Vol.7 No.2 (2024)

Journal of Applied Learning & Teaching

ISSN : **2591-801X**

Content Available at : **http://journals.sfu.ca/jalt/index.php/jalt/index**

Exploring the synergy of human and AI-driven approaches in thematic analysis for qualitative educational research

Soroush Sabbaghan[A] [A] *Associate Professor, Werklund School of Education, University of Calgary, Canada*

Keywords

AI;
artificial intelligence;
educational research;
generative artificial intelligence;
qualitative analysis.

Correspondence

ssabbagh@ucalgary.ca [A]

Article Info

Received 2 July 2024
Received in revised form 24 August 2024
Accepted 27 August 2024
Available online 4 September 2024

DOI: https://doi.org/10.37074/jalt.2024.7.2.32

Abstract

This paper investigates the integration of Generative Artificial Intelligence (GenAI), particularly Large Language Models (LLMs) such as GPT-4, into qualitative analysis in educational research. Utilizing TRACER (Transcript Analysis and Concept Extraction Resource), a GenAI-driven tool, the study evaluated its efficiency, reproducibility, and synergy with human analytical expertise. The research demonstrated that TRACER significantly streamlined thematic analysis, efficiently handled large data volumes, and maintained consistency in theme identification. The findings reveal that integrating TRACER's computational power with human interpretive skills enriches research outcomes, suggesting a collaborative approach for optimal results. Despite its efficacy, limitations such as data scope and current GenAI capabilities are acknowledged, indicating areas for future development. This paper contributes to the understanding of GenAI's role in qualitative research, proposing it as a valuable tool for overcoming traditional challenges in the field and highlighting the importance of human-AI collaboration for comprehensive and nuanced analyses in educational research.

Introduction

Qualitative analysis in educational research is pivotal for understanding complex phenomena through non-numerical data such as text, images, or videos. It captures subjective experiences, perceptions, and contexts, which are crucial for comprehending educational processes (Denzin & Lincoln, 2011). This approach explores teaching methods, learning experiences, policy impacts, and institutional dynamics, focusing on individuals' voices and perspectives (Bogdan & Biklen, 2008; Stake, 1995). By addressing key issues in learning and teaching, such as student engagement, instructional effectiveness, and educational equity, qualitative analysis provides valuable insights for improving educational practices.

A key strength of qualitative research is its interpretive capability, which is crucial for deciphering underlying meanings and motivations in education (Merriam & Tisdell, 2016). It offers flexibility and adaptability in research design, allowing for in-depth exploration of nuanced educational contexts (Marshall & Rossman, 2014). However, qualitative analysis faces challenges such as the time-intensive nature of data collection and analysis, potential bias, and issues of scalability and reproducibility. These challenges make it difficult to generalize findings and require careful consideration to maintain consistent interpretations across different studies (Bogdan & Biklen, 2008; Geertz, 1973; Stake, 1995).

The emergence of Generative Artificial Intelligence (GenAI) has significantly impacted various fields, including qualitative data analysis. GenAI systems, capable of generating new content by learning from existing datasets, have introduced efficiencies in data coding and analysis. Advanced Large Language Models (LLMs) like GPT-4, a subset of GenAI, excel at processing large textual datasets and identifying patterns, thus addressing time and capacity limitations inherent in manual coding (Gamieldien et al., 2023; Perkins & Roe, 2024). This capability is particularly relevant for analyzing extensive educational data, such as learner feedback, classroom observations, or policy documents.

These LLMs produce detailed codes, capturing nuances potentially missed in broader thematic analyses, enabling a more intricate understanding of the data. Studies such as Siiman et al. (2023) demonstrated how careful prompt design enables LLMs to code textual data with substantial agreement with human coders, significantly reducing manual coding time. Similarly, Dia et al. (2023) illustrated the use of GPT 3.5 Turbo as a "machine coder," automatically generating initial codes from text, guided by human-provided examples. This iterative process between human coders and LLMs refines themes, with LLMs offering rationale until a convergence of understanding is reached. Perkins and Roe (2024) highlight the synergy between human analysis and AI-supported inductive thematic analysis, emphasizing how the combination enhances data processing and interpretative depth in educational research contexts. Such advancements not only bolster the efficiency and breadth of qualitative analysis but also pave the way for innovative research methodologies in education. De Paoli (2024) conducted an experiment with GPT 3.5-Turbo to perform an inductive thematic analysis on semi-structured interviews, comparing the results to previous human analyses. The study found that the LLM was able to infer most of the main themes identified by human researchers, demonstrating the model's capacity to produce valid qualitative analyses, though with some limitations in handling certain themes and ethical concerns. Such advancements not only bolster the efficiency and breadth of qualitative analysis but also pave the way for innovative research methodologies in education.

GenAI's role in qualitative research addresses key challenges like the time-consuming nature of data collection and analysis, especially in extensive studies on learning and teaching. Its ability to apply consistent coding schemes across diverse datasets improves the reproducibility and generalizability of studies, overcoming human-induced variability. As GenAI continues to evolve, its integration into research is expected to streamline processes and introduce novel methodologies, enhancing the scope and depth of educational research (Ismail et al., 2023). This evolution suggests a future where GenAI's rapid data processing capabilities are seamlessly integrated with human expertise, offering a balanced approach to qualitative analysis in educational contexts.

TRACER (Transcript Analysis and Concept Extraction Resource) (Link: https://tracer-9pa5.onrender.com), developed by the author and free to use, is a GenAI-powered platform that utilizes GPT-4 API to perform thematic analysis on interview transcripts, a process traditionally dominated by manual human effort. The primary purpose of this paper is to conduct a detailed evaluation of GenAI via TRACER, focusing on its efficiency and reproducibility compared to traditional human-led thematic analysis in educational research. This evaluation aims to determine how effectively TRACER can streamline the qualitative research process, particularly regarding time and resource expenditure. Additionally, the paper seeks to explore the extent to which TRACER's findings are consistent and reliable across multiple runs, compared to the variability inherent in human analysis. Furthermore, this paper examines strategies for effectively integrating TRACER's GenAI-driven capabilities with human expertise. The dynamics of this integration are explored to understand how the collaboration between human analysts and TRACER can lead to more comprehensive, accurate, and nuanced results in qualitative research.

