Usability and User Experience, Vol. 194, 2025, 73–83
https://doi.org/10.54941/ahfe1006669

Enhancing Thematic Analysis With Local Large Language Models: A Scientific Evaluation of Prompt Engineering Techniques

Timothy Meyer, Carolyn Baker, and Jonathan Keefe

Pacific Science & Engineering Group, San Diego, CA, 92121, USA

ABSTRACT

Thematic Analysis (TA) is a powerful tool for human factors, HCI, and UX researchers to gather system usability insights from qualitative data like open-ended survey questions. However, TA is both time consuming and difficult, requiring researchers to review and compare hundreds, thousands, or even millions of pieces of text. Recently, this has driven many to explore using Large Language Models (LLMs) to support such an analysis. However, LLMs have their own processing limitations and usability challenges when implementing them reliably as part of a research process – especially when working with a large corpus of data that exceeds LLM context windows. These challenges are compounded when using locally hosted LLMs, which may be necessary to analyze sensitive and/or proprietary data. However, little human factors research has rigorously examined how various prompt engineering techniques can augment an LLM to overcome these limitations and improve usability. Accordingly, in the present paper, we investigate the impact of several prompt engineering techniques on the quality of LLM-mediated TA. Using a local LLM (Llama 3.1 8b) to ensure data privacy, we developed four LLM variants with progressively complex prompt engineering techniques and used them to extract themes from user feedback regarding the usability of a novel knowledge management system prototype. Contrary to conventional approaches to studying LLMs, which largely rely upon descriptive statistics (e.g., % improvement), we systematically applied a set of evaluation methods from behavioral science and human factors. We performed three stages of evaluation of the outputs of each LLM variant: we compared the LLM outputs to our team's original TA, we had human factors professionals ($N = 4$) rate the quality and usefulness of the outputs, and we compared the Inter-Rater Reliability (IRR) of other human factors professionals ($N = 2$) attempting to code the original data with the outputs generated by each variant. Results demonstrate that even small, locally deployed LLMs can produce high-quality TA when guided by appropriate prompts. While the "baseline" variant performed surprisingly well for small datasets, we found that the other, scalable methods were dependent upon advanced prompt engineering techniques to be successful. Only our novel "cognition-inspired" approach performed as well as the "baseline" variant in qualitative and quantitative comparisons of ratings and coding IRR. This research provides practical guidance for human factors researchers looking to integrate LLMs into their qualitative analysis workflows, disentangling and uncovering the importance of context window limitations, batch processing strategies, and advanced prompt engineering techniques. The findings suggest that local LLMs can serve as valuable and scalable tools in thematic analysis.

Keywords: Usability, Thematic analysis, Large language models, Prompt engineering

INTRODUCTION

Human Factors professionals are commonly tasked with improving the usability of a system. Understanding how to enhance a system or product's usability often requires analyzing qualitative user feedback to identify recurring patterns, uncover pain points, and gain meaningful insights into user interactions and experiences. Thematic Analysis (TA) is a widely used method for uncovering and analyzing such themes in qualitative data (McDonald et al., 2019). While TA is a versatile and effective tool for discovering insights, it can be time- and labor-intensive, especially when dealing with large or detailed datasets.

Advancements in Natural Language Processing (NLP), and most recently, Large Language Models (LLM) have sparked growing interest among researchers seeking to leverage these technologies for more efficient qualitative analysis. While LLMs have shown strong performance in various NLP tasks, their effectiveness in research applications is constrained by several factors. LLMs have finite context windows for how much information one can parse at once, which limits how well an LLM can describe a dataset that is larger than its context window. LLMs are stochastic in nature and sometime produce errant responses, which is compounded by the fact that those errant responses can be quite convincing and hard to detect. Additionally, to exert more control over the LLM and ensure data privacy, users must sometimes use locally hosted LLMs, which are typically much less powerful and do not have convenient methods for interacting with the model. Altogether, these limitations in LLMs create a set of human factors considerations for using LLMs for thematic analysis. Users must closely monitor the information being input into the model, meticulously screen its output for its consistency and veracity, and learn how to effectively communicate with the model – often through trial and error.

