LLM-supported Thematic Analysis: Evaluating GATOS Workflow on Complex Qualitative Data

Chaewon Kim
Florida State University
ck22j@fsu.edu

Fengfeng Ke
University of Maryland
fke@umd.edu

Nuodi Zhang
Florida State University
nzhang4@fsu.edu

Alex Barrett
Florida State University
abarrett3@fsu.edu

ABSTRACT

Inductive qualitative coding is a labor-intensive process that often is challenging to assess rigor or validity. The development of coding frameworks remains largely subjective, relying on researchers' interpretations. Recent advancements in machine learning, particularly the Generative AI-enabled Theme Organization and Structuring (GATOS) workflow, offer promising solutions to tackle this problem by simulating the inductive coding process while systematically evaluating the rigor of generated codes. However, GATOS has been tested only with synthetic datasets, leaving its effectiveness in real-world qualitative research unexplored. This study empirically evaluates the GATOS workflow by applying it to qualitative data from a research project on supporting neurodiversity in rural communities. The dataset includes open-ended survey responses and semi-structured interview transcripts from healthcare providers, teachers, and caregivers. After following the GATOS workflow for the open-ended survey responses, interview transcripts, and an aggregated dataset of both, the generated final codebook revealed that the themes captured are comprehensive and salient, but certain meaningful summary points were lost in the process of repetitive embedding and clustering. Such findings contribute to the broader discourse on AI integration in qualitative research, exploring how computational tools can complement human expertise while maintaining methodological rigor and interpretive depth.

Keywords

qualitative research, thematic analysis, human-AI collaboration, large language models

1. INTRODUCTION

Inductive qualitative coding, particularly in the stage of developing a coding framework, presents several challenges. First, it is inherently labor-intensive and often lacks a standardized, systematic approach for evaluation [9]. Furthermore, the traditional qualitative coding and thematic framework development process tends to be anecdotal, relying heavily on researchers' subjective interpretations without a structured mechanism for assessing trustworthiness [7]. In response to these challenges, many researchers are exploring how recent advancements in machine learning can enhance qualitative coding processes [1][2][4][6][8][10]. Among these innovations, the Generative AI-enabled Theme Organization and Structuring (GATOS) workflow [5] stands out as a promising approach. It is one of the few recent models that explicitly outlines a workflow designed to simulate the inductive qualitative coding process while systematically evaluating the rigor of the generated codes. The model imitates qualitative researchers' reasoning process, iteratively reviewing the data, codes, and critically assessing the emergent themes.

However, to date, the GATOS workflow has been tested only with synthetic datasets, and its effectiveness in real-world qualitative research remains unverified. In practical applications, qualitative coding is more complex than what synthetic data simulations may capture. Real-world data often reflect nuanced and diverse perspectives, particularly when addressing multifaceted social issues such as supporting neurodiversity in rural communities. Moreover, the qualitative research process frequently integrates multiple data sources—such as open-ended survey responses and in-depth interviews—to ensure a comprehensive understanding of shared opinions and unique narratives [3].

As such, this study seeks to empirically investigate the application of the GATOS workflow in constructing a comprehensive, well-structured codebook derived from both open-ended survey responses and interview data. Specifically, it aims to understand how this AI-assisted workflow compares to traditional, human-driven qualitative coding, examining its strengths, limitations, and potential for complementing or enhancing human analysis. By bridging the gap between computational assistance and human expertise, this research contributes to the broader discourse on integrating AI into qualitative research while maintaining methodological rigor and interpretive depth.

2. METHOD

2.1 Context

The dataset used for this study was collected as part of a bigger project that aims to investigate the care ecosystem for neurodiverse individuals in rural areas. The research team collected survey responses for open-ended questions from healthcare providers, teachers, and caregivers who support individuals with neurodiversity in rural areas and selectively invited them for an hour-long semi-structured interview to gain an in-depth understanding of their perspectives and experiences as participants in rural care ecosystems.

2.2 Data

Two types of qualitative datasets were used: open-ended survey responses and individual interviewing transcripts. The survey dataset consists of 86 sets of open-ended responses from 8 healthcare providers, 62 teachers, and 16 caregivers (i.e., parents or family members) of individuals with neurodiversity in rural areas. The interview dataset consists of five semi-structured interview transcripts from 1 healthcare provider, 2 teachers, and 2 caregivers. In both survey and interview, the participants were asked different

Chaewon Kim, Fengfeng Ke, Nuodi Zhang, and Alex Barrett. LLM-supported Thematic Analysis: Evaluating GATOS Workflow on Complex Qualitative Data. In Caitlin Mills, Giora Alexandron, Davide Taibi, Giosué Lo Bosco, and Luc Paquette (eds.) Proceedings of the 18th International Conference on Educational Data Mining, Palermo, Italy, July, 2025, pp. 604–607. International Educational Data Mining Society (2025).

https://doi.org/10.5281/zenodo.15870241

sets of questions tailored to their roles, primarily focusing on the challenges they faced in supporting individuals with neurodiversity.

GATOS workflow involves the set-up of preliminary, baseline codes in its inductive code generation step. To pursue a balance between data-driven and literature-driven qualitative codebook generation, we used the codes emerging from a recent systematic literature review on rural ecosystems for neurodiversity as our preliminary or baseline codebook.