To achieve these objectives, the paper is structured around two key research areas and four research questions:

1. Efficiency and Reproducibility of TRACER:
   - Does the use of TRACER significantly reduce the time and resources required for thematic analysis compared to traditional human-led methods?
   - How do the findings of TRACER compare to the consistency of human analysis?

2. Integration and Synergy of Human and AI Analysis:

- In what ways can human analysis and AI-driven tools such as TRACER complement each other to produce more comprehensive results in qualitative research?

- What strategies can be employed to effectively integrate human expertise and AI capabilities in analyzing qualitative data?

Through addressing these research questions, the paper intends to shed light on the transformative impact of GenAI in qualitative research, particularly in educational settings, and to propose actionable strategies for the effective integration of GenAI-powered tools in qualitative data analysis.

Literature review

In the evolving landscape of qualitative research, particularly within educational settings, the arrival of GenAI has ushered in a new era of innovation and efficiency. This literature review explores the growing field of GenAI applications in qualitative analysis, specifically focusing on the integration of LLMs to enhance the coding process, improve reproducibility, and foster a synergistic collaboration between human expertise and AI capabilities. The review then summarizes and synthesizes key findings from recent studies that explored the practical applications and implications of GenAI in qualitative research, highlighting both the advancements made and the challenges that need to be addressed.

Gamieldien et al. (2023) illustrate the efficiency of LLMs like GPT-3.5 in qualitative coding. Their method algorithmically clusters over 10,000 exam responses, allowing human analysts to focus on interpreting meanings rather than manually identifying themes. This approach enhanced reproducibility due to consistent machine coding whereas human expertise contextualized AI-generated codes into broader concepts. This synergy between algorithmic processing and human interpretation addressed limitations inherent in each method when used alone, supporting the view that qualitative coding should not be entirely automated but should instead enhance the researcher's skills (Katz et al., 2024; Liu et al., 2023; Xiao et al., 2023).

Dai et al. (2023) proposed a collaborative framework involving humans and LLMs for thematic analysis. The objective was to utilize LLMs' predictive power to reduce the time and labour in qualitative coding, ensuring analytical quality. Their framework, which facilitated bidirectional communication between human coders and the LLM, allowed for iterative refinement of codes and themes. Their study demonstrated high inter-annotator agreement between human and LLM analyses, indicating that AI-generated codes were as reliable as those produced by human-only coding.

This research signifies an efficient emulation of human qualitative analysis tasks by GenAI systems, balancing computational scalability with human judgment. However, the study identified the need for further refinement in prompt design and model variability to enhance the efficiency of this collaborative approach. Despite these challenges, the human-LLM framework offers a promising avenue for integrating the strengths of both human and AI methods in qualitative research.

Siiman et al. (2023) investigated the use of ChatGPT and GPT-4 for qualitative analysis of collaborative problem-solving discourse. They tested both deductive coding and inductive rubric generation, finding substantial interrater reliability (Cohen's $\kappa = 0.706$) when detailed prompts were used. However, inductive rubrics showed lower agreement, highlighting a need for further research in complex problem areas. The study suggests that AI-assisted analysis improves transparency and reproducibility, although human oversight is essential to mitigate bias and align with research objectives.

Chew et al. (2023) introduced "LLM-assisted content analysis" (LACA) to integrate LLMs into deductive coding in qualitative research. They compared GPT-3.5's performance with human coders across various datasets, finding that GPT-3.5 often matched human-level inter-rater reliability and was significantly more efficient. This finding suggests that LLMs can reduce the manual effort in deductive coding, although their serial API requests may not fully capture the potential efficiency gains.

Zhang et al. (2023) developed QualiGPT, a toolkit that utilizes LLMs for qualitative data analysis, to automate thematic analysis with customized prompts. Their tests showed that QualiGPT dramatically enhanced efficiency compared to manual coding and matched the accuracy of experienced researchers. The integration of techniques such as batching and role-playing in prompts addresses issues like context limitation and inconsistency, making QualiGPT a valuable tool for qualitative analysis and human-AI collaboration.

Gao et al. (2023) proposed CollabCoder, a system using LLMs for collaborative qualitative analysis. In a study with 16 participants, CollabCoder improved discussion quality and inter-rater reliability compared to traditional methods. The study noted that LLMs, as suggestion providers, reduce the cognitive load and improve efficiency. However, caution is needed to maintain human control and prevent overreliance on AI suggestions, ensuring balanced and unbiased analysis outcomes.

Finally, Perkins and Roe (2024) explored integrating Generative AI tools, specifically ChatGPT, into the inductive thematic analysis of qualitative data. They employed a dual-method approach where one researcher employs traditional manual coding, and another uses ChatGPT to assist in coding. The study highlighted the enhanced capacity for data processing and theme identification provided by GenAI, coupled with the interpretative depth of human analysis. Key findings included the increased efficiency and objectivity offered by GenAI tools and the challenges posed by inconsistencies and hallucinations in AI outputs. The research underscores the complementary relationship between GenAI and human expertise, advocating for their synergistic use to expedite analysis without compromising the essential role of human researchers.

Synthesis of the literature on GenAI in qualitative analysis

A common thread across studies, starting with Gamieldien et al. (2023) and extending through Dai et al. (2023) and Chew et al. (2023), is the significant efficiency gain offered by GenAI in the coding process. These studies collectively underscore how LLMs expedite the traditionally time-consuming task of coding without compromising the depth and quality of analysis. The ability of LLMs to rapidly process and cluster large volumes of data is a recurring highlight, suggesting a shift toward more scalable and less labour-intensive methods in qualitative research.

