Transcript-to-Text Thematic Analysis: Tool Development and Evaluation

Kevin Childs
k.childs@ufl.edu
University of Florida

ABSTRACT

Qualitative research helps to take natural language from survey or interview data and develop themes, ideas, and communicable lessons in a structured format. With developments in natural language processing and large language models, the status quo for conducting qualitative research is being challenged by LLM thematic analysis tools. This paper seeks to review the efficacy of existing techniques and develop tools to explore the landscape of qualitative analysis using LLMs. To perform this analysis, we use two existing tools on internal datasets that have previous qualitative analysis performed and develop ContextTA to match closely to the six stages of thematic analysis. In doing so, we find that the state of the art is not currently matched to expert researchers; however, for smaller tasks and accounting for bias, LLMs can offer a supplemental analysis vector. It remains important to continually evaluate the efficacy of LLM-based tools and account for implications to human researchers within qualitative research.

1 INTRODUCTION

Thematic analysis is a research method used within human-computer interaction and other qualitative research fields to systematically understand content from non-uniform human feedback[3, 4] (e.g., interview transcripts). The standard method involves six stages: (1) familiarization with data, (2) coding information into small ideas, (3) identification of themes or trends within codes, (4) refining themes with respect to research questions, (5) defining and naming themes, and (6) writing the report. This process starts with detail-oriented tasks of annotating, eventually working toward combining ideas into a larger theme and manuscript.

Large language models (LLMs) are positioned as a potential disruptor to how thematic analysis is conducted. Tasks such as qualitative coding are well suited for LLMs and, in some research settings, are delegated to crowdsourced individuals [12]. In these cases, expert experience and context are necessary to develop the protocol and interpret the analysis results; however, the smaller details of annotation are delegated. Given this existing framework, it is an open question whether LLMs are capable of performing these tasks effectively and to what extent they should be used.

In this work, we evaluate two state-of-the-art transcript-to-theme LLM tools published in medical domains [10, 15]. Additionally, we develop our own tool to learn the nuances of this development format and become intimately familiar with the potential pitfalls of LLM-based thematic analysis. Ultimately, we identify that transcript-to-theme LLM tools are not a replacement for expert research, lacking the research-specific context that can contextualize results. However, when used as a supplemental tool during the theme identification stage, these tools can aid in considering alternative themes and identifying bias within research. We also identify that these tools have development pitfalls and have the potential for overfitting to the research questions and background.

We make the following contributions:

- **Develop LLM thematic analysis tool:** We develop a thematic analysis tool called ContextTA that leverages contextual integrity. In developing this tool, we propose framework-specific analysis that can perform supplemental analysis. In evaluation, we recognize open opportunities to improve this tool for flexibility across different formats of data.
- **Evaluate LLM thematic analysis tools:** We evaluate two separate LLM thematic analysis tools in comparison to our own, noting how these tools are developed, the economic and environmental costs, and the accuracy of themes compared to prior research. We find that tools are not suited for identifying novel contributions, but still provide valuable insights.
- **Propose future frameworks:** We propose future work leverage LLMs for thematic analysis with a skeptical lens. We identify that over-reliance on LLMs may limit the value of novel research, but it may help identify missed themes and serve as an unbiased third party.

The remainder of this paper is organized as follows: Section 2 provides the necessary background; Section 3 describes the methodology for testing and developing our thematic analysis tool; Section 4 outlines the results with tool evaluation, usage analysis, and comparison with literature; Section 5 discusses the results; and Section 6 concludes.

2 BACKGROUND

Large language models (LLMs) are an emerging technology, providing new opportunities in software development and productivity, while presenting threats to education and many jobs. The research community is not immune to these impactful tools, with opportunities emerging in qualitative and big data analysis, while threats are emerging to the validity of some crowdsourced outputs [11]. This work explores techniques for effective LLM usage, their effectiveness in practice, and considerations and pitfalls when using LLMs. We conclude with a discussion on how LLMs are being used by the crowdsourced population and a big-picture discussion on their current impacts.

