets of data, such as individual focus group transcripts. This approach ensures that the LLM can process the input effectively without truncation or loss of context. Processing individual transcripts helps capture all important themes from every discussion without the risk of over-summarization/ merging at the

[1] https://azure.microsoft.com/en-us/blog/introducing-gpt4-in-azure-openai-service/
[2] https//aws.amazon.com/bedrock/claude/
[3] https//jupyter-notebook.readthedocs.io/en/latest/
[4] https://flask.palletsprojects.com/en/3.0.x/

next step. It also ensures traceability up to the source level, so the researcher can refer back to the original instances in the transcript for an enhanced understanding of the theme.

4.2. Step 2: Extracted Quote Validation and Theme Concatenation

In this step, the quotes are validated to ensure that they exist in the original dataset and are not a hallucination by the LLM. A fuzzy string matching algorithm is also used to match instances where small punctuation differences exist. As inductive themes are found for individual data sources, themes are concatenated into a single file with metadata.

4.3. Step 3: Theme Merging and Evaluation

As the initial theme finding was performed on individual datasets, the concatenated list of themes contains multiple semantically similar or overlapping themes. We first used a prompt to get the LLM to merge semantically related themes, but found that approximately a third of the themes were not included by both LLMs being evaluated. In this step a soft clustering algorithm, namely Non Negative Matrix Factorization is used to find clusters of semantically related themes while supporting overlap. NMF is chosen due to its ability to handle overlapping themes and provide an additive, parts-based representations of the data [18]. The number of clusters is chosen based on the maximum topic coherence metric. An LLM is used to validate the clustering as well as label and provide a rationale for the merged theme. The LLM has the ability to remove a theme from a cluster if it is not semantically related.

4.4. Step 4: Theme Categorization

An LLM is used to further group together the merged themes to provide another higher level of categorization. The refined themes are grouped into broader categories, forming a hierarchical structure that reflects the relationships among concepts in the data. This structure aids in the interpretation of the findings and supports the development of actionable insights.

4.5. Step 5: Exploration and visualization with Sankey Diagrams

An interactive Sankey diagram visualizes the flow and relationships between themes, categories, and source files. It enhances interpretability and shows pathways back to additional metadata (e.g., the individual focus group or cohort) - See Figure 2. The benefits of using Sankey diagrams include:

- Clarity: Providing a clear visual representation of how themes aggregate into categories.
- Traceability: Allowing analysts to trace specific quotes from source files through themes to categories.
- Integration of Metadata: Incorporating additional information such as focus group identifiers, locations, or demographic data.

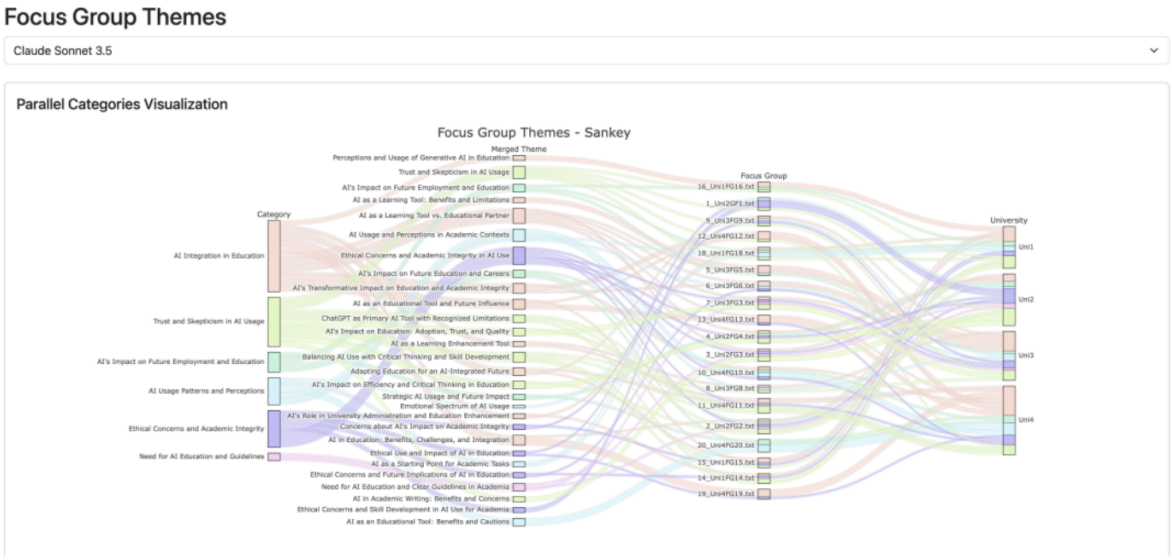

Figure 2: An interactive Sankey Diagram visualises the flow of data from source *Universities* and their *Focus Group Transcripts*, to *Merged Themes*, to higher order *Categories*.