The studies by Siiman et al. (2023) and Gao et al. (2023) explored the dynamics of human-AI collaboration. They revealed that while LLMs can autonomously perform certain analytical tasks, the integration of human expertise was crucial for contextually rich and nuanced interpretations. This interplay between AI's computational power and human cognitive skills points to an optimal model of qualitative research that leverages the strengths of both entities. The collaborative framework proposed by Dai et al. (2023) exemplifies this synergy, highlighting the iterative refinement process that benefits from both AI efficiency and human judgment.

The advanced applications of GenAI in qualitative research are vividly demonstrated in the studies by Zhang et al. (2023) with QualiGPT and Gao et al. (2023) with CollabCoder, and Perkins and Roe (2024) with their innovative dual-method approach. These studies showcase the cutting-edge uses of LLMs in thematic analysis and collaborative qualitative analysis opening new avenues for methodological innovations. The capability of LLMs to serve as independent analytical tools or as facilitators in collaborative settings suggests a future where qualitative research can be more inclusive, diverse, and comprehensive in its analytical scope.

The potential of GenAI in qualitative research is evident, however, studies have collectively highlighted ongoing challenges, particularly in optimizing prompt design (Liu et al., 2023), ensuring model variability (Dai et al., 2023), and managing the risks of AI-induced biases (Siiman et al., 2023). Perkins and Roe (2024) further emphasized the necessity of addressing AI hallucinations and maintaining rigorous validation processes to ensure research validity. These considerations indicate the need to continuously refine AI tools and methodologies, ensuring they complement rather than replace human expertise.

In conclusion, the synthesis of these studies reveals a consistent narrative: GenAI, and specifically LLMs, are not only enhancing the efficiency and scalability of qualitative research but also enriching its depth and quality through synergistic human-AI collaboration. The future of qualitative research in educational settings, shaped by these advancements, appears poised for groundbreaking developments that balance technological innovation with the irreplaceable value of human insight.

Methodology

This study employed a mixed-methods comparative data analysis to examine the differences in findings and methods between a human research team and TRACER powered by GPT-4 API (an LLM) when analyzing the same data set. The primary objective was to explore the effectiveness and efficiency of the GenAI system compared to the human team in analyzing the data and identifying patterns and insights. The data set consisted of interview transcripts with educators who implemented a new mathematics instructional model in research schools between 2013 and 2015. The decision to use a mixed-methods approach was grounded in the need to capture both the qualitative richness of human interpretation and the quantitative efficiency and consistency of AI analysis. By integrating both qualitative and quantitative methods, this approach offers a comprehensive examination of the effectiveness, efficiency, and reliability of the GenAI system compared to traditional human-led thematic analysis.

Data collection

The data collection process for the original interviews involved research team members conducting face-to-face interviews with educators who implemented a new mathematics instructional model in research schools between 2013 and 2015 (see Preciado-Babb et al., 2015). The interviews were conducted face-to-face by research team members, each lasting approximately one hour, and were subsequently transcribed verbatim. Ethical considerations, including informed consent and participant anonymity, were rigorously maintained throughout the data collection process.

Data sources

The data used in this research consists of semi-structured interview transcripts with six educators. The transcripts contain approximately 67,200 words, including the interview questions and the educators' responses.

The semi-structured interview questions include:

- What specific advice would you offer to new Math Minds teachers?
- Have you benefited from materials?
- What specific advice would you give to new teachers joining Math Minds?
- Have you found [JUMP Math] materials to be helpful?
- Restrictive or difficult? To what extent did you follow the teachers' guide? SmartBoard lessons? Workbook?
- In what ways did you improvise / extend / elaborate? Have you found [JUMP Math] principles

- helpful? Restrictive or difficult?
- What are your goals or priorities for improving your teaching of math?

Research team analysis (qualitative component)

The research team utilized NVivo to analyze the interview transcripts, as it enhanced inter-coder reliability by providing a platform for consistent coding practices among multiple researchers (Limna, 2023). The primary focus of this analysis was to explore the impact of a new mathematics teaching model, as highlighted in the work of Preciado-Babb et al. (2015). The team employed a multifaceted approach to ensure the rigour and trustworthiness of their qualitative analysis.

Initially, multiple researchers independently coded the interview transcripts using NVivo. This phase involved identifying significant phrases, concepts, and emerging themes directly from the data. Each researcher applied an interpretive lens, marking segments of the data they deemed relevant to the study's objectives and theoretical underpinnings. Establishing inter-coder reliability was crucial, as it is a key measure of consistency in qualitative research. This process took approximately a week to complete.

After the independent coding process, the research team engaged in comprehensive discussions to reconcile any coding discrepancies. These discussions were not merely for resolving differences but served as a collaborative effort to deepen the collective understanding of the data. Often, these deliberations necessitated revisiting the transcripts to reassess and reinterpret the data, enriched by the insights gained from the group's collective analysis. This process also took about a week to complete.

To further validate the findings, the team employed data triangulation, comparing the qualitative findings from the transcripts with observational field notes. These notes offered additional insights into the context and subtleties of the teachers' experiences and actions. This triangulation process corroborated the themes identified from the transcripts, ensuring that the findings were anchored in a robust and diverse evidence base. This process took approximately four days to complete.

The culmination of this meticulous process was the derivation of themes. This step involved synthesizing the coded data to identify consistent patterns, relationships, and overarching concepts. Four themes emerged from this analysis, reflecting key insights into educators' use of resources in their documentation process. Identifying these four themes required approximately 20 days.

