Evaluation and Best Practices of AI-Assisted Thematic Analysis

TEDDY WRIGHT*, Kahlert School of Computing, University of Utah, USA
NASTARAN JADIDI, Kahlert School of Computing, University of Utah, USA
SARA YEO, Department of Communication, University of Utah, USA
VINEET PANDEY, Kahlert School of Computing, University of Utah, USA

AI brings the possibility of performing qualitative research in fractions of the time at quality similar to or better than human researchers. Testing Chat GPT 5.5 and Opus 4.8's ability to construct themes from public comments submitted to a Food and Drug Administration (FDA) policy document, we find that AI's fit for qualitative analysis is largely a function of its configuration. Our results show that output quality increases with a thoughtfully engineered prompt and a more detailed research question. We also find that different research questions and positionality profiles surface different theme sets and that output quality holds. This work informs new hybrid human-AI workflows in qualitative research.

1 Introduction

Qualitative research underpins vast amounts of insight and action across a variety of contexts [6]. It shapes decisions on topics such as what changes to make in workplaces or how medical treatments might be improved [2]. However, a primary challenge of qualitative analysis is that it can be very time consuming to do well: one study reported 126 person-hours to analyze 22 hour-long transcriptions and 8 service documents [10]. AI brings the possibility of performing this qualitative research in fractions of the time at similar or better quality [3].

Currently, the viability of AI in assisting with qualitative research is unclear. Some studies find LLM output competitive with expert work [6, 11]. Others find that AI does not produce quality enough outputs to be useful or hallucinates outputs when assisting with qualitative research [4, 7]. Current research on AI as a tool for qualitative research has primarily relied on testing single or few configurations of the tool, which does not generalize to all AI use or sufficiently inform how to improve output. In practice, model output shifts substantially with prompt design, question framing, and the perspective that the AI is asked to inhabit [5, 8, 9]. This reframes the question from whether AI can perform qualitative analysis to which of its configurations fit specific analytical needs. To answer this question, we test top AI models with repeated runs across systematically varied configurations of prompt design, research questions (RQs), and positionalities (the values, knowledge, and external context a researcher brings), and we evaluate output with a standardized rubric.

2 Methods

We test ChatGPT 5.5 (High thinking mode) and Claude Opus 4.8 (Max thinking mode) on constructing themes in the context of thematic analysis, a specific qualitative research method [1]. We do this on comments submitted by the Amyotrophic Lateral Sclerosis (ALS) community in response to a draft industry guidance document issued by the Food and Drug Administration (FDA) in 2018. We test on the 577 unique comments out of the 611 submitted over the two month commenting period.[1]

We grade each AI theme on rubric fields on a scale from 1-5. The rubric is based on literature identifying what valid thematic analysis looks like [1]. Four of these fields are: grounding (is the theme supported by the data), RQ fit (is the theme relevant to the RQ), interpretation level (how deep/non obvious vs. surface the theme is), novelty (how

*Undergraduate author.
[1]Full prompt text, all RQs, and all positionalities tested are available in supplemental materials.

unpredictable is the theme). Higher grounding and RQ fit directly indicate output quality. Higher interpretation level and novelty can indicate higher output quality when it is expected for an RQ, but lower scores on these fields do not always mean lower quality.

We first test 6 prompt levers, informed by prompting best practices [8] and our own hypotheses, on a subset of 20 comments to identify which changes yield higher quality output, then use these findings to build a refined, engineered prompt. We then test across two RQs (RQ1: What are the experiences of the ALS community regarding their relationship with regulatory processes and institutions around treatment?; RQ2: What kinds, structures, and topologies of criticisms and recommendations do people make for policy?) and three positionalities (e.g., AI default, an FDA regulatory analyst, an ALS community advocate) on the full 577 comment corpus. In total, we rated 551 individual AI-generated themes on the rubric across our runs.

3 Results and Discussion

Two of the six prompt levers tested on the 20-comment subset improved output. A more detailed RQ (3-4 sentences instead of 1) produced the largest improvement, raising interpretation level and novelty while grounding held (Table 1). Using mathematical language for more precise instruction (e.g., orthogonal instead of independent themes) moderately improved novelty for both models. Richer background context, explicit theme organization instruction, and having the AI predict what a human expert would produce rather than inhabiting that expert each had negligible effects.