The full output from themes found in individual datasets to merged themes and higher-level categories can be explored along with extracted quotes (See Figure 3). The interactive application also provides a search option to find text within the selected transcripts for further reading and lookup the exact location of a quote within the transcript. A drop-down enables switching between LLMs (E.g. Claude vs GPT-4).

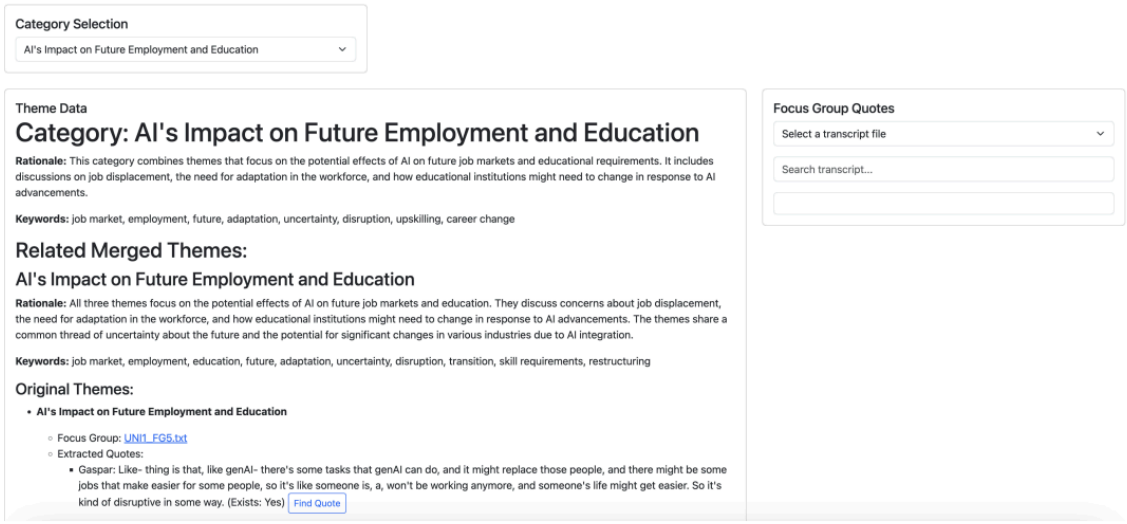

Figure 3: An interactive user interface enables exploration of themes and quotes, tracing them back to their original source files

5. Evaluation

The evaluation primarily focuses on core aspects of traceability and interpretability within the proposed workflow. This involves assessing the capabilities of large language models (LLMs) to accurately extract verbatim quotes for a theme and comparing the ease of interpretation of themes extracted by LLMs against those generated by traditional topic modeling algorithms. Our findings

indicate that level 4 LLMs such as GPT-4 and Claude Sonnet 3.5 have advanced capabilities. However, when experimenting with several smaller open-source models like Llama 3 and Mistral 7B, we encountered difficulties in implementing quote extraction through prompt engineering.

From our analysis using focus group datasets, GPT-4 exhibited an average of 8.3% invalid verbatim quotes, while Claude Sonnet 3.5 displayed slightly better performance with an average of 7.7%. It's crucial to note that these invalid quotes are not hallucinations. As illustrated in Figure 4, adjustments in the fuzzy search threshold reveal that minor discrepancies such as punctuation errors or the omission of filler words (e.g., 'like') used by the focus group participants indicate close matches, though not verbatim. A SequenceMatcher string comparison algorithm that identifies the longest contiguous matching subsequence between two strings was used in the analysis. This arises because the LLM regenerates the quote from the provided textual context.

