LLM-TA: An LLM-Enhanced Thematic Analysis Pipeline for Transcripts from Parents of Children with Congenital Heart Disease

Muhammad Zain Raza[1*], **Jiawei Xu**[1*], **Terence Lim**[2,3], **Lily Boddy**[2], **Carlos M. Mery**[4,5], **Andrew Well**[6,7], **Ying Ding**[1,7]

[1]School of Information, UT Austin
[2]College of Natural Sciences, UT Austin
[3]Graphen, Inc.
[4]Division of Pediatric Cardiac Surgery, Vanderbilt University Medical Center
[5]Pediatric Heart Institute at Monroe Carell Jr. Children's Hospital at Vanderbilt
[6]Texas Center for Pediatric and Congenital Heart Disease
[7]Dell Medical School, UT Austin

raza.zain08@austin.utexas.edu, {jiaweixu,terence.lim}@utexas.edu, lilyboddy212@gmail.com, carlos.mery@vumc.org, Andrew.Well@austin.utexas.edu, ying.ding@ischool.utexas.edu

Abstract

Thematic Analysis (TA) is a fundamental method in healthcare research for analyzing transcript data, but it is resource-intensive and difficult to scale for large, complex datasets. This study investigates the potential of large language models (LLMs) to augment the inductive TA process in high-stakes healthcare settings. Focusing on interview transcripts from parents of children with Anomalous Aortic Origin of a Coronary Artery (AAOCA)—a rare congenital heart disease—we propose an **LLM**-Enhanced **T**hematic **A**nalysis (**LLM-TA**) pipeline. Our pipeline integrates an affordable state-of-the-art LLM (GPT-4o mini), LangChain, and prompt engineering with chunking techniques to analyze nine detailed transcripts following the inductive TA framework. We evaluate the LLM-generated themes against human-generated results using thematic similarity metrics, LLM-assisted assessments, and expert reviews. Results demonstrate that our pipeline outperforms existing LLM-assisted TA methods significantly. While the pipeline alone has not yet reached human-level quality in inductive TA, it shows great potential to improve scalability, efficiency, and accuracy while reducing analyst workload when working collaboratively with domain experts. We provide practical recommendations for incorporating LLMs into high-stakes TA workflows and emphasize the importance of close collaboration with domain experts to address challenges related to real-world applicability and dataset complexity. https://github.com/jiaweixu98/LLM-TA

1 Introduction

Thematic Analysis (TA) is a widely employed qualitative research method, particularly prevalent in healthcare and related fields such as psychology (Braun and Clarke 2006), heart disease research (Mery et al. 2023), sport and exercise studies (Braun, Clarke, and Weate 2016), gender identity exploration (Bradford et al. 2020), and anorexia research (Tierney and Fox 2010). TA facilitates an inductive examination of participant perspectives, enabling the identification of unanticipated themes, patterns, and insights in textual datasets such as patient interview transcripts (Braun and Clarke 2006; Saldana 2011). The seminal work by Braun and Clarke (Braun and Clarke 2006) defines a six-phase process for inductive TA: (1) familiarization with the data, (2) generation of initial codes, (3) theme identification, (4) theme review, (5) theme definition and naming, and (6) report production. This iterative and reflective process involves continuous movement between these phases to ensure a thorough analysis (Nowell et al. 2017).

Inductive TA does not rely on pre-determined codes or themes; instead, analysts derive codes directly from the data without imposing external frameworks. However, this flexibility, while valuable for capturing nuanced insights, can lead to inconsistencies and challenges in coherence (Holloway and Todres 2003). Furthermore, the approach is resource-intensive, time-consuming, and does not scale effectively for large datasets. While small teams can manage hundreds of observations, analyzing thousands of data points introduces unique challenges related to consistency and resource allocation (Katz, Fleming, and Main 2024). Researchers have incorporated machine learning to assist humans during the data annotation process, such as by learning patterns in real-time as user annotated the data (Gebreegziabher et al. 2023) or leveraging rationale extraction models to generate theme recommendations (Overney et al. 2024). Some studies have explored automating the themes identification by using topic modeling techniques (Leeson et al. 2019; Guetterman et al. 2018). However, these methods are primarily statistical, only considering the prevalence of keywords in the corpus and do not capture the nuances of human researchers' results (Parfenova, Denzler, and Pfeffer 2024).

To address these limitations, several studies have explored the potential of large language models (LLMs) to automate inductive TA tasks. These efforts include evaluating LLMs' alignment with human annotations in diverse contexts such as video game players and data science educators (De Paoli 2024), psychiatric patient-clinician in-

[*]These authors contributed equally.