Across the full comment corpus, AI produced quality themes across different RQs and positionalities with average grounding per theme set (all runs contained 7-10 themes per set) staying within 4.78-5.00 and RQ fit within 3.4-4.7. The two research questions produced themes with different content, while positionality mainly affected how the data got sliced and organized and the prose and framing used, with Claude showing a larger positionality response than ChatGPT. Our engineered prompt averaged 3.03 on interpretation level compared to 2.14 using a non-engineered prompt, with the non-engineered output reading more like clustered summaries than analysis.

Table 1. Average scores per theme set, one single sentence research question compared theme set averages on the left compared to the mean of 2 more detailed research question theme set averages on the right (delta in parentheses). All scores out of 5.

	ChatGPT 5.5	Claude Opus 4.8
Grounding	4.67 → 4.84 (+0.17)	5.00 → 5.00 (0.00)
Research question fit	4.33 → 3.95 (−0.39)	4.20 → 4.70 (+0.50)
Interpretation level	2.00 → 2.89 (+0.89)	2.30 → 3.54 (+1.24)
Novelty	2.67 → 2.95 (+0.28)	3.10 → 3.54 (+0.44)

This work informs new hybrid human-AI workflows for qualitative research. We show that specific levers (an engineered prompt, a more detailed RQ, different RQs, and different positionalities) can tailor AI's output toward a given analytical need. This means researchers can generate themes across many questions and perspectives, and over far larger datasets in a fraction of the time, with the analyst's work shifting toward designing the AI configuration and evaluating and synthesizing AI output. It also means that reporting of AI's configuration is needed for interpreting results, which is important both for qualitative methods that prioritize reliability across analysts or methods that expect different interpretations of data, such as thematic analysis. One outcome of these workflow adjustments would be allowing institutions that receive thousands of public submissions, like the FDA, to thoroughly analyze every submission they receive — giving more public input a path into shaping policy decisions.

References

[1] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. doi:10.1191/1478088706qp063oa

[2] Virginia Braun and Victoria Clarke. 2024. A Critical Review of the Reporting of Reflexive Thematic Analysis in Health Promotion International. *Health Promotion International* 39, 3 (2024), daae049. doi:10.1093/heapro/daae049

[3] Stefano De Paoli. 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 (2024), 997–1019. doi:10.1177/08944393231220483

[4] Tanisha Jowsey, Peta Stapleton, Shawna Campbell, Alexandra Davidson, Cher McGillivray, Isabella Maugeri, Megan Lee, and Justin Keogh. 2025. Frankenstein, thematic analysis and generative artificial intelligence: Quality appraisal methods and considerations for qualitative research. *PLOS One* 20, 9 (Sept. 2025), e0330217. doi:10.1371/journal.pone.0330217

[5] Grgur Kovăč, Masahiro Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large Language Models as Superpositions of Cultural Perspectives. arXiv:2307.07870 doi:10.48550/arXiv.2307.07870

[6] Cristina Martinez Montes, Robert Feldt, Cristina Miguel Martos, Sofia Ouhbi, Shweta Premanandan, and Daniel Graziotin. 2025. Large Language Models in Thematic Analysis: Prompt Engineering, Evaluation, and Guidelines for Qualitative Software Engineering Research. arXiv:2510.18456 [cs.SE]. doi:10.48550/arXiv.2510.18456

[7] Duc Cuong Nguyen and Catherine Welch. 2026. Generative Artificial Intelligence in Qualitative Data Analysis: Analyzing—Or Just Chatting? *Organizational Research Methods* 29, 1 (Jan. 2026), 3–39. doi:10.1177/10944281251377154

[8] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=RIu5lyNXjT

[9] Siddharth Suresh, Kushin Mukherjee, Xizhao Yu, Wei-Chun Huang, Lisa Padua, and Timothy Rogers. 2023. Conceptual structure coheres in human cognition but not in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 722–738. doi:10.18653/v1/2023.emnlp-main.47

[10] Beck Taylor, Catherine Henshall, Sara Kenyon, Ian Litchfield, and Sheila Greenfield. 2018. Can rapid approaches to qualitative analysis deliver timely, valid findings to clinical leaders? A mixed methods study comparing rapid and thematic analysis. *BMJ Open* 8, 10 (Oct. 2018), e019993. doi:10.1136/bmjopen-2017-019993

[11] Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In *28th International Conference on Intelligent User Interfaces*. ACM, Sydney NSW Australia, 75–78. doi:10.1145/3581754.3584136