Prompt engineering techniques can be leveraged to provide a solution to these human factors issues for optimizing interactions with LLMs by designing clear, structured inputs that enhance usability, reliability, scalability and effectiveness of the model's responses (Brown et al., 2020; Chen, Zhang, Langrené, & Zhu 2023). In the context of TA, well-crafted prompts help ensure that LLMs produce meaningful, structured, and interpretable results that align with research objectives (Mathis et al., 2024). However, most users possess only a basic understanding of how LLMs function and what influences the quality of their outputs. This knowledge gap underscores the need for systematic evaluation of different prompting techniques to establish evidence-based best practices for research applications.

Several attempts have been made to integrate LLMs into processes of TA (Dai, Xiong & Lu, 2023; De Paoli, 2024; Douglas, 2023; Morgan, 2023; Torii, Murakami & Ochiai, 2024). While valuable these attempts come with a few caveats. First, most of these attempts use LLMs that cannot be run locally (e.g., ChatGPT 3.5). This makes their applications not transferable to anyone dealing with proprietary or further secured data that can't be passed over an internet connection to a third party. Additionally, we know of no other attempts to quantitatively measure how using techniques to create more effective prompts (i.e., prompt engineering) impacts the accuracy, quality, and effectiveness of LLM-generated TA of qualitative user feedback data.

Therefore, in this paper we investigate the impact of several prompt engineering techniques on the quality of LLM-mediated TA of qualitative data regarding the usability of a novel knowledge management system prototype. We created 4 prompt variants with differing prompt engineering and performed a systematic evaluation based on behavioral science and human factors principles. We performed three phases of evaluation of the outputs of each variant. First, we compared the variant outputs to our team's original TA; then, we asked four HFEs to rate the quality and usefulness of the outputs; and finally, we compared the Inter-Rater Reliability (IRR) of two more HFEs attempting to code the original data with the outputs generated by each variant. The confluence of these analyses provides a deeper understanding of how LLMs can support researchers using TA and provides best practices for integrating LLMs in an effective way. We also provide our prompts and scripts so that these can be used to guide researchers in creating their own prompts for local models (found at: https://github.com/Pacific-Science-Engineering/PE_for_LLM_TA).

METHODS AND RESULTS

Study Overview

To study the impact of advanced prompt engineering techniques on the quality of LLM-mediated TA, we developed four LLM prompt variants with progressively complex prompt engineering techniques (explained in depth below) and used them to extract themes from user feedback. We used data from an evaluation of an alpha build of a novel Knowledge Management software, which included data our team collected to evaluate the usability of the system (Eckroth et al., 2025). This included user feedback from 44 participants who used the system to complete a series of tasks in the oil and well-drilling domain and then voluntarily provided feedback on the system at the end regarding their experience. Each of these participants responded to an open-ended prompt asking for additional feedback on the system after completing their tasks and surveys.

LLM Prompt Variants

We developed four LLM prompt variants by progressively increasing the complexity of prompt engineering techniques. Each variant used a different combination of prompt engineering techniques used to direct the analysis of the LLM and scripting techniques used to control the way data was fed into the LLM. All prompts were processed using a local instance of Llama 3.1 8b. The prompts and accompanying Python scripts are available online (see above).

The "baseline" prompt variant simply asked the LLM to identify themes in the user feedback, analyzing all the user feedback at once without additional guidance or structure (i.e., without any advanced prompt engineering). While our dataset was relatively small and fit within the context window, this method is generally considered unscalable because in many cases the amount of text to be processed will exceed the length of the context window of a local model.

The second variant we developed is referred to as the “naïve batch processing” variant. This variant uses only the technique of batch processing with LLMs, which refers to the practice of using a script to divide large datasets into smaller segments that are processed sequentially to overcome context window limitations and improve scalability and efficiency in tasks such as thematic analysis (see Ou & Lapata, 2025 for a more thorough review of batch processing for LLMs). By avoiding context window limitations and iteratively analyzing data, we hypothesized that this variant could meaningfully improve output quality over the baseline.

The third variant we developed is referred to as the “advanced batch processing” variant. This prompt combined batch processing with advanced prompt engineering techniques, including role-based prompting, chain of thought (CoT) prompting, and self-consistency prompting (see Chen, Zhang, Langrené, & Zhu, 2023 for a thorough review of prompt engineering techniques). Role-based prompting instructs the LLM to adopt a specific role (e.g., UX researcher), which helps refine its output by aligning responses with the expected expertise, tone, and context of that role. CoT prompting guides the LLM to reason sequentially by detailing the series of intermediate steps taken to produce the answer, which has been shown to lead to more complete and accurate outputs (Wei et al., 2022). Finally, self-consistency prompting asks the LLM to make several responses for a given prompt and choose the one that is most consistent with the others, improving accuracy of responses by mitigating the effects of incorrect or random reasoning paths by selecting the majority answer, similar to resampling many times and taking the mean (Wang et al., 2022). We hypothesized that the incorporation of these advanced techniques would significantly improve TA output of this variant over the “naïve batch processing” variant.

