Evaluation and Best Practices of AI-Assisted Thematic Analysis

Teddy Wright[1*] · **Nastaran Jadidi**[1] · **Sara Yeo**[2] · **Vineet Pandey**[1]

[1] Kahlert School of Computing, University of Utah [2] Department of Communication, University of Utah * Undergraduate author

This work was supported by SPUR from the Office of Undergraduate Research at the University of Utah awarded to Teddy Wright.

Problem

A primary challenge of qualitative analysis is that it can be very time consuming to do well. AI brings the possibility of performing this research in fractions of the time at similar or better quality [2]. Underlying the literature's mixed results is an assumption that AI can largely be characterized as a fixed instrument whose viability can be settled by testing one configuration of it [2, 3, 4]. In practice, the question is not whether AI can perform qualitative analysis, but which configurations of it fit which analytical needs.

Methods

We use comments submitted by the Amyotrophic Lateral Sclerosis (ALS) community on a draft industry guidance document issued by the Food and Drug Administration (FDA) in 2018 — 611 comments totaling around 200 pages; we test on the 577 unique comments.

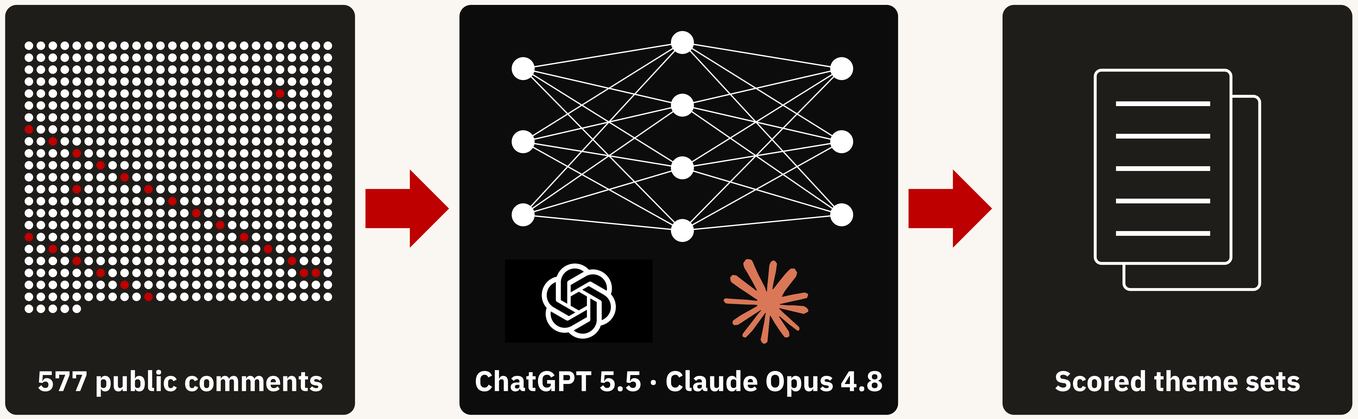
We test ChatGPT 5.5 and Claude Opus 4.8 on their highest thinking modes on constructing themes via thematic analysis [1]. We grade each AI theme on rubric fields on a scale from 1-5, based on literature identifying what valid thematic analysis looks like [1, 4].

We first test 6 prompt levers on a subset of 20 comments, then use these findings to build a refined, engineered prompt. We then test across varied conditions of research questions (RQs) and positionalities on the full 577 comment corpus.