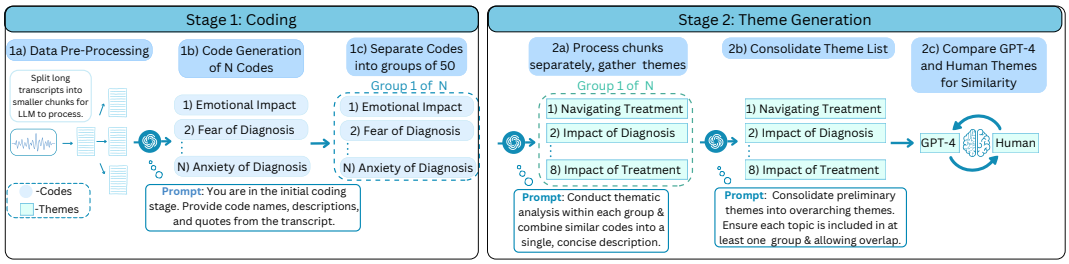

Figure 1: Pipeline for LLM-enhanced thematic analysis (LLM-TA) of transcripts from parents of children with AAOCA.

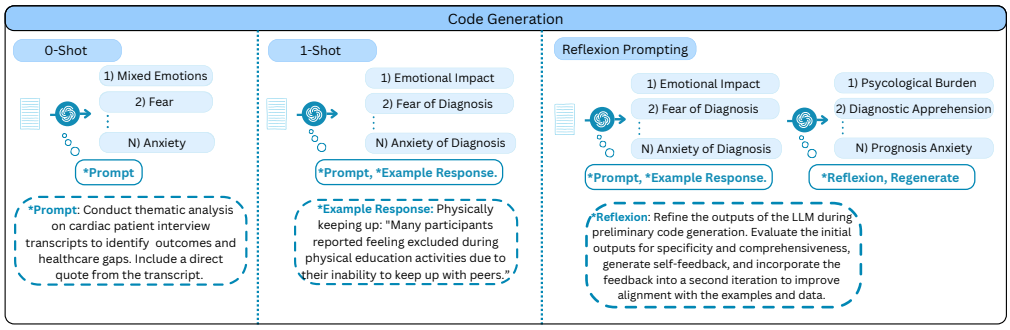

Figure 2: Prompting strategies in the LLM-TA code generation stage. These strategies are also used for theme generation.

teractions (Mathis et al. 2024), pandemic-era team feedback (Katz, Fleming, and Main 2024), barriers to arthroplasty (Mannstadt et al. 2024), vaccine rhetoric on Twitter (Deiner et al. 2024), media coverage of a controversial financial scandal (Khan et al. 2024), facts descriptions from criminal court opinions regarding thefts (Drápal, Westermann, and Savelka 2023), and SMS health intervention prompts (Prescott et al. 2024). Collaborative frameworks integrating human expertise with LLM capabilities have also been proposed to enhance the TA process (Dai, Xiong, and Ku 2023; Zhang et al. 2023; Gao et al. 2024). Both open-source models (e.g., Llama 2, Mistral) (Mathis et al. 2024; Katz, Fleming, and Main 2024) and proprietary models (e.g., GPT-3.5, GPT-4, Claude) (De Paoli 2024; Dai, Xiong, and Ku 2023; Singh et al. 2024; Mannstadt et al. 2024) have been utilized, demonstrating significant time savings and scalability while maintaining relevance (Mathis et al. 2024; Prescott et al. 2024). Additionally, some studies have evaluated LLMs' capacity for deductive qualitative research (Gao et al. 2024; Xiao et al. 2023; Singh et al. 2024).

Despite significant advances, LLMs' ability to perform inductive TA on real-world healthcare interview transcripts remains under-explored. Prior studies have primarily focused on lower-stakes domains like video games or music (Dai, Xiong, and Ku 2023; De Paoli 2024). Even within healthcare, existing analyses often use small datasets with shorter transcripts, failing to capture the complexity of real-world scenarios (Mathis et al. 2024; Mannstadt et al. 2024). Moreover, evaluations of LLM-generated themes have not involved researchers who conducted the original inductive TA on the same dataset (Mathis et al. 2024; Mannstadt et al. 2024). Advanced LLMs and robust prompt engineering techniques have yet to be fully leveraged in this context (Katz, Fleming, and Main 2024; Mathis et al. 2024). In this study, we introduce an **LLM**-Enhanced **T**hematic **A**nalysis (**LLM-TA**) pipeline and address these gaps with three key contributions:

- We present the first LLM-enhanced TA pipeline tailored for high-stakes, lengthy, real-world, de-identified transcripts. Specifically, we analyze interview transcripts with parents of children diagnosed with Anomalous Aortic Origin of a Coronary Artery (AAOCA), a type of congenital heart disease.
- We apply thematic similarity analysis, LLM-based judgments, and assessments by a TA expert who previously worked on the dataset to compare LLM-generated themes with human-coded ground truth. Employing chunking strategies, we test various prompt engineering techniques—including zero-shot, one-shot, and reflection—on contextually rich AAOCA interview transcripts. Our LLM-TA pipeline outperforms existing LLM-augmented TA methods in thematic accuracy, LLM assessment, and expert review.
- By closely collaborating with the inductive TA expert, we identified gaps of LLM-gernated themes compared to human researchers. We provide preliminary insights on integrating LLMs into real-world workflows to enhance the efficiency and scalability of inductive TA.