The final variant we developed introduced a novel approach by prompting the LLM to create and maintain a living set of notes that it would update after reviewing each piece of user feedback, which we refer to as the “Cognition-inspired” model. Mirroring how humans perform TA, we scripted this model to consider each piece of feedback individually and keep “rolling notes” of the themes encountered in feedback to progressively extract a full set of themes. Though this variant applied the same advanced prompt engineering techniques used in the “advanced batch processing” variant, we hypothesized that this variant could yield improvements given that its process more closely follows that of a trained human analyst.

Phase 1: Qualitative Comparison to Team of HF Professionals

In Phase 1, we conducted a qualitative comparison to evaluate the agreement between the themes generated by the original team of HFEs (i.e., the authors on this paper) and those generated by each prompt variant. Our TA identified six themes from user feedback. Three of the themes had to do with specific software features, which included ‘Filters’, ‘Search’, and ‘Insights’ (a feature that provided tips to improve task performance). The fourth theme centered on the system's 'Evaluation', where users interacted with the software to complete various tasks. The final two themes were broader and included 'General' feedback unrelated to specific software features and 'User Recommendations' that included suggested system improvements.

All variants showed some alignment with the original themes identified. The "Baseline" variant identified several overlapping categories by capturing users' frustration with search functionality, and correctly identified mixed opinions on the filters and insights features. Although this variant did not explicitly recognize the role of the evaluation, likely because the evaluation was never explicitly mentioned in the user feedback, its fourth theme did highlight a cluster of issues stemming from it.

Several themes from the "Naïve batch processing" variant also overlapped with the original TA, including users' varying opinions of the Insights feature. Additionally, this variant was able to draw out the impact of the testing environment as a theme. However, the themes from this variant were somewhat vague, lacking descriptive clarity in comparison to the original themes.

Thematic alignment improved significantly with the "Advanced batch processing" variant, particularly in its identification of topics related to the search and filtering features. However, unlike in the original TA, it did not identify insights as a theme. This variant also identified several new areas of interest that were not initially identified by the original researchers such as 'System Design and Customization' and 'Database Navigation and Complexity'.

The "Cognition-inspired" variant produced the most comprehensive set of themes, covering all the original themes and adding more detailed ones. This set included themes for filtering, insights, and navigation and usability, which contained feedback originally coded under the evaluation theme. This variant also introduced some new themes not captured in the original TA; however, many of these could be merged into a single 'search' theme, making the overall theme sets quite similar.

Overall, our qualitative assessment of each variants' output revealed that there were clear benefits to incorporating scripted batch processing into prompt engineering for TA, but that these benefits were only noteworthy when paired with advanced prompt engineering techniques.

Phase 2: Quantitative Assessment of LLM Output Quality

In Phase 2, we looked to collect independent evaluations of each prompt variant's output and quantify the impact of each prompt engineering approach as a complement to our qualitative analysis. To this end, we assessed the quality of each variant's TA output by recruiting four HFEs to provide quantitative ratings of quality using a survey designed to evaluate the accuracy, completeness, and usefulness of each variant's output.

Raters were provided with a document containing all the user feedback, followed by the four sets of themes generated by the LLM variants. After reviewing each set, participants responded to seven questions using a 5-point Likert scale, ranging from strongly disagree to strongly agree:

1. The identified themes are clear and easy to understand.
2. The identified themes are presented concisely without losing important information.
3. The thematic analysis is comparable to what I would expect from a human researcher.
4. The identified themes completely describe all pieces of user feedback.
5. The identified themes accurately reflect the main ideas in the user feedback.
6. The identified themes provide actionable insights for improving the system.
7. I would feel confident using the themes for further analysis or decision-making.