Effective use of LLMs starts with understanding the capabilities of the technology. A key area of proficiency is with text-related tasks and natural language processing. On paper, these goals align with many qualitative analysis techniques, specifically those performed in crowdsourcing [13]. Crowdsourcing techniques leverage a large pool of human workers to perform small tasks, allowing for time-consuming tasks to be completed rapidly. For example, it has been proven that identifying different elements of contextual integrity can be performed by crowd workers, effectively distributing the task of

reviewing privacy policies to human workers [12]. These simple, bite-sized tasks are perfectly suited for artificial intelligence. The combination of generating, evaluating, merging, improving, focusing, and partitioning are techniques to create an effective pipeline that can leverage the scale of crowd workers and translate to simple prompts [6]. Among these tasks, LLMs fall short when it comes to producing unique and culturally representative answers, and for evaluation, there is a bias toward affirmation that in some cases can be difficult to overcome. Given this, guidelines for using LLMs in qualitative tasks begin with creating a structured pipeline with both human and AI agents, keeping prompts brief, and evaluating whether the human element is necessary. When giving prompts, there should be examples of successful identification of codes or tasks.

These approaches have been applied in two of our selected studies. In 2023, Xiao et al. [14] used GPT-3 to study the capability of LLMs to perform deductive coding. While these results clearly demonstrated that LLMs were not as effective as human coders, more recently, results presented at CHI 2025 [1] show higher efficacy, demonstrating that for a ranking task (ranking the worst experience described by a quote), GPT-4 and Gemini perform similarly to humans. Between these papers, many improvements are represented. First, the underlying models have rapidly improved in response quality, which suggests even more potential to overcome shortcomings such as validation struggles in critical tasks. Second, there has been a change in approaches: the first study provided a whole codebook as a prompt, while the more successful study asked simple questions such as "Which participant quote represents a more positive experience?" Translating back to the codebooks, this suggests that simple, clear codebooks—or even breaking down coding tasks—can be an effective approach, while blanket coding tasks for large data sets may prove too vague to implement successfully.

Within the crowdsourcing space, researchers appear to have options: is this a crowdsourcing task, or can it be performed by artificial intelligence? In reality, the choice may not be that simple, as crowdsourcing humans may be implementing AI within their task performance. In both conversation and review of the literature [11], Bentley of Google described how crowdsourcing users are using LLMs to qualify for research studies. This can reduce cognitive load and increase earning potential for users while coexisting with ethical issues on the users' part. While some use it to create whole ideas, others use it to overcome language barriers and refine responses to questions. The current best approach is to ask users to agree to acceptable use statements, motivate participants not to use AI through appeals to the research mission, and outline potential consequences. When high assurance that AI is not being used is required, having a face-to-face Zoom call—or, as deepfake techniques improve, in-person conversations—will become increasingly important.

When reviewing the impact of LLMs on qualitative research, we find that effective implementation can have promising results while representing consequences to academic validity when not properly executed. In the qualitative space, there has always been a need to be skeptical of interview questions, variance among coders, and the high level of subjective reporting. Future work using AI adds a layer of scrutiny toward AI implementation. As described by Eschrich et al. [5], much of this has followed a positivist approach—trying to understand how new technology can slot into existing workflows for improvements. There will also be more novel research methodologies created using a constructivist approach, where new technology allows for research methods previously not possible or conceivable. Representing a changing field, there is value in being fluid and open to new ideas, but there also needs to be an understanding of how technology impacts researchers. AI unlearning is being studied, representing that reliance on technology may reduce proficiency. From a research perspective, in many qualitative tasks, there is benefit to spending time with the data. A researcher is not only an individual who touches the frontiers of knowledge; they are also stewards of this new information. The potential for LLMs and AI to generate new insights with proper techniques is present, but ensuring that the human in the loop can be present—sit with the data, ruminate, wrestle, and foster a passion for what new information means—should not be lost and, above all else, should be maintained.

3 METHODOLOGY

Within this section, we discuss the development methodology for our LLM thematic analysis tool as well as the evaluation against two open-source tools. This forms the basis for the evaluation that we outline in Section 4, the results.

3.1 Tool Development - ContextTA

In developing ContextTA, we followed Braun and Clarke's six stages of thematic analysis[1]. This is represented visually in Figure 1. The code base is designed as a pipeline where the input from each stage is piped into the next stage. During the theme reduction stage, the output from each transcript is combined.

Familiarization: The familiarization stage in traditional thematic analysis involves reading and note-taking for transcripts. When translated to a computational version, we mark this stage as data formatting and preparation. The input is split into individual paragraphs and stored within a dictionary for later analysis. The splitting, coding, and theme identification are done in parallel to increase efficiency. Refinements to this process are possible with labeling the subject and interviewer; however, to remain flexible to different input formats that researchers may use, we did not implement this step.