TRACER

TRACER is a web-based application designed to utilizes GenAI to perform thematic analysis on interview transcripts. The program employs the GPT-4 API to identify recurring themes or patterns within the transcripts. To identify themes, the transcripts undergo several stages of processing (see Figure 1). Initially, the user must input contextual information about the transcripts, such as the number of themes TRACER is instructed to find, who is being interviewed and the topic of the interview. This information is necessary for the model to accurately identify themes. The raw transcripts, which may be presented in various formats (e.g., .docx, .txt, .pdf), are read and converted into plain text format. In earlier versions of the program, all data were consolidated into a single large text file. However, upon examination, it was observed that the results exhibited a bias against the contents of the second half of the text file and disproportionately included themes from the first half of the text file. Consequently, the program was modified to ensure all transcripts were converted into separate text files, eliminating sectional bias. Subsequently, the text files are divided into smaller segments and then inputted into the GPT-4 API for vector space indexing.

TRACER analysis (quantitative component)

To perform the thematic analysis, TRACER is given a query string that defines the task. Prompt 1 is designed to analyze the transcripts, identify key themes, and support its analysis with a brief explanation for each transcript. The application then retrieves the identified themes from the vector space index of every transcript and saves them all into a single text file. To identify recurring themes from this text file, TRACER is given a new query with Prompt 2, which tasks the GPT-4 API with identifying the common threads connecting various ideas within a specific context. With each execution, the final output is saved with a timestamp, enabling easy identification and comparison with the research team's analysis.

A critical aspect of this method involves addressing the non-deterministic nature of the GPT-4 API. In GenAI systems, non-determinism implies that identical inputs might not always produce identical outputs due to inherent variations in the model's response generation. To counteract this variability and enhance the reliability of the thematic analysis, the temperature parameter in the GPT-4 API was set to zero. This setting minimizes the randomness in the LLM's responses, striving for more consistent outputs.

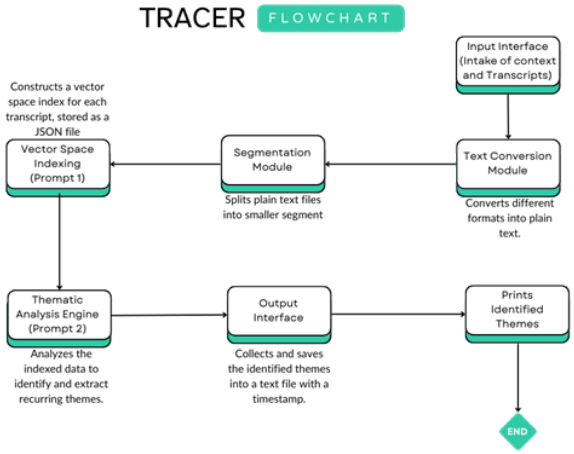

Figure 1. The workflow of TRACER.

The decision to run TRACER 100 times for each analysis task is particularly noteworthy. This repetition was not arbitrary but a carefully considered methodological choice, crucial for several reasons:

1. Enhancing Reliability: Running the program multiple times provided a broad sample of outputs, allowing the researcher to assess the consistency of the themes identified by TRACER. This approach was essential in evaluating the reliability of the AI tool, particularly given the non-deterministic nature of AI responses.

2. Data Validity: Multiple runs enabled the researcher to gauge the validity of the data. By analyzing the variance and similarities in the themes identified across different runs, the researcher determined the extent to which TRACER's analysis aligned with the qualitative data's inherent patterns and nuances.

3. Statistical Robustness: From a statistical standpoint, conducting numerous runs offered a substantial dataset for analysis. This robust dataset was pivotal in drawing reliable conclusions about the thematic trends and patterns identified by TRACER.

4. Mitigating AI Model Variability: Given the variability in GenAI model responses, multiple runs ensured a comprehensive exploration of the thematic space. This was particularly important for capturing a wide range of potential themes that a single run might not reveal.

Results

Following separate analyses conducted by the research team and the AI system, the researcher performed a comparative data analysis to evaluate the similarities and differences in the findings. This comparison aimed to assess the effectiveness and efficiency of the AI program in relation to the research team and to identify any unique insights or patterns that either method might have uncovered.

Results of the research team's analysis

Preciado-Babb et al. (2015) organized the results into four sections, discussing how educators utilized the JUMP Math materials and applied Math Minds principles. The authors highlighted continuous assessment, bonusing, document genesis, and inquiry and problem-solving as themes extracted from analyzing the transcripts.

Continuous assessment was a recurring theme in the interviews, with educators using small whiteboards for in-the-moment assessments. The authors highlighted quotes from educators stating that this practice helped them understand students' challenges and informed their decisions on whether to skip parts of the material.

Bonusing was also mentioned by all interviewed educators. The authors noted that many found it challenging to create bonus questions or tasks. Some educators believed that it was essential to personalize bonus questions for each student, while others found it easier to create bonus questions based on the structure of the material. For document genesis, the authors noted that teachers claimed to follow the teachers' guide and use JUMP Math materials consistently. However, the study also noted that the teachers sometimes adapted the materials based on various factors, such as their experience, time constraints, familiarity with other resources, or adherence to Math Minds principles. Lastly, Preciado-Babb et al. (2015) stated that teachers consistently perceived a lack of opportunity for problem-solving or inquiry in the JUMP Math approach. While teachers acknowledged the importance of building a foundation through mini-steps, some expressed concerns about students not being accustomed to multi-step problems or inquiry-based learning.

Results of TRACER's analysis

In this section, TRACER's findings on the same data that the research team analyzed are presented. The AI analysis underwent several iterations, refining the prompts to align the outcomes with the research team's findings (see Table 1 for prompt evaluation). This iterative process allowed for a more comprehensive comparison between the human and AI analyses.

Table 1. Evaluation of prompts.