The average score across all questions and participants, plotted in Figure 1A, ranges from 1 (entirely negative) to 5 (entirely positive). The average scores for the "Advanced batch processing" (M = 3.9) and "Cognition-inspired" (M = 3.9) variants were identical and similar to the "Baseline" variant (M = 4.1), and demonstrably higher than the "Naïve batch process" variant (M = 1.14). Therefore, as in our qualitative analysis, the quantitative analysis found the outputs of the variants with advanced prompt engineering techniques to be of much higher quality than the "Naïve batch processing" variant without them.

To further examine the relationship between the variant type and participants' ratings, we sought a purely quantitative analysis to more objectively measure the rated quality of each variant's output. Accordingly, we conducted an ordered logistic regression. In the regression model, represented by the equation below, the outcome variable (y) is defined as the participant-provided rating for *k* ordinal values. The predictor variables include the LLM variant (v) and the specific survey question (*q)* to control for inter-question variability.

$$\log \frac{\Pr(y_i \ \leq k)}{1 - \Pr(y_i \ \leq k)} = \alpha_k - \phi_i$$

$$\phi_i = \beta v_i + \beta q_i$$

The model parameters were estimated using Bayesian methods implemented via the rethinking package (McElreath, 2020) to estimate the phi parameter, which represents participants' response probabilities relative to the variant type and question. In this analysis, higher phi values indicate higher ratings on the survey questions being associated with the outputs of a variant. The analysis results showed that that phi is roughly equivalent and positive for the "Baseline" ($\beta = 1.07$, 95% confidence interval [CI]: 0.36 to 1.81), "Advanced batch processing" ($\beta = 0.30$, 95% CI: −0.02 to 0.61), and "Cognition-inspired" variants ($\beta = 0.21$, 95% CI: −0.02 to 0.61). The positively skewed distributions of ratings for these variants indicated favorable participant responses. The "Naïve batch processing" variant had a distinctly negative β estimate of -1.75 (95% CI: -2.31 to -1.23) that was far removed from the other variants, clearly indicating its lower performance. This distinction is shown in Figure 1B, where posterior distributions illustrate the notably lower rating for Variant 2 compared to others.

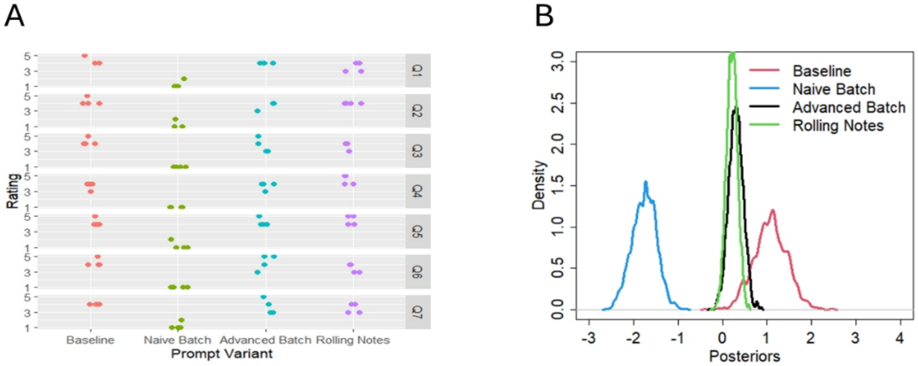

Figure 1: (A) shows subject ratings for each question and variant, ranging from 1 (strongly disagree) to 5 (strongly agree). (B) shows the posterior distribution for each βv_i.

We also asked the raters to rank the four variants by their perceived usefulness as a complementary, straightforward measure of output quality. In line with the results above, all four participants ranked "Naïve batch processing" as the least useful variant. The "Baseline" and "Cognition-inspired" variants each received two rankings as most useful, one as second-most useful, and one ranking as third-most useful. The "Advanced batch processing" variant was consistently ranked in the middle, with two rankings as second most useful and two as third-most useful. These rankings align with earlier qualitative and quantitative findings showing that the "Baseline" and "Cognition-inspired" variants were more effective than the batch processing variants.

Phase 3: Coding and Inter-Rater Reliability Analysis

In Phase 3, we looked to collect a fine-grained and ecologically valid analysis of the usefulness of each variant's output. To this end, we recruited two HFEs to use the sets of themes generated by each LLM to code the user feedback. These professionals were given 6 documents: a master document containing all user feedback presented in a numbered list, four separate theme documents each containing one of the sets of themes extracted by the variants, and a corresponding coding table with one row per user feedback entry. Researchers marked which themes they thought were appropriate for each piece of user feedback, allowing for an assessment of coding consistency across raters. We reasoned that if the themes extracted from the feedback are useful, then we should see that independent coders can use them consistently to code individual feedback. This analysis takes us beyond reasoning about how good output seems and provides an objective measure of how good the TA output is in practice.