2 Methodology

We developed a **LLM**-Enhanced **T**hematic **A**nalysis pipeline (Figure 1) to perform inductive TA (Braun and Clarke 2006) on de-identified transcripts from nine focus group sessions involving 42 parents. These transcripts, with a median word count of 11,457, document conversations between interviewers and parents of children with AAOCA. Traditional inductive TA involves six stages: *1) Familiarization with the data, 2) Generation of initial codes, 3) Theme identification, 4) Theme review, 5) Final theme definition and naming, and, 6) Report production.* Stages *(2)* through *(5)* typically require at least two human researchers to independently generate and refine codes and themes, followed by collaborative discussions to finalize results. This iterative process is both time- and labor-intensive.

Formally, given a dataset of de-identified transcripts

$$T = \{T_1, T_2, \ldots, T_n\}$$

where $n = 9$ focus group sessions, involving 42 parents of children with AAOCA, our objective is to automate the inductive TA process using an LLM. Each transcript T_i is lengthy (median word count of 11,457) and contains rich conversational data between parents. The goal is to generate a set of initial Codes and and Themes:

$$C = \{c_1, c_2, \ldots, c_k\}, \Theta = \{\theta_1, \theta_2, \ldots, \theta_k\},$$

that capture significant patterns in the data. There are two primary stages in the LLM-TA pipeline (Figure 1):

Stage 1: Initial code generation on chunked transcripts. Codes are the foundational units of inductive TA, capturing significant concepts and ideas from the transcripts. Each code includes a concise name, a meaningful description, and representative quotes from the transcripts.

- **1(a) Splitting transcripts into smaller chunks.** To enable fine-grained analysis, we divided each transcript into smaller chunks of up to 1,500 words. Each chunk preserves the integrity of conversational context. This approach improves the LLM's ability to generate detailed and accurate codes, compared to directly processing entire transcripts as described by Mathis et al. (2024).
- **1(b) LLM-based initial coding.** We prompted the LLM to roleplay as an inductive TA researcher tasked with generating initial codes for each chunk. The prompt provided detailed context about the transcripts and instructed the LLM to identify exhaustive codes, each accompanied by a name, description, and representative quotes. On average, the LLM generated one code for every 225 words of transcript.
- **1(c) Grouping initial codes.** After generating codes for all chunks, we concatenated the codes (including their names, descriptions, and quotes) and divided them into N sequential groups. This divide-and-conquer strategy mitigates the LLM's limitations in processing large amounts of information at once (Liu et al. 2023; Wang et al. 2024), ensuring that the grouped codes retain fine-grained details for subsequent theme generation.

Stage 2: Theme Generation. Themes represent overarching concepts synthesized from multiple related codes. Each theme includes a short title, a detailed description, and the associated codes. These themes are critical for understanding participants' perspectives and are used to inform the research findings.

- **2(a) Preliminary theme generation.** For each group of codes, we prompted the LLM to synthesize preliminary themes. Acting as an inductive TA researcher, the LLM analyzed the grouped codes and their descriptions to identify themes. Each preliminary theme included a title, a detailed description, and the associated codes. This process was repeated for all N groups, resulting in N sets of preliminary themes. Notably, a single code could belong to multiple themes.
- **2(b) Final theme generation.** In the final stage, the LLM reviewed all preliminary themes and their associated codes across groups. It synthesized overarching themes by considering the participants' experiences, needs, and meaningful outcomes related to living with children diagnosed with AAOCA. The final themes included detailed descriptions and were designed to reflect the most prominent insights from the data.

Prompting Strategies. To implement the pipeline, we employed three distinct prompting strategies of increasing complexity (Figure 2): zero-shot, one-shot, and Reflexion (Shinn et al. 2024). All detailed prompts can be accessed through our public GitHub repository.

- **Zero-shot prompting.** In the zero-shot setting, we used straightforward prompts to guide the LLM in identifying codes from transcript chunks and generating the final themes, as outlined in the pipeline. These prompts included a detailed explanation of the inductive TA methodology. Additionally, the prompts incorporated comprehensive background information on the AAOCA transcripts, aligning with the familiarization stage of traditional inductive TA. This approach allowed the LLM to process the data without requiring prior examples.
- **One-shot prompting.** Building on the zero-shot approach, the one-shot setting introduced curated examples from a related inductive TA study. These examples served as templates to guide the LLM in generating code names and theme descriptions that adhered to the desired format and context.
- **Reflexion prompting.** In the Reflexion setting (Shinn et al. 2024), we extended the one-shot approach by incorporating iterative feedback to refine the LLM's outputs. After the LLM generated preliminary themes, we prompted it to critically evaluate its outputs, focusing on the specificity and comprehensiveness of the theme names and descriptions. The LLM then generated feedback on its own outputs, identifying areas for improvement. Using this feedback, we conducted a second round of theme generation, refining the preliminary themes to ensure greater alignment with the the underlying data. A similar Reflexion-based strategy was employed during Stage 2(b) (Final theme generation) to synthesize final

themes that were both coherent and representative of the participants' experiences.