Coding: Our novelty in this research lies in the use of contextual integrity, specifically leveraging Kumar et al.'s [7] framework for integrating contextual integrity and thematic analysis. To perform this analysis, information is coded based on the subject, sender, receiver, information type, and transmission principles. From these codes, themes of use can be developed. To implement this in code, we use a sliding window of paragraphs to determine if they describe the flow of information in a computer system. This is determined by having an LLM output which, if any, of five paragraphs represent an information flow numerically. Five paragraphs are chosen to provide adequate context. The window is sliding—paragraphs 1–5 are evaluated, then 2–6, 3–7, and so on. Each paragraph receives five votes, and if the total number of votes exceeds 2.5, it is considered for further analysis. In this further analysis, the paragraph is then coded based on the six parameters of contextual integrity, noting both descriptions and raw quotes of different elements.

[1]The full code base is attached in the companion zip file

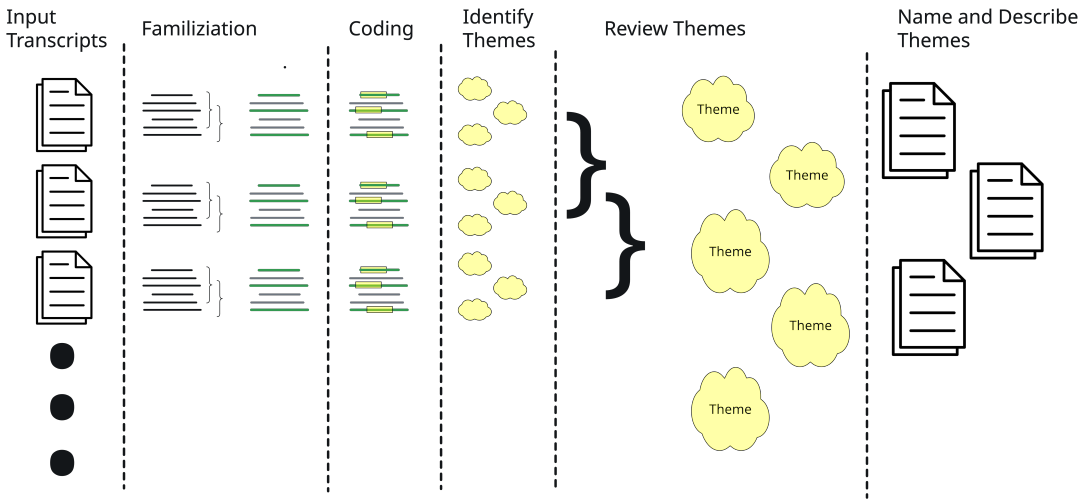

Figure 1: Our transcript-to-theme LLM analysis tool (ContextTA) follows five of the six thematic analysis steps, leaving the final stage of writing a document to the user of the tool. Familiarization is represented by processing and organizing the files, coding is conducted using the contextual integrity framework, and the combination of themes, refining, and naming is done by converging ideas into larger groups and across transcripts.

Theme Identification: These descriptions and quotes related to contextual integrity elements are then viewed holistically across a transcript to identify themes. These themes are refined to reduce overlap, and are then combined during the theme reduction stage. Quotes representing these themes are maintained.

Theme Reduction: In the theme reduction stage, each theme is evaluated to see if it matches others in the list. If the theme is unique, it is considered a final theme; if it can be combined with an existing final theme, then the themes are merged. A threshold of 10 themes is enforced, with the least relevant themes removed or combined. The best quotes representative of the themes are retained.

Naming and Defining Themes: The 10 final themes, along with their descriptions and quotes, are finally interpreted into short paragraphs for communication and easy identification. These paragraphs represent the final output of the transcript-to-theme LLM tool.

3.2 Tool Evaluation

For this research, we selected two open-source LLM qualitative analysis tools [10, 15]. The tools we evaluate are required to have stated goals of performing thematic analysis on qualitative interview transcripts. Interview transcripts are typically conversations between two individuals but can also occur in focus groups or other multi-party scenarios. For our evaluation, we focus on the base case involving one interview subject and one interview participant. The evaluation material is drawn from the first author's prior work. Each of the following two studies involves one researcher and one participant discussing technology usage and privacy. The studies are as follows:

Research study 1 [2]: This research study included 23 individuals, each of whom had experienced some uncomfortable situation related to continuous location-sharing applications. These participants were asked a series of questions related to their current and/or former usage of continuous location-sharing applications, what motivated their discomfort with these applications, and suggestions related to those experiences. The findings and themes centered around three ideas: that some individuals felt minor discomfort, others experienced ongoing discomfort, and some reported more significant, life-changing experiences.