	Iteration 1	Iteration 2	Iteration 3
Prompt 1	You are a thematic analysis bot. You analyze content from {transcripts}. The goal is to analyze the frequency of specific words, themes, or other characteristics in a set of content. Given this information, please answer the question: {what are the main themes?}. Always list your response	You are a content analysis bot. You analyze content from {transcripts}. The goal is to analyze the frequency of specific words, themes, or other characteristics in a set of content. Given this information, please answer the question: {what are four main themes?}. Always list your response.	You are a content analysis bot. You analyze content from {transcripts}. The goal is to analyze the frequency of specific words, themes, or other characteristics in a set of content. Given this information, please answer the question: {what are four main themes?}. Always list your response.
Prompt 2			You are given content containing a list of themes gathered from multiple transcripts. Your task is to analyze these themes and identify the four most frequent and central aspects related to teaching, materials, and assessment. Please consider the following instructions: Recognize semantically similar themes, even if they have different wording, and group them together. Focus on the common threads connecting various ideas within the context of teaching and assessment. Rank the themes based on their frequency, and present the top 4 most common themes. For each of the top 4 themes, provide a brief explanation of how it relates to teaching and assessment. Here is the content: {themes} Analyze the themes, and provide your insights on the most frequent and central aspects related to teaching and assessment.

Initially, TRACER was not given specific instructions on the number of themes to identify. The 100 executions of the application cumulatively lasted 3,760 seconds, meaning each execution averaged 37.6 seconds. After the executions, themes were extracted, and the researcher manually clustered the themes based on semantic similarity. This process lasted approximately an hour and involved a careful examination of the underlying ideas and patterns within the themes, enabling the researcher to categorize them into distinct clusters that represented broader, cohesive thematic concepts. Five themes were identified from 100 executions (see Table 2 for findings).

Table 2. Finding of TRACER after 100 executions first iteration.

Theme no.	Theme	Number of appearances
1	Breaking down Math into micro lessons	56
	The importance of breaking down concepts into microsteps	44
2	Constant assessment	46
	The importance of frequent assessment	39
	The use of frequent assessment	15
3	The need to create an environment where students feel comfortable and can experience success, regardless of their range of abilities	4
4	The importance of providing students with constant praise and encouragement when they are successful	2
5	The need for teachers to have a strong understanding of mathematics and the ability to unpack and provide answers	1

Themes that were present in every execution of TRACER are considered valid. Semantically similar variations of themes 1 and 2 appeared in 100% of the extracted themes. Consequently, TRACER identified only two themes: Micro-steps and Continuous Assessment. Although these themes were relevant, the findings were not as extensive as the research team's analysis, which identified four primary themes.

In response, Prompt 1 was revised and TRACER was instructed to find four main themes. This time, the 100 executions of the application cumulatively lasted 3802 seconds, with each execution averaging 38.02 seconds. The researcher then clustered the themes based on semantic similarity, and the process lasted approximately one hour. Table 3 displays the findings.

The results indicated that some semantically similar variations of the two themes of Micro-steps and Continuous Assessment appear in 100% of the findings. Semantically similar variations of the theme Bonus Questions appeared 66 times and were identified as the third theme due to their high frequency of occurrence. However, based on the results, it was not possible to identify the fourth theme. It seemed that TRACER's focus shifted primarily towards the student experience rather than the JUMP Math material and teaching strategies. Although these findings provided valuable insights into the student perspective, they did not completely align with the human research team's themes. Furthermore, there were too many variations in the findings, which impacted the validity of the results.

Table 3. Finding of TRACER after 100 executions second iteration.

Theme no.	Theme	Number of appearances
1	Breaking down Math into micro lessons	34
	The importance of breaking down concepts into microsteps	41
	The importance of breaking down concepts into smaller steps and providing clear instructions to students	25
2	Constant assessment	45
	The importance of frequent assessment	40
	The use of frequent assessment	15
3	The use of bonus questions	57
	The importance of using bonus questions	9
4	The importance of attention to mental math	7
5	The need to create an environment where students feel comfortable and can experience success, regardless of their range of abilities	5
6	The importance of providing students with constant praise and encouragement when they are successful	5
7	The need for teachers to have a strong understanding of mathematics and the ability to unpack and provide answers	4
8	The need to use experiences from the current year to shape teaching for the next year, and the potential for the JUMP program to be successful in helping students find success	4
9	The need to assess student understanding and address any gaps in knowledge	4
10	The importance of creating a lesson structure that encourages students to discover and explore together	3
11	The need for abstract thinking	2

In a final revision, the code was modified to instruct TRACER to save the themes gathered from each execution into one file. So, Prompt 1 remained unchanged. However, a second process was added in Prompt 2, which instructed TRACER to recognize semantically similar themes by reviewing all the themes identified by GPT-4 after each execution. The 100 executions of this version of TRACER lasted 4,288 seconds. Once again the researcher clustered the themes based on semantic similarity. This process lasted approximately 90 minutes. The headings are displayed in Table 3.

Table 4. Finding of TRACER after 100 executions third iteration.

Theme no.	Theme	Number of appearances
1	Breaking down tasks into smaller, more manageable steps	37
	Microsteps	31
	Breaking down tasks into microsteps	32
2	Constant assessment	84
	The importance of frequent assessment	12
	Assessment	4
3	Engaging students	54
	Engagement and motivation	46
4	Teacher preparation and comfort	31
	Teacher preparation	27
5	Comfort with new methods of teaching	26
	Comfort level	16

The results indicated that semantically similar variations of micro-steps appear in 100% of the findings. Similarly, semantically similar variations of continuous assessment also appeared in 100% of the findings. Moreover, semantically similar variations of student engagement also appeared in 100% of the findings. Semantically similar variations of

teacher preparation appear in 58% of the findings, and semantically similar variations of flexibility with new teaching methods appeared in 42% of the findings.