3 Experiment and Evaluation

Qualitative research does not always have a definitive ground truth. In this study, to evaluate the performance of LLM-TA pipeline, we employ two complementary methods. First, we invited a core researcher, responsible for generating the human themes using the same dataset (see AAOCA patients focus group Transcript Dataset section), to provide qualitative feedback regarding the accuracy and helpfulness of the LLM-generated themes and descriptions. We also utilize embedding similarity and LLM judgment methods to evaluate the similarity between the LLM-generated and human-generated themes and theme descriptions.

Similarity Metrics. We employed both embedding and LLM-based metrics to validate the LLM-TA pipeline by evaluating the similarity between human and LLM-generated descriptions. For each method, pairwise similarity scores between theme descriptions were aggregated into similarity matrices. Following Mathis et al. (2024), we converted continuous similarity scores into binary classifications (similar or not) by setting a threshold through sensitivity analysis (Figure 3). Based on the similarity matrices, we calculated the *Jaccard Similarity* and the *Hit Rate*. These metrics provide insights into the similarity between human-generated themes H and LLM-generated themes L. The Jaccard Similarity quantifies the proportion of similar theme pairs while the Hit Rate indicates how many human themes are adequately captured by the LLM.

Embedding-Based Semantic Similarity. We employed the sentence transformer models: `all-MiniLM-L6-v2`, `all-mpnet-base-v2`, and `sentence-T5-xxl`, to encode the theme descriptions into high-dimensional embeddings. Pairwise cosine similarity scores were computed between human-generated and LLM-generated embeddings.

LLM-Based Similarity. The LLM acts as a judge to assess the similarity of ideas between human-generated and LLM-generated themes (Liu et al. 2024). For each pair of themes, the LLM assigned a similarity score between 0 and 1 based on its understanding of the conceptual overlap, with a score of 0 meaning completely different and 1 meaning completely overlapping. We prompted the LLM to apply a penalty if one description is very specific and the other is very general, ensuring a more balanced evaluation.

Baselines. We used the method proposed by Mathis et al. (2024) as our baseline, which incorporates iterative refinement, chain-of-thought, and Reflexion prompting. We also performed a simplified version of their method without the Reflexion module, while keeping all other settings identical.

Experimental Settings. For all our proposed methods, we use OpenAI's `GPT-4o-mini-2024-07-18` with a temperature of 0 for reproducibility. We build the pipeline using `langchain v0.3.21`. For the baseline methods proposed by Mathis et al. (2024), we follow their strategy and set the temperature to 1.0.

AAOCA Patients Focus Group Transcript Dataset. The de-identified transcript corpus was derived from nine focus group sessions involving 42 parents of children with AAOCA. These sessions were lightly moderated by a non-clinical facilitator, encouraging open discussions that allowed parents to express previously unarticulated needs and experiences related to living with the condition. The nine transcripts had a median word count of 11,457. Using traditional inductive TA methods, three study team members independently coded the transcripts and developed themes based on the data. They engaged in iterative discussions to review and refine codes and identify themes. This process was both time- and resource-intensive and required approximately 30 person-hours. The final analysis identified twelve meaningful outcomes for individuals and parents.

For this study, we used these nine deidentified transcripts as our dataset. Additionally, one of the experts from the original inductive TA team collaborated closely with us to evaluate the LLM-generated results. The twelve themes representing meaningful outcomes, as articulated by the participants and identified by the human research team, serve as the ground truth data. The theme titles are listed in Figure 3.

4 Results

Efficiency Analysis. Conducting inductive TA on nine lengthy transcripts required approximately 30 person-hours for human researchers. In contrast, the LLM-TA pipeline completed the task in under 10 minutes for all methods except the 1-shot + Reflexion approach, which took 90 minutes (1.5 hours). Even this most time-intensive method reduced task duration by 97% compared to manual analysis. While some quality gaps remain between LLM- and human-generated themes, the results highlight the potential for integrating the LLM-TA method into human-led inductive TA workflows to enhance scalability and efficiency.