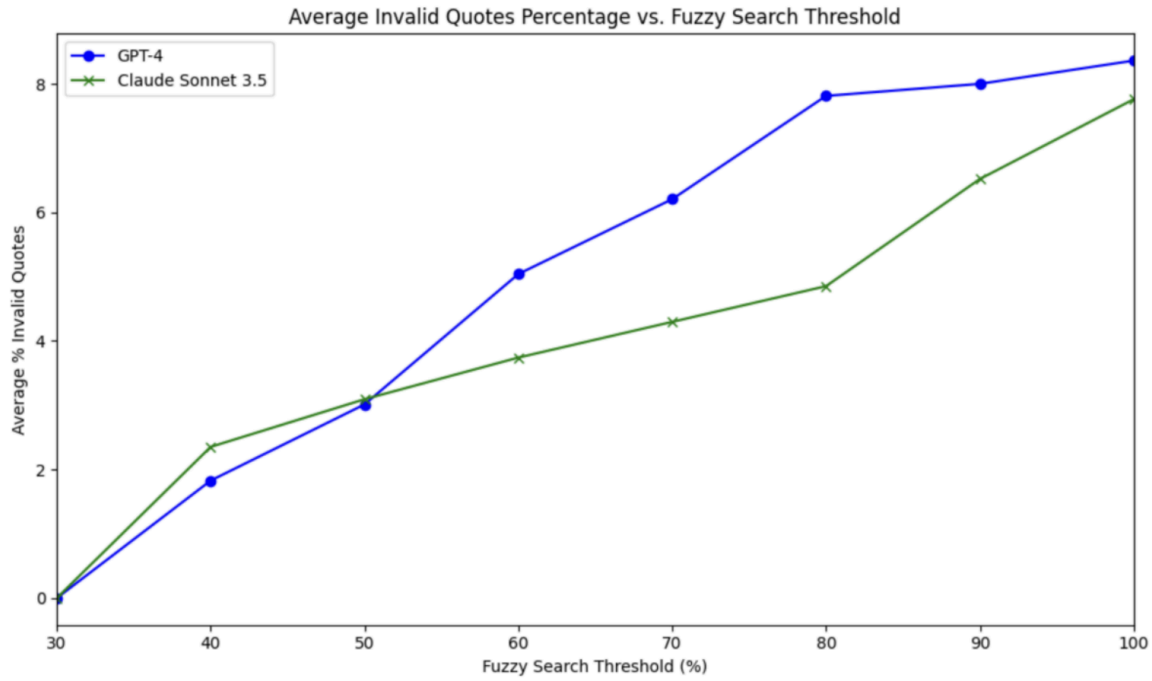

Figure 4: Impact of fuzzy search threshold reduction on the percentage of invalid extracted quotes.

Addressing the question of whether themes extracted by LLMs are easier to interpret compared to those from traditional topic modeling algorithms is more complex. Through targeted prompting, LLMs can perform tasks beyond the capabilities of conventional algorithms, such as labeling, crafting rationales, and embedding descriptive keywords into the extraction process. In contrast, traditional topic modeling techniques, including NMF and LDA, were less effective, particularly on the focus group dataset. For instance, as demonstrated in Table 1, Claude Sonnet 3.5 was able to produce more coherent and contextually relevant theme labels and keywords compared to the generic outputs from NMF. Our trials with different data cleaning methods and text structuring (e.g., providing complete focus group texts versus splitting texts by speaker) further highlighted the superior flexibility and adaptability of LLMs, which could dynamically adjust to instructions such as removing facilitator contributions by providing their names.

Table 1
Comparison of selected topics and main keywords between Claude Sonnet 3.5 and NMF

Claude Sonnet 3.5	NMF
Trust and Skepticism in AI Usage: trust, skepticism, verification, accuracy, reliability, critical thinking, ethics, academic integrity, fact-checking, human oversight	trust, chat, response, interesting, stuff, teacher, datum, level, type, moment

AI's Impact on Future Employment and Education: job market, employment, education, future, adaptation, uncertainty, disruption, transition, skill requirements, restructuring	profession, definition, article, draft, evaluate, solely, perform, graduate, pretty, judge
AI as a Learning Enhancement Tool: learning, acceleration, tool, understanding, writing skills, feed- back, refining, exploring	assessment, read, change, article, sum- marize, education, undergrad, blah, uni, mention

6. Discussion

Our study presents an LLM-based workflow for performing qualitative inductive and deductive thematic analysis using a trustworthy, human-centred approach with a researcher in the loop. While we demonstrate the approach using an educational data set examining student perspectives on generative AI, the workflow has already been adapted to the analysis of textual data from other contexts. Although some recent work using large language models highlights the potential of LLMS to enhance the efficiency and efficacy of qualitative analysis [19], our aim is not to simply reduce human effort and time spent on the process. We want to leverage the potential of LLMs to gain deep insights into the data, and the improved efficiency is a by-product that can offer the researcher new, diverse ways to engage in a deeper exploration of the data (For example, viewing from a different theoretical angle, or a specific research question). This aligns with the human-centered view of using AI as a thinking companion, potentially as a partner in cognition to augment their thinking [20]. The in-built validations in our approach improve the reliability and verification of using LLMs in the qualitative data analysis process, while meeting the requirements we set out for quality assurance.

