LLM-powered Workflow to Support Qualitative Data Analysis: Lessons Learned from an Applied Research Project

Clara-Maria Barth[1] and Anton Fedosov[2]

[1] Institut für Informatik - University of Zurich, Switzerland
cbarth@ifi.uzh.ch
[2] Institut für Interaktive Technologien - Fachhochschule Nordwestschweiz, Switzerland
anton.fedosov@fhnw.ch

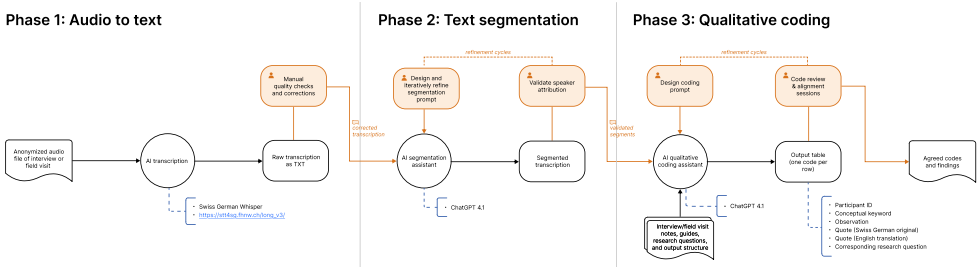


Fig. 1. LLM-based Workflow for Qualitative Data Analysis of Interviews and Field Visits

Abstract. Large Language Models (LLMs) can optimize data analysis in time-consuming industry projects. We present an AI-powered workflow that we developed and adopted in an applied research project on active and assisted living (AAL) to facilitate qualitative data analysis of semi-structured interviews. We report the challenges encountered and reflect on the trade-offs qualitative researchers should consider when using such tools in practice.

Keywords: Qualitative Text Analysis, LLM, Qualitative Coding, Interviews, Qualitative Data, Swiss German

1 Introduction

Prior work has surveyed (e.g., [1]) and tested LLMs for qualitative analysis, including inductive thematic analysis [2], multi-agent workflows [3], and codebook development [4]. We contribute a novel workflow that combines specialized and off-the-shelf LLM-powered tools, namely STT4SG[3] and ChatGPT, tailored to the linguistic context of German-speaking Switzerland, and we reflect on its limitations. The workflow [5] was developed in a project on active and assisted living (AAL) through semi-structured interviews lasting 30 to 90 minutes with older adults ($N = 8$) and their caregivers (formal: $N = 7$, informal: $N = 4$), complemented by two 3.5-hour visits to retirement homes.

[3] https://stt4sg.fhnw.ch/long_v3/

2 Workflow

We propose a three-phase linear-iterative workflow in which each phase's output serves as input to the next (Figure 1).

1. **Audio-to-Text Transcription** Audio recordings were transcribed using STT4SG, an FHNW speech-to-text system adapted for Swiss German dialect variations and based on OpenAI Whisper. Human quality checks helped avoid propagating observed model limitations, including erroneous transcriptions in dialect-heavy passages, long sentences, and overlapping speech.
2. **AI-Assisted Text Segmentation** Through multiple refinement rounds, researchers developed a system prompt [5] to separate interviewer and participant and group consecutive utterances, prohibiting paraphrasing, interpreting, summarizing, or adding text. Output required human validation due to challenges with contextualization and the absence of a fixed conversational flow in semi-structured interviews.
3. **AI-Assisted Qualitative Coding** We enriched the data with the interview guide, field notes, and research questions. Researchers developed a system prompt for the AI analysis agent, including detailed instructions, constraints (e.g., one code per row, no generic codes), few-shot examples, and instructions to produce a structured table detailed in Figure 1. Human code reviews and alignment sessions were required.

3 Discussion and Conclusion

Our experience suggests that LLM-supported workflows shift researchers' work from coding to prompt development, validation, interpretation, and output alignment. First, *over time*, LLM updates changed reasoning style, abstraction level, and phrasing, requiring continuous human prompt adaptations to maintain comparable results. Second, *inter-interview* coding varied in wording, emphasis, and granularity, requiring human mapping and prompt refinements to align style and conceptual framing. Third, *in prompt design* more constraining prompts produced surface-level text extraction resembling descriptive coding, while under-constraining prompts improved interpretative richness resembling interpretative coding but reduced consistency and comparability. Finding a balance required human testing, validation, and adjustment. Fourth, *interpretative decisions* remained inconsistent across runs, because the criteria for "analyzing" and "interpreting" remained implicit, resembling critiques of GenAI imitating without engaging in human meaning-making [6]. Finally, *for researchers*, work shifted from meaning-making to validation, and prior work suggests that critical thinking may decline when users expect GenAI outputs to be reliable [7]. We see opportunities for LLM-assisted qualitative analysis in approaches using a human-generated predefined codebook or word counting approaches (e.g., qualitative content analysis); however, when it comes to human meaning-making, it is questionable to what extent, if at all, LLMs should be used, in line with Jowsey et al. [6].

4 Acknowledgment

This project was supported by Innosuisse – Swiss Innovation Agency Innovation Project 123.167 IP-ICT "Serai care: Turning ordinary homes to smart homes to detect and prevent falls for older adults."

References

1. Barros, C.F., Azevedo, B.B., Graciano Neto, V.V., Kassab, M., Kalinowski, M., Do Nascimento, H.A.D., Bandeira, M.C.: Large language model for qualitative research: A systematic mapping study. In: 2025 IEEE/ACM International Workshop on Methodological Issues with Empirical Studies in Software Engineering (WSESE), IEEE (2025) 48–55
2. De Paoli, S.: 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
3. Rasheed, Z., Waseem, M., Ahmad, A., Kemell, K.K., Xiaofeng, W., Duc, A.N., Abrahamsson, P.: Can Large Language Models Serve as Data Analysts? A Multi-Agent Assisted Approach for Qualitative Data Analysis (October 2025) arXiv:2402.01386v2 [cs.SE].
4. Gao, J., Guo, Y., Lim, G., Zhang, T., Zhang, Z., Li, T.J.J., Perrault, S.T.: CollabCoder: A lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. CHI '24, New York, NY, USA, Association for Computing Machinery (2024)
5. Fedosov, A., Kern, D., Barth, C.M.: LLM-powered workflow to support qualitative data analysis: Lessons learned from an applied research project. In: Proceedings of AlpCHI'26 Workshop on 'Crossing Lenses – Exploring HCI/UX Projects Through Academic and Industry Perspectives'. 1st International Conference on Human-Computer Interaction in the Alps (AlpCHI'26). (2026)
6. Jowsey, T., Braun, V., Clarke, V., Lupton, D., Fine, M.: We reject the use of generative artificial intelligence for reflexive qualitative research. Qualitative Inquiry **0**(0) (2025) 10778004251401851
7. Lee, H.P.H., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., Wilson, N.: The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. CHI '25, New York, NY, USA, Association for Computing Machinery (2025)