Quantitative Evaluation. *Jaccard Similarity* and *hit-rate* were used for evaluation. Table 1 summarizes the comparison between human- and LLM-generated themes, utilizing five prompting and chunking strategies. Two baseline methods were replicated from the (Mathis et al. 2024) study, alongside three variants of our proposed LLM-TA pipeline. Our approaches showed clear improvements across both metrics. For *Jaccard Similarity*, the zero-shot strategy achieved the highest score in three of four evaluation methods, as well as the highest overall average. For *hit-rate*, the one-shot method performed the best, obtaining the highest or joint-highest scores across all evaluation methods. Performance declined slightly in the one-shot + Reflexion setting. Compared to baseline methods, our pipeline achieved a 216% improvement in average *Jaccard Similarity* and 45% improvement in average *hit-rate*.

Two key factors underpinned the superior performance of our approach. First, the chunking strategy played a pivotal role. Baseline methods (Mathis et al. 2024) analyzed shorter transcripts with a median length of 3,200 words and did not utilize chunking. While this approach is adequate for smaller datasets, it struggles with large, real-world transcripts. By dividing longer transcripts into manageable chunks, our method prevented information loss and facilitated the extraction of more nuanced insights, thereby improving theme

	Similarity Metrics	Human-labeled themes vs. LLM-labeled themes Jaccard Similarity	Hit Rate
Mathis et al. (2024) (w/o Reflexion)	sentence-t5-xxl (> 0.82)	0.111	0.667
	all-mpnet-base-v2 (> 0.62)	0.146	0.750
	all-MiniLM-L6-v2 (> 0.56)	0.097	0.500
	LLM as a judge (> 0.5)	0.056	0.667
Mathis et al. (2024) (w/ Reflexion)	sentence-t5-xxl (> 0.82)	0.139	0.583
	all-mpnet-base-v2 (> 0.62)	0.139	0.583
	all-MiniLM-L6-v2 (> 0.56)	0.083	0.417
	LLM as a judge (> 0.5)	0.042	0.750
LLM-TA (0-Shot)	sentence-t5-xxl (> 0.82)	**0.410**	**1.000**
	all-mpnet-base-v2 (> 0.62)	**0.410**	**0.917**
	all-MiniLM-L6-v2 (> 0.56)	**0.389**	**0.917**
	LLM as a judge (> 0.5)	0.118	0.750
LLM-TA (1-Shot)	sentence-t5-xxl (> 0.82)	0.396	**1.000**
	all-mpnet-base-v2 (> 0.62)	0.326	**0.917**
	all-MiniLM-L6-v2 (> 0.56)	0.285	**0.917**
	LLM as a judge (> 0.5)	**0.174**	**0.917**
LLM-TA (1-Shot + Reflexion)	sentence-t5-xxl (> 0.82)	0.222	**1.000**
	all-mpnet-base-v2 (> 0.62)	0.291	0.833
	all-MiniLM-L6-v2 (> 0.56)	0.215	0.833
	LLM as a judge (> 0.5)	0.152	**0.917**
Ground Truth (Human)	sentence-t5-xxl (> 0.82)	0.583	1.000
	all-mpnet-base-v2 (> 0.62)	0.486	1.000
	all-MiniLM-L6-v2 (> 0.56)	0.500	1.000
	LLM as a judge (> 0.5)	0.104	1.000

Table 1: Performance comparison of human-generated and LLM-generated themes and descriptions. Jaccard similarity and Hit rate are used to measure Human-labeled themes vs. LLM-labeled themes based on pair-wise similarity in Figure 3.

Methods	Similarity with Human Coding Concepts Level	Theme Level	Specificity	Usefulness
Mathis et al. (2024) (w/o Reflexion)	High	1 (Low)	1 (Low)	Not Very Helpful
Mathis et al. (2024) (w/ Reflexion)	High	1 (Low)	2 (Medium)	Not Very Helpful
LLM-TA (0-Shot)	High	2 (High)	4 (High)	Helpful
LLM-TA (1-Shot)	High	2 (High)	3 (Moderate-High)	Moderately Helpful
LLM-TA (1-Shot + Reflexion)	High	3 (Highest)	5 (Highest)	Helpful

Table 2: Expert evaluation.The LLM-TA pipeline outperformed baselines in thematic similarity, specificity, and usefulness.

generation. Second, the use of detailed prompts significantly enhanced outcomes. Unlike the generic prompts used in baseline methods, our prompts explicitly outlined the research context, objectives, and task-specific instructions, such as generating theme names, descriptions, and identifying relevant quotes. This precision enabled the LLM to produce outputs more aligned with human-generated themes. Moreover, in the one-shot setting, we further enriched contextual understanding by incorporating a real-world example from Inductive TA conducted on similar transcripts. This example clarified expectations and reinforced the LLM's ability to structure outputs effectively.

Expert Evaluation. To assess the LLM-generated themes, we engaged a human researcher who had previously conducted the ground truth inductive TA on the same AAOCA dataset. This expert's domain familiarity enabled a precise evaluation of thematic accuracy and relevance.