TRACER identified key themes in effective mathematics teaching and assessment, including the use of micro-steps for simplifying complex concepts, the role of continuous assessment in monitoring student progress, and the importance of student engagement and motivation through techniques like bonus questions. Another critical theme was teacher preparation, emphasizing the need for educators to be well-versed in new teaching methods and adaptable in material presentation.

TRACER's analysis, refined through several iterations and prompt adjustments, closely aligns with the research team's findings. This iterative process highlighted the importance of specific prompts in guiding GenAI's analysis to desired outcomes. Tailoring prompts to focus on a number of themes, subject matter, and analysis goals significantly improved the GenAI's relevance and accuracy. The study illustrates GenAI's potential in qualitative research with precise instructions and underlines the necessity of continuous human oversight to ensure contextually accurate AI analysis.

Discussion

Efficiency of TRACER in thematic analysis

The adoption of GenAI via tools such as TRACER in educational research has potentially and significantly improved efficiency in thematic analysis, particularly in terms of time and resource utilization. This development addresses a key challenge in qualitative research: the labour-intensive nature of data analysis (Bogdan & Biklen, 2008). In this study, TRACER significantly expedited the thematic analysis process compared to the human research team. For instance, tasks that took the human team 20 days were completed by TRACER and one researcher in mere hours. This efficiency gain echoed the findings of Gamieldien et al. (2023) regarding the time-saving capabilities of Large Language Models in data coding.

This enhanced efficiency has substantial implications for larger-scale qualitative studies in education. It enables researchers to handle more extensive datasets more feasibly, potentially leading to richer insights and a deeper understanding of complex educational phenomena. Therefore, the integration of TRACER marks a significant advancement in educational research methodologies, addressing time and resource constraints and facilitating more comprehensive qualitative analyses.

Reproducibility and consistency of TRACER findings

The comparative analysis of themes identified by TRACER and the research team offers a unique perspective on the reproducibility and consistency of GenAI in qualitative analysis. This comparison is crucial in understanding the reliability of GenAI tools such as TRACER in capturing the nuances of educational research data.

Thematic alignment between TRACER and the human research team

Both TRACER and the research team identified key themes related to educational practices and teaching methodologies, although they varied in labelling and focus. For instance, both TRACER and the human team recognized the theme of continuous assessment, which is crucial for understanding teaching dynamics. This alignment signifies a level of consistency in GenAI's ability to identify major themes that are also discernible to human analysts. Similarly, the theme of microsteps, although not initially labelled as a distinct theme by the research team, was later acknowledged in their subsequent research (Metz et al., 2016). TRACER's identification of this theme aligns with the human team's findings, indicating GenAI's potential to uncover underlying patterns that might not be immediately apparent even to experienced researchers. However, discrepancies were also noted. For instance, the research team's identification of the theme related to inquiry and problem-solving was not mirrored in TRACER's analysis. This divergence could stem from differences in the analytical focus and the specific prompts guiding TRACER, highlighting the importance of prompt design in directing AI analysis (Short & Short, 2023).

Reliability of GenAI in qualitative analysis

The reproducibility and consistency of findings are critical metrics in assessing the reliability of any analytical tool. In the case of TRACER, the repeated identification of certain key themes across multiple analyses suggests a high degree of reproducibility. This consistency is particularly noteworthy given the non-deterministic nature of AI responses (Yang & Menczer, 2023). It demonstrates that with carefully designed prompts and a structured analytical approach, GenAI can reliably identify major themes in qualitative data.

However, the lack of contextual understanding and the dependence on specific prompts are limitations that need to be acknowledged (Ray, 2023; Sun & Hoelscher, 2023). Although TRACER effectively identified several key themes, its analysis might overlook subtleties that human analysts, with their contextual knowledge and interpretive skills, could capture (Byrne, 2022; Joffe, 2011). This aspect underscores the complementary roles of human analysts and GenAI in qualitative research.

Integration and synergy of human and AI analysis

The integration of GenAI tools such as TRACER with human expertise in qualitative analysis reveals a complementary relationship that enhances the overall depth and breadth of research. This synergy leverages the strengths of both AI-driven analysis and human interpretation, facilitating a more nuanced and comprehensive understanding of educational research data.

TRACER's AI-driven analysis excels in efficiently processing large volumes of data, identifying recurring themes, and providing consistent results. Its capability to rapidly analyze and code data allows for the handling of extensive datasets,

which might be challenging and time-consuming for human researchers (Haleem et al., 2022; Lund et al., 2023; Perkins & Roe, 2024). However, GenAI tools like TRACER may lack the nuanced understanding and contextual awareness that human analysts bring. Researchers possess the ability to interpret data beyond its explicit content, drawing on their expertise, experience, and understanding of the educational context (Byrne, 2022; Joffe, 2011).

Reflecting on the experience of designing, developing, and deploying TRACER to harness the full potential of both GenAI and human expertise, several strategies can be adopted:

1. Iterative Collaboration: Researchers should implement an iterative process to review and refine initial AI-generated themes. This approach contextualizes AI findings and ensures that the final themes accurately reflect the depth of the data

2. Prompt Design and Calibration: AI prompts must be designed carefully to align with the research objectives and context. Regular calibration of these prompts based on feedback guides the AI analysis toward more relevant and contextually appropriate themes.

3. Blended Analysis Teams: Teams comprising both AI tools and analysts should be established. AI handles the initial data processing, allowing researchers to focus on interpreting and contextualizing the findings, thus creating a balanced and efficient workflow.

4. Training and Sensitization: Researchers should receive training to interact effectively with AI tools, understanding their capabilities and limitations. Simultaneously, AI systems should be sensitized to the specific nuances of educational research through continuous learning and feedback loops.

Limitations of the current study

This study demonstrates TRACER's capabilities in educational research, yet it has limitations that need acknowledgment for a comprehensive understanding of its results and implications. The focus on a specific set of educational research transcripts limits generalizability, as educational research varies across subjects, educational levels, and cultural contexts. Thus, TRACER's efficacy in different educational settings remains untested, potentially limiting its applicability across diverse educational research scenarios.