We compared IRR between the two raters for each set of themes using an adjusted Krippendorff's alpha, which we developed to accommodate the non-mutually exclusive nature of the themes found in feedback. The code and a summary of the technique are available at the linked GitHub. We found that adjusted Krippendorff's alphas were relatively similar across the "Baseline" ($\alpha = 0.67$), "Naïve batch processing" ($\alpha = 0.67$), and "Cognition-inspired" ($\alpha = 0.71$) variants; however, the "Advanced batch processing" variant had a lower alpha of 0.58. This difference between the "Advanced batch processing" variant and the others isn't large, but it may still reflect meaningful variability in agreement among raters and clearly suggests that the raters had less agreement regarding its usefulness. This aligns with results from Phase 2 indicating that the "Advanced batch processing" variant, despite generally positive ratings, consistently ranked below the "Baseline" and "Cognition-inspired" variants in its usefulness.

DISCUSSION

Consistent with recent studies (Dai et al., 2023; De Paoli, 2024; Douglas, 2023; Morgan, 2023; Torii et al., 2024; Wittmann, 2024) our findings provide evidence for the beneficial role that LLMs can play in qualitative analysis and highlight how and when prompt engineering can further enhance their performance and scalability. In Phase 1, we found that the "Baseline", "Advanced batch processing", and "Cognition-inspired" prompt variants produced theme sets that were qualitatively similar to those generated during our initial TA, whereas the "Naïve batch processing" variant showed less alignment with the original themes. Similar results were found in Phase 2 when comparing the perceived quality across variants. The "Baseline", "Advanced batch processing", and "Cognition-inspired" prompt variants patterned together and were rated higher in perceived accuracy and usefulness compared to the "Naïve batch processing" variant. Ranking data helped to slightly differentiate the top three variants by showing a split preference for both "Baseline" and "Cognition-inspired" variants over "Advanced batch processing." Lastly, results from the IRR analysis showed that raters most consistently applied themes from the "Baseline", "Naïve batch processing", and "Cognition-inspired" variants when coding individual feedback.

Taken together, these findings suggest that the "Baseline" and "Cognition-inspired" variants consistently produced the highest-quality and most reliable results. While it may be surprising that the "Baseline" model performed on par with the more advanced "Cognition-inspired" variant, the explanation lies in the scale of the data used in this evaluation. Simpler prompts can be effective when the data fits within the context window, as was the case with our relatively small dataset, but larger datasets require scalable solutions. However, the poorer performance of the "Naïve batch processing" variant across evaluations suggests that only addressing scalability using scripting techniques such as batch processing is not sufficient. To improve effectiveness, batch processing should be paired with advanced prompt engineering – as evidenced by the higher quality of the "Advanced batch processing" (vs. Naïve) variant's output. Thus, we advocate for researchers to incorporate CoT, self-consistency, and role-based prompting at a minimum into their LLM prompts to support effective TA.

In reviewing the full set of results presented in the current paper, it becomes clear that the "Cognition-inspired" variant offered the best balance of scalability, usefulness, and usability of the evaluated LLM prompt variants. Its "rolling notes" approach enabled the LLM to track the frequency of each theme's appearance and iteratively refine the full list of identified themes. Our evaluation indicated that this approach resulted in clearer, more useful themes extracted from the data when applied concurrently with advanced prompt engineering techniques. Therefore, we can conclude that forcing the LLM to describe its evidence, as well as its reasoning, is an additionally useful type of prompt engineering to deploy in support of TA.

Additionally, and perhaps more importantly, the more transparent and tractable nature of the approach embodied by the "Cognition-inspired" variant is clearly more conducive to an effective human-AI pairing. Many papers have collectively emerged to not only reveal the benefits of pairing LLMs and humans together to perform TA, but also to warn against using LLMs to perform TA on their own (Khan et al., 2024). Indeed, we see clear evidence of LLMs' limitations in our qualitative analysis of each LLM variant's output, where none of the variants explicitly identified a category for the feedback that was related to the evaluation. This theme represents a grouping that human researchers were able to infer only because of the implicit information they already knew about the context of the collected feedback. Indeed, understanding the context of the data will always be paramount for performing TA, and therefore we advocate for a "human-in-the-loop" approach to applying LLMs for thematic analysis so that a researcher can provide and interpret implicit contextual information for the LLM.