Table 2 highlights the expert's overall evaluation. The LLM-TA pipeline outperformed baselines in thematic similarity, specificity, and usefulness. However, the expert noted limitations in the utility of themes generated by the one-shot LLM-TA method, despite its superior hit rate. In contrast, themes produced by the zero-shot setting were qualitatively more useful, underscoring the insufficiency of similarity-based metrics for assessing thematic quality.

The expert also identified some shortcomings in LLM-generated themes. **(1) Lack of representativeness.** Certain themes overemphasized rare, dramatic experiences rather than reflecting broader trends. For instance, the theme *"Finding relief in a diagnosis after health crises"* disproportionately highlighted cardiac arrest cases, whereas most parents of incidentally diagnosed children reported heightened anxiety instead of relief. Similarly, themes like *"Managing the financial challenges of care"* overrepresented financial concerns, which were rarely mentioned across transcripts. **(2) Inaccurate interpretations.** Some themes misrepresented transcript content. For example, the theme *"Seeking simplicity in discussing my child's heart condition"* misinterpreted parents' preferences for a comprehensive understanding as a desire for simplified information. Similarly, the overly broad theme *"Desiring clear communication and understanding from medical professionals"* conflated distinct

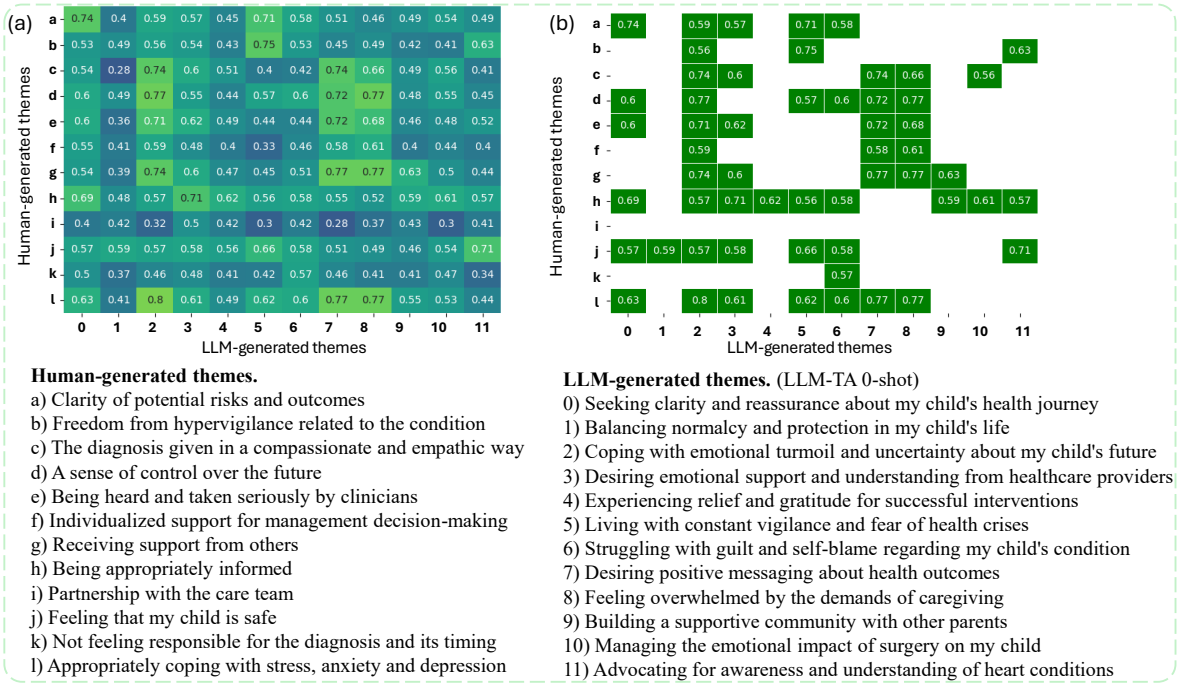

Figure 3: (a) Similarity heatmap of themes generated by the LLM-TA (zero-shot) method compared with human themes using all-MiniLM-L6-v2. (b) Similarity heatmap of themes generated by the LLM-TA (zero-shot) method compared with human themes using all-MiniLM-L6-v2, with a threshold score of 0.56.

issues such as empathy and clinical communication into a single category. **(3) Missing clinical context.** The LLM lacked knowledge of clinical context external to the transcripts, such as long-term outcomes for AAOCA or management strategy nuances. For example, parents' desire for more information reflects a lack of clinical evidence on optimal management strategies of AAOCA, rather than mere gaps in communication. Those clinical contexts are not in the transcripts, and providing the model with such clinical context could improve the accuracy of future LLM-generated themes.

These findings underscore the necessity of close collaboration with domain experts to iteratively refine prompts based on research goals and contextual knowledge. High-stakes datasets often require expertise-derived context beyond what transcripts explicitly provide, which LLMs inherently lack. Our results reinforce the irreplaceable role of expert human evaluation in TA, particularly in complex healthcare datasets. While embedding-based and LLM-based similarity metrics provide valuable quantitative insights, only human reviewers can, as of now, qualitatively assess thematic relevance and accuracy.