Moreover, although TRACER efficiently identifies key themes, its performance in deeper interpretive tasks, such as understanding context and subtle nuances, is less robust compared to human analysts. The tool's reliance on precisely designed prompts also raises concerns about its autonomy in theme identification. Methodologically, the study's comparative analysis might not fully reflect the complexities of qualitative data, emphasizing thematic consistency over thematic richness and depth. Furthermore, the dynamics of human-AI collaboration in qualitative analysis are not thoroughly explored, particularly the integration of human intuition and expertise with AI-generated themes. Understanding how human and AI analyses can effectively complement each other remains a crucial area for further research.

Implications for research

The findings of this study have significant implications for the integration of GenAI into qualitative research in educational settings. By demonstrating that TRACER can efficiently and consistently perform thematic analysis on qualitative data, this study suggests that AI tools have the potential to significantly enhance the research process. However, the implications extend beyond mere efficiency gains, encompassing broader considerations for the future of qualitative research and the role of AI in academic inquiry.

1. Enhanced Efficiency and Scalability: The use of TRACER has shown that GenAI can drastically reduce the time and resources required for thematic analysis. This efficiency enables researchers to handle larger datasets more feasibly, potentially leading to richer insights and a deeper understanding of complex educational phenomena. As educational research often involves extensive qualitative data, such as interviews and classroom observations, the ability to process this data quickly and accurately is invaluable.

2. Reproducibility and Consistency: The study highlights the consistency of AI-driven analysis, as TRACER was able to reproduce key themes across multiple executions. This reproducibility addresses a critical challenge in qualitative research, where human analysis can be subject to variability. The ability to achieve consistent results across different iterations of analysis strengthens the reliability of research findings and supports the generalizability of the results.

3. Complementary Role of Human-AI Collaboration: The findings underscore the potential for a synergistic relationship between human researchers and AI tools. While AI can efficiently identify and categorize themes, human researchers bring essential contextual understanding and interpretative depth. This collaboration can lead to more comprehensive and nuanced analyses, suggesting a future where AI assists in the initial stages of analysis, allowing researchers to focus on deeper interpretative tasks.

4. Implications for Training and Methodological Development: As AI tools like TRACER become more integrated into research, there will be a need for researchers to develop new skills in interacting with these tools. This includes understanding how to design effective prompts, interpret AI-generated outputs, and integrate these findings with traditional qualitative methods. Institutions

may need to offer training and support to help researchers effectively use AI in their work, ensuring that the tools complement rather than replace human expertise.

Potential for further research

The integration of GenAI tools such as TRACER into educational research presents several avenues for future exploration to enhance their application and effectiveness. Key areas for future research include:

1. Diversifying Educational Contexts: Researchers should investigate GenAI tools across various academic disciplines, educational stages, and cultural settings. This research will provide insights into the adaptability and scalability of GenAI in different educational environments, thereby assessing its flexibility and effectiveness more broadly.

2. Enhancing Interpretive Capabilities: Developing advanced natural language processing techniques to improve GenAI's understanding of context, subtleties, and implicit meanings in qualitative data is crucial. Research aimed at enabling GenAI tools to mimic human-like interpretive skills will significantly advance the field.

3. Integrating GenAI with Human Expertise: Collaborative frameworks that combine AI-generated themes with human analytical depth should be investigated. This approach, leveraging AI's efficiency and human interpretive skills, will ensure a balanced and comprehensive analysis

4. Training and Continuous Learning: Researchers must be equipped with the skills to effectively use GenAI tools. Concurrently, refining GenAI through diverse data inputs and human feedback will enhance its accuracy and relevance in educational research.

5. Addressing Ethical Implications: As GenAI gains prominence in educational research, understanding its ethical implications becomes critical. Future research should focus on issues such as data privacy, consent, and potential biases in AI algorithms, thereby establishing ethical guidelines for GenAI use in research.

These future directions underscore the need for continuous innovation, ethical consideration, and collaborative efforts in harnessing GenAI's potential in educational research.

Conclusion

This study has explored the integration of GenAI, specifically through the TRACER tool, into the qualitative analysis of educational research data. By comparing TRACER's AI-driven analysis with traditional human-led thematic analysis, the research highlights both the potential and limitations of using AI in qualitative research. TRACER demonstrated significant efficiency, consistently identifying key themes across multiple runs and reducing the time and resources typically required for qualitative analysis. This efficiency enables researchers to manage larger datasets, leading to potentially richer insights and more comprehensive understandings of complex educational phenomena.

However, the study also underscored the importance of human involvement in qualitative research. While TRACER was effective in identifying broad themes, the depth of interpretation and contextual understanding provided by human researchers remains irreplaceable. The findings suggest that a synergistic approach, where AI tools like TRACER are used in tandem with human expertise, can enhance the overall quality of qualitative research. This collaboration allows for the strengths of both AI efficiency and human interpretative depth to be fully realized.

The implications of this study extend to the broader field of educational research, suggesting that the future of qualitative analysis may lie in the integration of AI and human capabilities. As AI tools become more sophisticated, there will be a growing need for researchers to develop the skills necessary to interact effectively with these tools, ensuring that they complement rather than replace human expertise. Additionally, the ethical considerations surrounding AI in research, such as data privacy and algorithmic biases, must be carefully addressed to maintain the integrity of academic inquiry.

In conclusion, the study contributes to the ongoing dialogue about the role of AI in qualitative research, particularly in educational settings. It offers insights into how AI can be harnessed to enhance research processes while also emphasizing the continued importance of human judgment and expertise. As the field evolves, further research is needed to explore the full potential of AI-human collaboration in qualitative analysis, ensuring that both technology and human insight are leveraged to their fullest extent in advancing educational research.