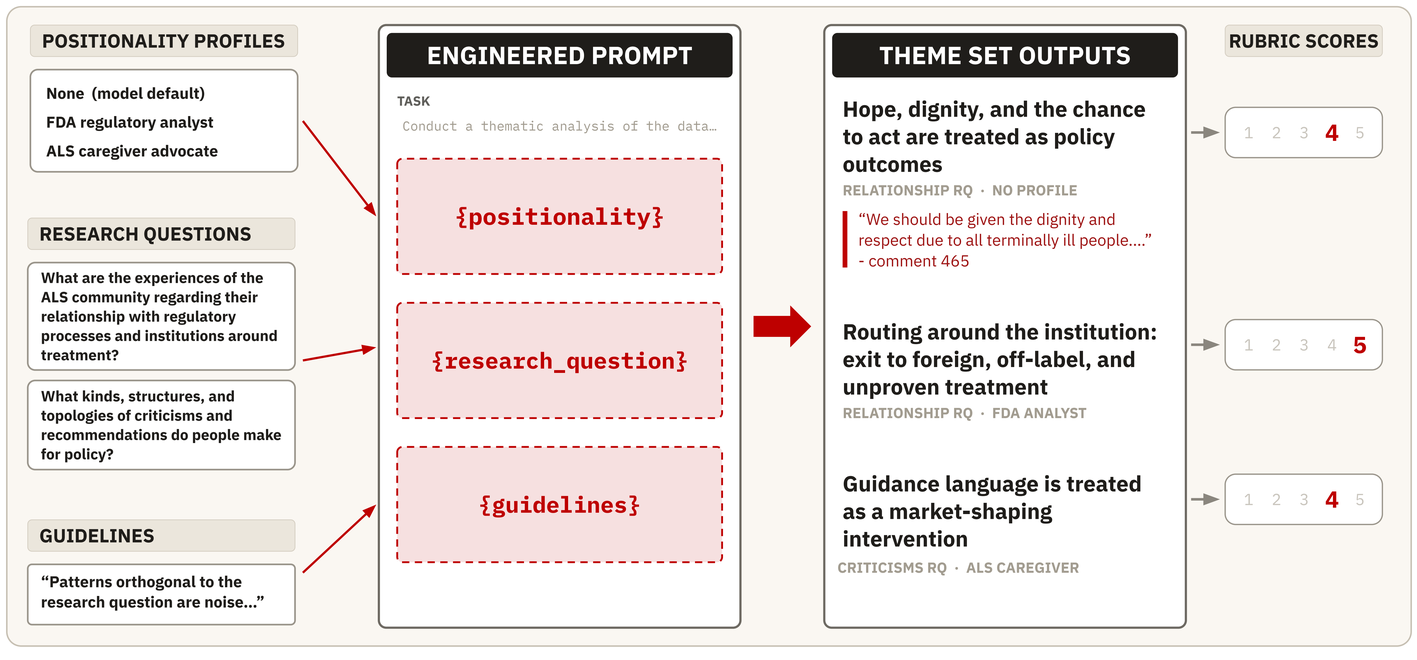
PROMPT LEVERS TESTED

Elaborating the research question (3–4 sentences vs 1)	▲ Large gain
Mathematical phrasing of the instructions	▲ Moderate gain

● No change — extra independence instruction, directing salience vs RQ relevance, richer background on ALS and the FDA, predicting a person vs being the person



Qualitative research done with AI depends on AI's configuration



STANDARDIZED RUBRIC

Grounding	Research Question Fit	Interpretation Level	Novelty	Data Contribution	Positionality Contribution
1 2 3 4 **5**	1 2 3 **4** 5	1 2 3 **4** 5	1 2 **3** 4 5	1 2 3 **4** 5	1 2 **3** 4 5
How supported by the data	Fits the depth and specificity the RQ demands	How far beyond surface text	How unexpected the insight is	How different theme is from no data baseline	How different theme is from no positionality baseline

Results

551
AI-generated themes rated on the rubric

+0.9
Average interpretation level gain from a less to more engineered prompt

4.8–5 out of 5
Average grounding per theme set

KEY FINDINGS

Engineering the prompt matters	Without it themes stayed grounded but read like a clustered summary with little interpretation; most negative results in the literature use less engineered prompts
Quality held across every configuration	Grounding 4.8–5, RQ fit 4–4.9 (2 outliers at 3.4)
The research question is the bigger lever	Positionality mostly sliced, framed, and organized the data differently rather than adding new content
A second pass adds novelty	More novel themes, often more interpretive, with higher data contribution
No-data themes scored higher on RQ fit	Models overfit to the comments' wording, pulling output away from RQ relevant language

Conclusion

This work informs new hybrid human-AI workflows where researchers can generate themes across many questions and perspectives, with the analyst's work shifting toward designing the AI configuration and evaluating and synthesizing AI output. One outcome of this is allowing institutions like the FDA to thoroughly analyze every submission they receive — giving more public input a path into shaping policy decisions.

References

[1] Braun, V. and Clarke, V. 2006. Using thematic analysis in psychology. Qualitative Research in Psychology 3, 2.
[2] De Paoli, S. 2024. Performing an Inductive Thematic Analysis of Semi-Structured Interviews with a Large Language Model. Social Science Computer Review 42, 4.
[3] Jowsey, T. et al. 2025. Frankenstein, thematic analysis and generative artificial intelligence. PLOS One 20, 9.
[4] Martinez Montes, C. et al. 2025. Large Language Models in Thematic Analysis: Prompt Engineering, Evaluation, and Guidelines. arXiv:2510.18456.

AFFILIATIONS

Teddy Wright | teddy.wright@utah.edu | linkedin.com/in/teddywright

Office of Undergraduate Studies
UNDERGRADUATE RESEARCH
THE UNIVERSITY OF UTAH