5 Conclusion

Motivated by the need to reduce the time-consuming inductive TA process in complex healthcare transcript data, we designed an LLM-enhanced inductive TA pipeline to automate stages (1) to (5) of the traditional TA process. Using chunking, we performed fine-grained coding and theme identification. Additionally, we developed detailed instructions on the steps of TA and the research goals, specifically regarding the meaningful outcomes for parents of children with AAOCA, and incorporated these into various LLM prompting strategies. By evaluating the pipeline with domain expertise and LLM-based metrics, we significantly improved the performance of existing LLM-enhanced TA methods. We emphasize that the evaluation of LLM-enhanced inductive TA on high-stakes data must involve domain experts. While automated methods can assist in analyzing correlations and provide quantitative insights, expert evaluation remains irreplaceable and invaluable in ensuring the accuracy and relevance of themes derived from such datasets.

Through close collaboration with an inductive TA expert, we identified key limitations of current LLM-generated themes, primarily arising from the lack of clinical context that domain experts possess and cannot be directly inferred from transcripts. Based on these findings, we strongly recommend that future researchers collaborate closely with domain experts to design prompts and interpret results at each stage of TA. To further enhance reliability, future work should: (1) engage multiple experts to mitigate individual biases and broaden interpretive perspectives, (2) incorporate clinical guidelines or regulatory documents into prompts to enrich LLMs' contextual understanding, and (3) validate the pipeline across diverse medical conditions to assess generalizability. Such steps will help fully harness LLMs' potential to scale inductive TA while preserving analytical rigor in sensitive healthcare domains.

Acknowledgements

We would like to acknowledge the following funding supports: NIH OT2OD032581, NIH OTA-21-008, NIH 1OT2OD032742-01.

References

Bradford, N. J.; Rider, G. N.; Catalpa, J. M.; Morrow, Q. J.; Berg, D. R.; Spencer, K. G.; and McGuire, J. K. 2020. Creating gender: A thematic analysis of genderqueer narratives. In *Non-binary and Genderqueer Genders*, 37–50. Routledge.

Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2): 77–101.

Braun, V.; Clarke, V.; and Weate, P. 2016. Using thematic analysis in sport and exercise research. In *Routledge handbook of qualitative research in sport and exercise*, 213–227. Routledge.

Dai, S.-C.; Xiong, A.; and Ku, L.-W. 2023. LLM-in-the-loop: Leveraging Large Language Model for Thematic Analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 9993–10001.

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.

Deiner, M. S.; Honcharov, V.; Li, J.; Mackey, T. K.; Porco, T. C.; and Sarkar, U. 2024. Large Language Models Can Enable Inductive Thematic Analysis of a Social Media Corpus in a Single Prompt: Human Validation Study. *JMIR infodemiology*, 4(1): e59641.

Drápal, J.; Westermann, H.; and Savelka, J. 2023. Using Large Language Models to Support Thematic Analysis in Empirical Legal Studies. arXiv:2310.18729.

Gao, J.; Guo, Y.; Lim, G.; Zhang, T.; Zhang, Z.; Li, T. J.-J.; and Perrault, S. T. 2024. CollabCoder: a lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–29.

Gebreegziabher, S. A.; Zhang, Z.; Tang, X.; Meng, Y.; Glassman, E. L.; and Li, T. J.-J. 2023. PaTAT: Human-AI Collaborative Qualitative Coding with Explainable Interactive Rule Synthesis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394215.

Guetterman, T. C.; Chang, T.; DeJonckheere, M.; Basu, T.; Scruggs, E.; and Vydiswaran, V. V. 2018. Augmenting Qualitative Text Analysis with Natural Language Processing: Methodological Study. *J Med Internet Res*, 20(6): e231.

Holloway, I.; and Todres, L. 2003. The status of method: flexibility, consistency and coherence. *Qualitative research*, 3(3): 345–357.

Katz, A.; Fleming, G. C.; and Main, J. 2024. Thematic Analysis with Open-Source Generative AI and Machine Learning: A New Method for Inductive Qualitative Codebook Development. *arXiv preprint arXiv:2410.03721*.

Khan, A. H.; Kegalle, H.; D'Silva, R.; Watt, N.; Whelan-Shamy, D.; Ghahremanlou, L.; and Magee, L. 2024. Automating Thematic Analysis: How LLMs Analyse Controversial Topics. arXiv:2405.06919.

Leeson, W.; Resnick, A.; Alexander, D.; and Rovers, J. 2019. Natural Language Processing (NLP) in Qualitative Public Health Research: A Proof of Concept Study. *International Journal of Qualitative Methods*, 18: 1609406919887021.