Research study 2 [9]: This research study included 20 individuals, each of whom were elite collegiate athletes. These participants were asked about what technology they personally use, what their coaches use, the challenges or benefits of such technologies, and how technology should be used within athletics. The overall themes centered around four engagement types: directive, reflective, non-engagement, and coach-delegated usage of technology. Additional themes emerged around the use of multiple forms of one technology for the same metric, as well as the mental aspect of technology usage.

Our evaluation is based on three different criteria. First, we evaluate what is necessary to run the tools: does the research tool require prerequisite information or can it operate independently? Second, we evaluate how these applications function. Beyond the underlying prompt engineering, we consider estimated power usage and time to complete tasks using the same testbed. Finally, we evaluate the efficacy of the results in comparison to the research led by the first author. The themes are evaluated based on their accuracy and level of insight.

4 RESULTS

Within this results section, we evaluate three models (AutoThemeGenerator [15], LLM-TA [10], and ContextTA (our tool)). We evaluate these tools across three vectors: reproducibility, token usage and runtime, and performance compared to baseline thematic analysis performed by the lead author prior to this project.

Table 1: Token Usage, Inference Time, Energy Consumption, Cost, and Throughput Across Models

Model	Use Case	Tokens	Time	Tokens/sec	Total Joules	Price (USD)	Phone Charges
AutoThemeGenerator	Student-Athletes	332713	33:44	164.38	64879	0.33	0.90
AutoThemeGenerator	Continuous-Sharing	285564	25:52	184.00	55684	0.29	0.77
LLM-TA	Student-Athletes	325833	34:17	158.40	63537	0.33	0.88
LLM-TA	Continuous-Sharing	215277	41:52	85.70	41979	0.22	0.58
ContextTA	Student-Athletes	5467200	1:46:05	858.95	1066104	5.47	14.81
ContextTA	Continuous-Sharing	5888400	1:49:12	898.72	1148238	5.89	15.95

4.1 Reproducibility and Extendability

AutoThemeGenerator is a strong exemplar of extendability. There is a comprehensive README file, as well as input variables that are left empty with examples for the researcher to replace. These variables focus on (1) the background, (2) the research questions, and (3) the interview questions. Once these variables were input, the default model was changed, and an API key was inserted, the program ran effectively. The terminal output did not accurately display an estimated time, producing a question mark instead. While easily reproducible and extendable to other research questions, the tool requires contextual input. This raises a question about the integrity of the output and if it is overfitting to the data. This inspired the development of ContextTA to use a universal prompt and conduct analysis purely based on the content, independent of provided context.

LLM-TA provides a strong case for reproducibility without being easily extendable. The files provided were able to run successfully and output results that are comparable with the academic publication. However, when attempting to adapt the tool to our research use case, there were deeply embedded prompts containing the original author's context. As a result, we had to review each prompt and replace the original context with information relevant to our research. Additionally, this program required formatted output from the LLM, and while the tools mostly adhered, there were cases where output was lost or could not be parsed. The first test did not complete due to a JSON parsing error. This highlights that LLM-based tools should not rely heavily on parsing, once again informing the design of ContextTA.

Based on the lessons learned from these tools, ContextTA was built to run across different datasets without the need for reconfiguration. For reuse, researchers only need to update API keys and models. Outside of that, the code should be reproducible. Further testing by an external committee would be necessary to confirm this.

4.2 Basic Analytics

For basic analytics, we measured token usage, runtime, and estimated electricity cost. The estimated electricity cost per token is 0.195 Joules [8]. We used the Navigator Toolkit API from the University of Florida, priced at $1 per 1 million tokens. All research was conducted using llama3-70b parameter. Based on this information, we provide a breakdown of usage and cost for each of the three models across both datasets (Table 1).