Our work highlights another key caveat for using LLMs for TA. Our 'Baseline' results exemplify the double-edge sword of using LLMs: they easily produce quality outputs but can create a false sense of security when initial test cases appear acceptable at face value. As a result, users must closely inspect LLM output for its quality. Here, the value of the "Cognition-inspired" variant's explicit logging of evidence becomes even clearer, as it fits nicely with a human-in-the-loop system for LLM supervision. Instead of simply producing a set of themes that the researcher would have to re-evaluate or compare to their own TA, the "Cognition-inspired" variant makes that comparison accessible by putting the information the researcher must review among the notes for easy supervision.

However, some limitations of using LLMs cannot be overcome with prompt engineering and must simply be kept in mind as users interpret LLM output. For example, Braun and Clarke (2006) note that researchers make implicit decisions during thematic analysis; and we argue that LLMs make many of these decisions automatically without explicit disclosure. For example, using an LLM inherently forces a deductive analysis approach rather than an inductive one (Alhojailan, 2012) because LLMs have already been trained on relevant literature. Furthermore, since conventional theories dominate LLM training data, LLMs provide a maximally conventional theoretical approach rather than an atheoretical one (Eschrich & Sterman, 2024). Future work should attempt to identify and quantify the impact of these types of biases and implicit decisions and perhaps seek prompt engineering solutions to these issues to the extent it is possible.

In conclusion, this work demonstrates that LLMs can be a valuable tool in thematic analysis when appropriately engineered and integrated with humans. Prompt engineering techniques can help to maximize the quality, usefulness, and scalability of LLM-mediated thematic analysis; however, LLMs must be effectively paired with researchers to ensure the accuracy and interpretability of the TA output. In this way, LLMs can augment and extend the capabilities of researchers, allowing them to focus their expertise where it matters most—on understanding context, making nuanced interpretations, and generating meaningful insights that improve user experiences.

ACKNOWLEDGMENT

The work reported in this paper was supported by U.S. Army Research Office (ARO) and Defense Advanced Research Projects Agency (DARPA), Contract No. W911NF22C0048. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. The authors also thank colleagues who contributed to and supported this work including Karen Myers, Michael Wessel, Manuela Jaramillo, & Harvey Smallman.

REFERENCES

Alhojailan, M. I. (2012, October). Thematic analysis: A critical review of its process and evaluation. In *WEI international European academic conference proceedings, Zagreb, Croatia*.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C.,... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). Unleashing the potential of prompt engineering in large language models: A comprehensive review. *arXiv preprint arXiv:2310.14735*.

Dai, S. C., Xiong, A., & Ku, L. W. (2023). LLM-in-the-loop: Leveraging large language model for thematic analysis. arXiv preprint arXiv:2310.15100.

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.

Douglas, M. R. (2023). Large language models. *arXiv preprint arXiv:2307.05782*.

Eckroth, J., Freitag, D., Keefe, J., Meyer, T., Myers, K., Schoen, E., Sequeira, P., & Smith, R. (2025). *ATHENA: A Virtual Member of a Community of Practice*. [manuscript submitted for publication].

Eschrich, J., & Sterman, S. (2024). A framework for discussing llms as tools for qualitative analysis. *arXiv preprint arXiv:2407.11198*. Joffe, H. (2011). Thematic analysis. Qualitative research methods in mental health and psychotherapy: A guide for students and practitioners, 209–223.

Jones, K. S. (1994). Natural language processing: a historical review. *Current issues in computational linguistics: In honour of Don Walker*, 3–16.

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

Mathis, W. S., Zhao, S., Pratt, N., Weleff, J., & 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.

McDonald, N., Schoenebeck, S., & Forte, A. (2019). Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–23.

McElreath, R. (2020b). *Rethinking R package*.

Morgan, D. L. (2023). Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. International Journal of Qualitative Methods, 22, 16094069231211248.

Torii, M. G., Murakami, T., & Ochiai, Y. (2024). Expanding Horizons in HCI Research Through LLM-Driven Qualitative Analysis. arXiv preprint arXiv:2401.04138.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837.

Wittmann, F. H. (2024). Enhancing Thematic Analysis with Large Language Models: A Comparative Study of Structured Prompting Techniques.