Liu, J. L.; Wang, Y.; Lyu, Y.; Su, Y.; Niu, S.; Xu, X. O.; and Zhang, Y. 2024. Harnessing LLMs for Automated Video Content Analysis: An Exploratory Workflow of Short Videos on Depression. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW Companion '24, 190–196. New York, NY, USA: Association for Computing Machinery. ISBN 9798400711145.

Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2023. Lost in the Middle: How Language Models Use Long Contexts. arXiv:2307.03172.

Mannstadt, I.; Goodman, S. M.; Rajan, M.; Young, S. R.; Wang, F.; Navarro-Millán, I.; and Mehta, B. 2024. A Novel Approach for Mixed-Methods Research Using Large Language Models: A Report Using Patients' Perspectives on Barriers to Arthroplasty. *ACR Open Rheumatology*, 6(6): 375–379.

Mathis, W. S.; Zhao, S.; Pratt, N.; Weleff, J.; and De Paoli, S. 2024. Inductive thematic analysis of healthcare qualitative interviews using open-source large language models: How does it compare to traditional methods? *Computer Methods and Programs in Biomedicine*, 255: 108356.

Mery, C. M.; Well, A.; Taylor, K.; Carberry, K.; Colucci, J.; Ulack, C.; Zeiner, A.; Mizrahi, M.; Stewart, E.; Dillingham, C.; et al. 2023. Examining the Real-Life Journey of Individuals and Families Affected by Single-Ventricle Congenital Heart Disease. *Journal of the American Heart Association*, 12(5): e027556.

Nowell, L. S.; Norris, J. M.; White, D. E.; and Moules, N. J. 2017. Thematic analysis: Striving to meet the trustworthiness criteria. *International journal of qualitative methods*, 16(1): 1609406917733847.

Overney, C.; Saldías, B.; Dimitrakopoulou, D.; and Roy, D. 2024. SenseMate: An Accessible and Beginner-Friendly Human-AI Platform for Qualitative Data Analysis. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, 922–939. New York, NY, USA: Association for Computing Machinery. ISBN 9798400705083.

Parfenova, A.; Denzler, A.; and Pfeffer, J. 2024. Automating Qualitative Data Analysis with Large Language Models. In Fu, X.; and Fleisig, E., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, 83–91. Bangkok, Thailand: Association for Computational Linguistics.

Prescott, M. R.; Yeager, S.; Ham, L.; Rivera Saldana, C. D.; Serrano, V.; Narez, J.; Paltin, D.; Delgado, J.; Moore, D. J.; and Montoya, J. 2024. Comparing the efficacy and efficiency

of human and generative AI: Qualitative thematic analyses. *JMIR AI*, 3: e54482.

Saldana, J. 2011. *Fundamentals of qualitative research*. Oxford University Press.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Singh, S. H.; Jiang, K.; Bhasin, K.; Sabharwal, A.; Moukaddam, N.; and Patel, A. B. 2024. RACER: An LLM-powered Methodology for Scalable Analysis of Semi-structured Mental Health Interviews. *arXiv preprint arXiv:2402.02656*.

Tierney, S.; and Fox, J. R. 2010. Living with the anorexic voice: A thematic analysis. *Psychology and Psychotherapy: Theory, Research and Practice*, 83(3): 243–254.

Wang, Y.; Cai, Y.; Chen, M.; Liang, Y.; and Hooi, B. 2024. Primacy Effect of ChatGPT. arXiv:2310.13206.

Xiao, Z.; Yuan, X.; Liao, Q. V.; Abdelghani, R.; and 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*, 75–78.

Zhang, H.; Wu, C.; Xie, J.; Lyu, Y.; Cai, J.; and Carroll, J. M. 2023. Redefining qualitative analysis in the AI era: Utilizing ChatGPT for efficient thematic analysis. *arXiv preprint arXiv:2309.10771*.

Details about evaluation metrics

Let $H = \{h_1, h_2, \ldots, h_n\}$ represent the set of human-generated themes, and $L = \{l_1, l_2, \ldots, l_m\}$ represent the set of LLM-generated themes. For each pair (h_i, l_j) in $H \times L$, we compute a similarity score $s(h_i, l_j)$. We define $S_\theta = \{(h_i, l_j) \in H \times L \mid s(h_i, l_j) \geq \theta\}$, where θ is the similarity threshold determined via sensitivity analysis and kept consistent across different baselines.

The *Jaccard Similarity* is defined as the proportion of theme pairs considered similar out of all possible pairs:

$$\text{Jaccard Similarity} = \frac{|S_\theta|}{|H \times L|} = \frac{|S_\theta|}{n \times m} \tag{1}$$

The *Hit Rate* measures the proportion of human-generated themes that find a highly similar mapping in the LLM-generated themes:

$$\text{Hit Rate} = \frac{|H_s|}{n} \tag{2}$$

where $H_s = \{h \in H \mid \exists\, l \in L,\, s(h, l) \geq \theta\}$.