Reviewing the table, we observe that AutoThemeGenerator and LLM-TA use approximately 300,000 tokens, which is roughly equivalent to analyzing 20 research articles. ContextTA, in contrast, uses an order of magnitude more tokens due to its multi-stage filtering and a resource-intensive theme-combination process with an On^2 runtime. Removing annotation stages and reducing text per iteration helped make the other tools more efficient. ContextTA achieves much higher throughput via parallelization.

When reviewing the electricity cost, it reveals that despite a large amount of processing time, these tools can be efficent with electricty, using less than a phone battrery's worth of energy for a full run. This does not include the ecological cost of creating and training these initial datasets. Additionally, other models, such as ChatGPT 4o is an order of magnitude more energy [8].

4.3 Model Performance

To evaluate each tool, the lead author revisited two datasets they previously analyzed using thematic analysis and compared tool outputs to the findings from the resulting academic publications.

With AutoThemeGenerator, the themes were not fully representative of the original findings. For the student-athletes study, our research team identified four patterns of use and a focus on the mental aspects of sports. The tool recovered three of the four relevant cases. For the continuous-sharing study, the resulting themes were overly general and failed to capture the distinct categories identified in the original study, namely minor, ongoing, and life-altering discomfort. The tool instead reported on average experiences.

When evaluating LLM-TA, we encountered a formatting error in the first test run. After 30 minutes, the program failed due to strict formatting requirements. We extracted the 34,000-character output and used GPT-4o to clarify the findings. The generated themes closely aligned with our research,mentioning barriers to data engagement and a holistic view of physical, mental, and lifestyle metrics. However, these themes aligned more with the research and interview questions than the participant data itself. Notably, although the theme development prompt had the correct dataset context, the coding prompt still referenced collegiate athletics. Consequently, the final codes included mentions of student-athletes despite their absence in the dataset. This hallucination points to overfitting the themes to the provided context rather than deriving them from the data.

ContextTA performed similarly to the other tools in identifying valid themes. Its most valuable asset was its quote identification

feature. During development, representative quotes were preserved, helping to uncover relevant quotes that didn't make it into the final publications. This retention provides a helpful way to map themes back to specific participant responses. A major flaw in ContextTA was that it extracted information from both the participant and the researcher. Some representative quotes and themes reflected interview questions rather than participant sentiment. Additionally, the focus on data exchange helped generate more technologically-relevant themes compared to other tools, which surfaced more varied and sometimes tangential themes.

5 DISCUSSION

Within this section, we discuss the development considerations when working with LLMs, the risks of transcript-to-theme LLMs to research integrity, and the potential benefits these tools offer for qualitative analysis. We conclude with a reflection on this summer independent study course and its limitations.

5.1 Development Pitfalls with LLMs

In each of the two tested tools, as well as in the initial iteration of our internal development, transcript-to-theme LLM tools are specifically tailored to the research questions, interview content, and researcher-provided context. The model by Yang et al. [5] had this functionality explicitly stated as an input parameter, whereas Raza et al. [10] embedded this within their prompts. During the first round of development, we also followed this methodology to better extract content relevant to our research questions around technology in athletics.

Including this information was helpful in generating contextually relevant responses, consistent with prior work by Shvartzshnaider et al. [12], where short, relevant prompts produced more relevant results. However, this methodology introduces a risk of researcher bias. A core principle of thematic analysis is maintaining fidelity to the data without inserting undue bias. There is a risk of overfitting themes to the questions themselves, for instance, asking about risks and benefits may naturally lead to identifying a theme about trade-offs. As a proof of concept, we used the context from one study and paired it with the data from a different study. The resulting themes reflected the context, not the data. Initially, we were unsure whether the correct dataset had been used until further analysis revealed the prompt was the source of the suprising results. This represents a serious concern regarding the validity of these tools.

In addition, although best practices recommend keeping prompts short and focused [5], both tested tools often included thousands of tokens per request. Including more context can improve relevance and reduce the number of required requests, but it also increases the likelihood of hallucinations and off-topic responses. As LLM tools continue to evolve, it remains an open question whether small or large context windows are more effective.

5.2 Risk to Research Integrity

Academic research aims to generate new knowledge and effectively communicate that knowledge to the broader community. While transcript-to-theme LLMs may improve and eventually produce compelling and plausible themes, there is a risk of distancing researchers from their data. The first step in thematic analysis is familiarization, becoming immersed in the data. This immersion helps researchers understand and communicate their findings. As these tools become more prevalent, researchers may become increasingly disconnected from the data, potentially undermining the educational and interpretive value of qualitative research.