While our workflow for automating qualitative thematic analysis is made generic with no specific lenses/ human thinking built into it, it is important to note that parts of our workflow do embed decisions made by the research team who created it. This includes the crafting of prompts, selection of the clustering algorithm and its parameters, and the choice of LLMs. Changes in some, or all of these, are likely to change the results of the analysis. However, our aim is not to come up with absolute findings that remain the same for any given set of data through a fully automated analysis pipeline, but to aid researchers in their analysis process so they can then bring their specific lenses to drill into insights. Though the LLMs might lack consistency in their outputs, so do humans, in their subjective interpretation of qualitative data. This is a well-recognised phenomenon in qualitative research, and is not necessarily a limitation, but a characteristic of the process. Robustness of findings and research integrity are as important as other forms of research, but there is a preference for 'verification' rather than 'validity' and 'reliability', which are deemed contentious among qualitative researchers [21]. With the lack of 'gold standard' coding in thematic analysis of open-ended text, our approach presents a way to verify results using methods described above.

Ethical questions remain, and we acknowledge that there may not be one right answer to this evolving dilemma. Engaging with the data through reading and re-reading transcripts in its most human form is a useful process, and automating parts of it might take away important elements of qualitative research that build the capabilities of researchers, leading to an undesirable outcome in the long run. The LLM might be incapable of capturing latent meaning that is teased out by the human [13]. Introducing an LLM might also inadvertently steer the results in a direction that is different from what a researcher might have taken by themselves, potentially leading to homogeneity of results and a reduction in human intuition. On the other hand, as LLMs improve in their capabilities (such as the newest LLM o1 released by OpenAI that can perform reasoning tasks[5]), many attributes previously thought to be uniquely human may not remain that way and can be

[5] https//www.wired.com/story/openai-o1-strawberry-problem-reasoning/

attained to a reasonable degree by AI. In these cases, we believe it is best to work with AI to augment human intelligence, rather than working against it.

7. Limitations and future work

We have developed a fully working prototype of an automated qualitative analysis pipeline, but it requires some technical knowledge and setup before it can be used. Key limitations in the current workflow that we hope to resolve in future work are below:

Usability for qualitative researchers. We are working towards a user-friendly web application that allows anyone to create thematic analysis outputs by uploading a set of input text files. We are testing its usefulness and usability with researchers.

LLM bias. We also acknowledge that biases may exist in LLM outputs due to their training data. Our future work will involve implementing additional checks to identify and correct biased or inappropriate content generated by the LLMs to minimise the propagation of biases.

Validating the coverage of themes. The workflow ensures that the LLM generated themes are traceable back to the original quotes to ensure trustworthy outputs. However, the lack of comparable human analysis in inductive coding meant that we cannot ensure that the themes have comprehensive coverage; i.e, all relevant themes and quotes have been captured from the data set. Future work will include human-AI comparative analysis to identify gaps and opportunities for LLM analysis. This way, researchers can use such LLM workflows to complement their qualitative analysis process in appropriate ways by noting their limitations.

8. Conclusion

Our study showcased a novel workflow that integrates the strengths of LLMs with traditional topic modeling techniques to aid qualitative analysis in learning analytics. It supports inductive analyses to derive themes from open text data, providing a more comprehensive understanding of rich learner generated data. By incorporating validation processes and visualization techniques, the approach improves transparency, verifiability, and interpretability, while addressing limitations of previous methods and enhancing researcher processes in qualitative thematic analysis. Preliminary findings show that LLMs can be used to augment researchers' qualitative analysis workflows, and future work will explore the nuances of practical usage along with Human vs AI validations.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 (via Azure) and Claude Sonnet 3.5 (via AWS) to assist with theme extraction, labeling, and quote identification for qualitative data analysis. After using these tools, the authors reviewed and edited all outputs as needed and take full responsibility for the publication's content.

References

[1] M. Hammersley, What is qualitative research?, Bloomsbury Academic, 2012.

[2] A. Queirós, D. Faria, F. Almeida, Strengths and limitations of qualitative and quantitative research methods, European journal of education studies (2017).

[3] C. Schumacher, D. Ifenthaler, Features students really expect from learning analytics, Computers in human behavior 78 (2018) 397–407.

[4] L. Lim, S. Dawson, D. Gašević, S. Joksimović, A. Pardo, A. Fudge, S. Gentili, Students' perceptions of, and emotional responses to, personalised la-based feedback: An exploratory study of four courses, Assessment Evaluation in Higher Education 46 (2021) 339–359. doi:10.1080/02602938. 2020.1782831.