Acknowledgment

I gratefully acknowledge TD Bank Group's TD Ready Challenge Grant generous sponsorship of the Math Minds Initiative.

References

Bogdan, R., & Biklen, S. K. (2008). *Qualitative research for education: An introduction to theory and methods*. Pearson.

Byrne, D. (2022). A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & Quantity, 56*(3), 1391–1412. https://doi.org/10.1007/s11135-021-01182-y

Chew, R., Bollenbacher, J., Wenger, M., Speer, J., & Kim, A. (2023). *LLM-assisted content analysis: Using large language models to support deductive coding*. https://doi.

org/10.48550/arXiv.2306.14924

Dai, S.-C., Xiong, A., & Ku, L.-W. (2023). *LLLM-in-the-loop: Leveraging large language model for thematic analysis* (2310.15100). arXiv. https://doi.org/10.48550/arXiv.2310.15100

Denzin, N. K., & Lincoln, Y. S. (2011). T*he Sage handbook of qualitative research.* Sage.

De Paoli, S. (2024). Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review, 42*(4), 997–1019. https://doi.org/10.1177/08944393231220483

Gamieldien, Y., Case, J. M., & Katz, A. (2023). *Advancing qualitative analysis: An exploration of the potential of generative AI and NLP in thematic coding*. http://dx.doi.org/10.2139/ssrn.4487768

Gao, J., Guo, Y., Lim, G., Zhang, T., Zhang, Z., Li, T. J.-J., & Perrault, S. (2023). *CollabCoder: A GPT-powered workflow for collaborative qualitative analysis*. https://doi.org/10.48550/arXiv.2304.07366

Geertz, C. (1973). *The interpretation of cultures* (Vol. 5019). Basic books.

Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations, 2*(4), 100089. https://doi.org/10.1016/j.tbench.2023.100089

Ismail, F., Tan, E., Rudolph, J., Crawford, J., & Tan, S. (2023). Artificial intelligence in higher education. A protocol paper for a systematic literature review. *Journal of Applied Learning and Teaching, 6*(2), 56–63. https://doi.org/10.37074/jalt.2023.6.2.34

Joffe, H. (2011). Thematic Analysis. In D. Harper & A. Thompson (Eds.), *Qualitative research methods in mental health and psychotherapy* (pp. 209–223). Wiley. https://doi.org/10.1002/9781119973249.ch15

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). GPT-4 passes the bar exam. *Philosophical Transactions of the Royal Society, 382*(2270), 2–17. https://doi.org/10.1098/rsta.2023.0254

Limna, P. (2023). The impact of NVivo in qualitative research: Perspectives from graduate students. *Journal of Applied Learning & Teaching, 6*(2), 271–282. https://doi.org/10.37074/jalt.2023.6.2.17

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., & Tang, J. (2023). *GPT understands, too. AI Open*. https://doi.org/10.1016/j.aiopen.2023.08.012

Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology, 74*(5), 570–581. https://doi.org/10.1002/asi.24750

Marshall, C., & Rossman, G. B. (2014). *Designing qualitative research*. Sage publications.

Merriam, S. B., & Tisdell, E. J. (2016). *Qualitative research: A guide to design and implementation* (4th ed.). John Wiley & Sons.

Metz, M., Preciado Babb, P. A., Sabbaghan, S., Pinchbeck, G., Aljarrah, A., & Davis, B. (2016). Teachers' awareness of variation. *IDEAS 2016: Designing for Innovation Selected Proceedings* (pp. 182–191). http://hdl.handle.net/1880/51222

Perkins, M., & Roe, J. (2024). The use of Generative AI in qualitative analysis: Inductive thematic analysis with ChatGPT. *Journal of Applied Learning & Teaching, 7*(1), 390–395. https://doi.org/10.37074/jalt.2024.7.1.22

Preciado-Babb, P. A., Metz, M., Sabbaghan, S., & Davis, B. (2015). Insights on the relationships between mathematics knowledge for teachers and curricular material. In T. G. Bartell, K. N. Bieda, R. T. Putnam, K. Bradfield, & H. Dominguez (Eds.), *Proceedings of the 37th annual meeting of the North American chapter of the international group for the psychology of mathematics education* (pp. 796–803). https://eric.ed.gov/?id=ED584331

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems, 3,* 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Short, C. E., & Short, J. C. (2023). The artificially intelligent entrepreneur: ChatGPT, prompt engineering, and entrepreneurial rhetoric creation. *Journal of Business Venturing Insights, 19*, e00388. https://doi.org/10.1016/j.jbvi.2023.e00388

Siiman, L. A., Rannastu-Avalos, M., Pöysä-Tarhonen, J., Häkkinen, P., & Pedaste, M. (2023). Opportunities and challenges for AI-assisted qualitative data analysis: An example from collaborative problem-solving discourse data. In Y. Huang & T. Rocha (Eds.), *Innovative technologies and learning. ICITL 2023* (Vol. 14099, pp. 87–96). Springer. https://doi.org/10.1007/978-3-031-40113-8_9

Stake, R. E. (1995). *The art of case study research.* Sage.

Sun, G. H., & Hoelscher, S. H. (2023). The ChatGPT storm and what faculty can do. *Nurse Educator, 48*(3). https://doi.org/10.1097/NNE.0000000000001390

Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., & Oudeyer, P.-Y. (2023). 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* (pp. 75–78). https://doi.org/10.1145/3581754.3584136

Yang, K.-C., & Menczer, F. (2023). *Large language models can rate news outlet credibility.* arXiv. https://doi.org/10.48550/arXiv.2304.00228

Zhang, H., Wu, C., Xie, J., Kim, C., & Carroll, J. M. (2023). *QualiGPT: GPT as an easy-to-use tool for qualitative coding.* arXiv Preprint arXiv:2310.07061.