Additionally, transcript-to-theme tools require validation. A well-known characteristic of LLMs is that their output often sounds realistic, much like SCIGen, which can generate grammatically correct but meaningless scientific papers. With these tools, outputs are not only fluent but also contextually relevant. However, there's a risk that these results might be hallucinated or based on prior assumptions embedded in the model's training, rather than grounded in the research data itself.

5.3 Supplemental Analysis With LLMs

Our transcript-to-theme LLM tool is specifically designed to analyze elements of contextual integrity. This lens emphasizes the flow of information within technological systems and highlights relevant stakeholders, offering a different perspective from traditional thematic analysis. Contextual integrity is just one framework; using multiple frameworks across multiple tools may help researchers interpret data more deeply and in more structured ways, ultimately enhancing traditional thematic analysis.

Outside of frameworks, these tools also offer value in challenging existing expectations post-analysis. Researchers can inadvertently introduce bias, emphasizing familiar themes while overlooking others. While this human perspective is valuable, being aware of this bias is crucial in qualitative research. LLMs offer an opportunity to surface alternative or underrepresented themes, serving as a secondary lens through which researchers can review their interpretations. Transcript-to-theme LLMs should be viewed as supplemental tools that augment, but do not replace, traditional thematic analysis.

5.4 Lessons Learned

From a personal standpoint, this project has helped me understand how to develop and work with LLMs, especially within the transcript-to-theme space. This experience provided hands-on exposure to the limitations of models such as the LLaMA 70B parameter version, which we used for testing. I was impressed by the model's accuracy in identifying paragraphs related to continuous wearable technology. In my own testing, the accuracy of paragraph identification was nearly 100%. I was also impressed by the annotation capabilities of these tools. However, for more complex tasks such as theme creation and combination, the tools were less successful. For example, a human might label a theme as *athlete rumination from technology*, whereas an LLM may classify it more generically as *athlete technology usage*.

When combining codes across transcripts, themes often aligned closely with research questions or anticipated insights, rather than identifying novel information. A human researcher may detect subtle and unexpected insights due to their deeper familiarity with the field and data. LLMs lack this context; instead weighting all tokens in the analysis, even if they align with known theories.

In practice, I do see value for LLMs at smaller scales. Qualitative research often includes repetitive tasks that lead to mental fatigue, increasing the likelihood of errors. LLMs can support researchers by

handling these lower-level tasks, allowing researchers to focus their energy on higher-order thinking. For instance, in a recent qualitative study on continuous wearable devices, I manually annotated information flows across dozens of transcripts. As the task continued, the quality of my work declined due to fatigue. In the future, LLMs could perform such annotation tasks, allowing me to focus on the later, more interpretive stages of thematic analysis.

5.5 Limitations and Future Work

In this study, we reviewed two open-source tools and developed our own. Given the rapid pace of development in this field, other academic tools have emerged since our evaluation. As a result, we do not claim that our tools are the most up-to-date, but rather that they provide a representative snapshot of the state of the art at the time of testing. Moving forward, measurement platforms and standardized testbeds could help improve consistency in tool evaluation and development.

We are also limited in the breadth of LLM backends tested. Due to the sensitive and potentially identifiable nature of our reference datasets, our analysis could only be conducted locally or on secure internal servers. As such, we were limited to testing with the LLaMA 70B model provided through the UF Navigator Toolkit.

This work also opens the door for future research in several areas:

(1) **Adversarial testing:** Use the same dataset with different research questions and contexts to validate whether LLMs generate themes based on input data or on the provided context alone.
(2) **Development of a standardized testbed:** Create a common input/output format to promote consistency and facilitate community-driven benchmarking.
(3) **Further development of ContextTA:** Explore the small-context approach used in ContextTA and improve it through more rigorous engineering and testing.
(4)
(5) **Ecological analysis of LLM tools in academic settings:** Compare transcript-to-theme tools across datasets and model configurations to better understand the power usage, and ecological impacts of using such tools.
(6) **Annotation support:** Investigate how LLMs can be used to annotate qualitative data in a way that supports researchers without detaching them from the source material, effectively elevating the intelect of researchers opposed to replacing them.