[5] A. Shibani, S. Knight, S. B. Shum, Educator perspectives on learning analytics in classroom practice, The Internet and Higher Education 46 (2020) 100730.

[6] A. Bakharia, P. Bruza, J. Watters, B. Narayan, L. Sitbon, Interactive topic modeling for aiding qual itative content analysis, in: Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval, 2016, pp. 213–222.
[7] B. Sherin, N. Kersting, M. Berland, Learning analytics in support of qualitative analysis, in: ICLS 2018 Proceedings, International Society of the Learning Sciences, Inc.[ISLS]., 2018.
[8] Z. Cai, A. Siebert-Evenstone, B. Eagan, D. W. Shaffer, X. Hu, A. C. Graesser, ncoder+: a semantic tool for improving recall of ncoder coding, in: International Conference on Quantitative Ethnography, Springer, 2019, pp. 41–54.
[9] A. Shibani, E. Koh, V. Lai, K. J. Shim, Assessing the language of chat for teamwork dialogue, Journal of Educational Technology & Society 20 (2017) 224–237.
[10] L. Yan, R. Martinez-Maldonado, D. Gasevic, Generative artificial intelligence in learning analytics: Contextualising opportunities and challenges through the learning analytics cycle, in: Proceedings of the 14th Learning Analytics and Knowledge Conference, LAK '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 101–111. URL: https://doi.org/10.1145/3636555.3636856. doi:10.1145/3636555.3636856.
[11] Z. Xiao, X. Yuan, Q. V. Liao, R. Abdelghani, P.-Y. Oudeyer, Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding, in: Companion Proceedings of the 28th International Conference on Intelligent User Interfaces, IUI '23 Companion, Association for Computing Machinery, New York, NY, USA, 2023, p. 75–78. URL: https://doi.org/ 10.1145/3581754.3584136. doi:10.1145/3581754.3584136.
[12] C. Hou, G. Zhu, J. Zheng, L. Zhang, X. Huang, T. Zhong, S. Li, H. Du, C. L. Ker, Prompt-based and Fine-tuned GPT Models for Context-Dependent and -Independent Deductive Coding in Social Annotation, Association for Computing Machinery, Kyoto, Japan, 2024, p. 518–528. URL: https://doi.org/10.1145/3636555.3636910. doi:10.1145/3636555.3636910.
[13] R. M. Davison, H. Chughtai, P. Nielsen, M. Marabelli, F. Iannacci, M. van Offenbeek, M. Tarafdar, M. Trenz, A. A. Techatassanasoontorn, A. Díaz Andrade, N. Panteli, The ethics of using generative ai for qualitative data analysis, Information Systems Journal 34 (2024) 1433–1439. URL: https: //onlinelibrary.wiley.com/doi/abs/10.1111/isj.12504. doi:https://doi.org/10.1111/isj.12504.
[14] V. N. Rao, E. Agarwal, S. Dalal, D. Calacci, A. Monroy-Hernández, Quallm: An llm-based framework to extract quantitative insights from online forums, arXiv preprint arXiv:2405.05345 (2024).
[15] C. Pham, A. Hoyle, S. Sun, P. Resnik, M. Iyyer, TopicGPT: A prompt-based topic modeling framework, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 2956–2984. URL: https://aclanthology.org/2024.naacl-long.164. doi:10.18653/ v1/2024.naacl- long.164.
[16] T. Fawns, M. Henderson, K. Matthews, G. Oberg, Y. Liang, J. Walton, T. Corbin, M. Bearman, S. B. Shum, T. McCluskey, et al., Gen ai and student perspectives of use and ambiguity: A multi-institutional study, in: Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education 2024: Navigating the Terrain: Emerging Frontiers in Learning Spaces, Pedagogies, and Technologies, 2024.
[17] S. Barab, K. Squire, Design-based research: Putting a stake in the ground, in: Design-based Research, Psychology Press, 2016, pp. 1–14.
[18] D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization, nature 401 (1999) 788–791.
[19] Y. Gamieldien, J. M. Case, A. Katz, Advancing qualitative analysis: An exploration of the potential of generative ai and nlp in thematic coding, Available at SSRN http://dx.doi.org/10.2139/ssrn.4487768 (2023).
[20] G. Salomon, D. N. Perkins, T. Globerson, Partners in cognition: Extending human intelligence with intelligent technologies, Educational researcher 20 (1991) 2–9.

[21] K. Hammarberg, M. Kirkman, S. De Lacey, Qualitative research methods: when to use them and how to judge them, Human reproduction 31 (2016) 498–501.