6 CONCLUSION

Within this research, we evaluate and develop transcript-to-theme thematic analysis tools that leverage large language models. In doing so, we identify both challenges and opportunities. Notably, the tendency of outputs to overfit to researcher-provided context, and runtime and efficiency concerns. While these issues present real limitations, there are promising capabilities, particularly in quote identification and contextual relevance. In response to these findings, we developed ContextTA, a tool designed to leverage the contextual integrity framework without overfitting to background context. Our tool is capable of producing relevant themes grounded in the data, without requiring user-provided background inputs. These themes often center around technology usage patterns but still struggle to distinguish between participant and interviewer response. Ultimately, state-of-the-art thematic analysis tools that incorporate LLMs are not yet capable of replacing expert researchers. However, they offer significant potential as supplemental tools that can enhance the rigor, speed, and depth of qualitative analysis when used appropriately.

REFERENCES

[1] Kimberley Beaumont, Martin Oravec, Harry Emerson, Ian Penton-Voak, and Conor Houghton. Can llms be used to quantify the emotional salience of text statements using an elo rating system? In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–10, 2025.

[2] Kevin Childs, Cassidy Gibson, Anna Crowder, Kevin Warren, Carson Stillman, Elissa M Redmiles, Eakta Jain, Patrick Traynor, and Kevin RB Butler. " i had sort of a sense that i was always being watched... since i was": Examining interpersonal discomfort from continuous location-sharing applications. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 4197–4211, 2024.

[3] Victoria Clarke and Virginia Braun. Thematic analysis. In *Encyclopedia of critical psychology*, pages 1947–1952. Springer, 2014.

[4] Noah H Crampton, Shmuel Reis, and Aviv Shachak. Computers in the clinical encounter: a scoping review and thematic analysis. *Journal of the American Medical Informatics Association*, 23(3):654–665, 2016.

[5] James Eschrich and Sarah Sterman. A framework for discussing llms as tools for qualitative analysis. *arXiv preprint arXiv:2407.11198*, 2024.

[6] Madeleine Grunde-McLaughlin, Michelle S Lam, Ranjay Krishna, Daniel S Weld, and Jeffrey Heer. Designing llm chains by adapting techniques from crowdsourcing workflows. *arXiv preprint arXiv:2312.11681*, 2023.

[7] Priya C Kumar, Michael Zimmer, and Jessica Vitak. A roadmap for applying the contextual integrity framework in qualitative privacy research. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–29, 2024.

[8] L. H. Lin. Internal user test: Llama3-70b inference efficiency on h100. https://gist.github.com/bf81a9c7dfc4244c974335e1605dcf22, 2025. GitHub Gist.

[9] brewer Mollie, Kevin Childs, Jennifer Nichols, Kevin RB Butler, and Kristy Elizabeth Boyer. Student-athlete technology. In *Proceedings of the 2026 on ACM CHI Conference on Human Computer Interaction*, 2026.

[10] Muhammad Zain Raza, Jiawei Xu, Terence Lim, Lily Boddy, Carlos M. Mery, Andrew Well, and Ying Ding. Llm-ta: An llm-enhanced thematic analysis pipeline for transcripts from parents of children with congenital heart disease, 2025. Accepted at GenAI for Health Workshop, AAAI 2025, Philadelphia.

[11] Steven Schirra, Sasha G Volkov, and Frank Bentley. " it's something to polish your own thoughts, rather than create thoughts for you": Understanding participants' use of chatbots and llms during online research participation. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2025.

[12] Yan Shvartzshnaider, Noah Apthorpe, Nick Feamster, and Helen Nissenbaum. Going against the (appropriate) flow: A contextual integrity approach to privacy policy analysis. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 162–170, 2019.

[13] Tongshuang Wu, Haiyi Zhu, Maya Albayrak, Alexis Axon, Amanda Bertsch, Wenxing Deng, Ziqi Ding, Bill Guo, Sireesh Gururaja, Tzu-Sheng Kuo, et al. Llms as workers in human-computational algorithms? replicating crowdsourcing pipelines with llms. *arXiv preprint arXiv:2307.10168*, 2023.

[14] Ziang Xiao, Xingdi Yuan, Q Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding. In *Companion proceedings of the 28th international conference on intelligent user interfaces*, pages 75–78, 2023.

[15] Yuyi Yang, Charles Alba, Chenyu Wang, Xi Wang, Jami Anderson, and Ruopeng An. Gpt models can perform thematic analysis in public health studies, akin to qualitative researchers. *Journal of Social Computing*, 5(4):293–312, 2024.