2.3 GATOS Workflow

The GATOS workflow consists of five steps, leveraging machine learning techniques and generative AI tools. First, researchers summarize the raw data into text units of analysis. For example, open-ended survey responses and interview transcripts are summarized into individual points or sentences by prompting large language models (LLMs). Second, a text embedding model is used to convert the summary points into high-dimensional numeric representations. Third, the dimensionality of the embeddings is reduced to simplify further processing. Fourth, the reduced embeddings are clustered, with similar points grouped together based on semantic meaning. Finally, the clusters are identified by finding their nearest neighbor codes from the extant, baseline codebook. The cosine similarity between the embeddings of the summary points and the codes is calculated, and these nearest neighbors are aggregated, removing redundancies to generate a list of unique codes. The resulting codes are then used in the generative LLM model's prompt to decide whether a new code should be created and added to the extant codebook. Detailed stepwise instructions and prompts used while interacting with the LLM can be found in [5].

In this study, we adapted the workflow to process and fuse the survey and interview data. In the first step, we summarized both datasets individually, obtaining 68 summary points from survey responses and 102 from interview transcripts. The summary points were then converted into high-dimensional numeric representations using the *mixedbread-ai/mxbai-embed-large-v1* model, a pre-trained open-source word embedding model. After dimensional reduction, K-nearest neighbor clustering created 20 clusters from the survey data and 29 from the interview data. In the final step, these clusters were identified by finding the nearest neighbor codes from the predeveloped, baseline codebook based on a systematic literature review on the topic. The two sets of clusters were then aggregated to generate a set of unique codes. We followed and customized the prompts in [5] for iterative prompting of the large LLM models to refine the final codebook with themes and codes. The detailed workflow is illustrated in Figure 1.

3. RESULTS

Following the GATOS workflow, we generated three codebooks, based on 1) only survey open-ended responses, 2) only interview transcripts, and 3) an aggregated dataset. The same preliminary/baseline codebook was used to generate all three codebooks. The generated codebooks, respectively, consisted of 4 themes and 25 codes based on the survey open-ended responses, 7 themes and 20 codes based on the interview transcripts, and 7 themes and 22 codes based on the aggregated dataset. See Table 1 for the final codebook for the aggregated dataset.

While the themes and codes from three codebooks generally overlap and capture various perspectives from the stakeholders, some codes that emerged from the survey or interview data only did not appear in the codebook based on the aggregated dataset. For example, 'importance of involving neurodiverse individuals and families in decision-making' and 'the need of smaller class sizes and additional classroom support staff' emerged from the open-ended survey responses but was not explicated in the final codebook provided in Table 1. Some codes based on the interview transcript, such as 'challenges in transitioning from school to adulthood' and 'need for early intervention programs' were not highlighted in the final codebook.

Importantly, the final codebook extends the literature-informed codebook by capturg localized specifications. For example, "Environmental and Sensory Needs" was a new theme that was not represented in the literature, reflecting unique lived experiences and expectations of the participants. At the same time, this localization was also limited by scope. In the literature-driven codebook, we had a theme named "Lack of inclusive opportunities for improving essential skills," which was not included in the final codebook as our participants were mostly caregivers and teachers of neurodiverse children, the experiences of neurodiverse adults and their long-term well-being and independence was not the primary focus of our participants.

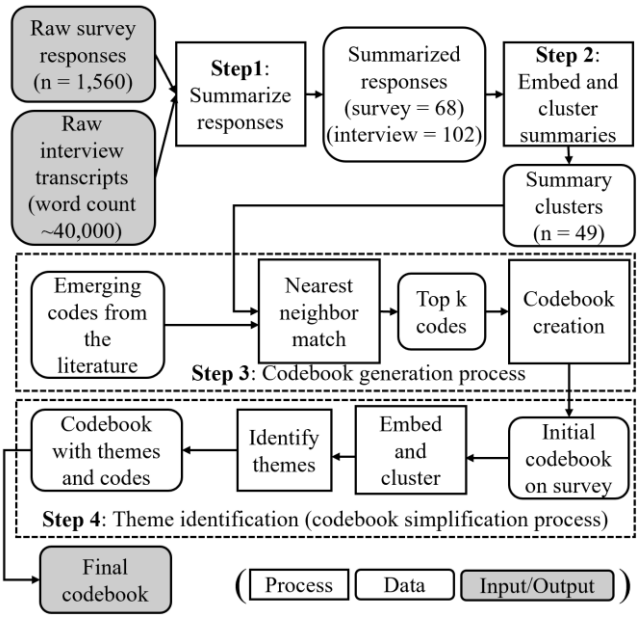

Figure 1. Modified GATOS workflow

4. DISCUSSION

LLMs are expanding the ways we analyze, conduct research, and understand societal issues. The GATOS workflow and this study's findings suggest that AI and human researchers can collaborate to gain rich qualitative insights into complex topics. In this study, we applied the GATOS workflow to the empirical survey and interview datasets, making necessary adjustments to accommodate and fuse two data types. The generated final codebook identified 7 themes supported by a total of 22 codes, capturing comprehensive perspectives from teachers, caregivers, and healthcare providers.

Notably, we observed that the final codebook resembled the one based on the interview data more than that on the survey data. This suggests that interview responses had a stronger influence on the coding process, potentially due to the volume and richness of qualitative interviews compared to open-ended survey responses. One point of concern is the disappearance of certain codes during

Table 1. Final codebook

Themes	Concept	Codes
Resource Limitations	Challenges related to the lack of funding, staff, and specialized services in rural areas.	- Limited access to specialized services - Financial constraints in rural schools - Lack of trained staff - High burnout rates among educators - Limited availability of mental health services - Lack of funding for community programs
Professional Development and Training	The need for ongoing training and skill development for educators, caregivers, and healthcare providers.	- Need for professional development - Importance of caregiver training - Importance of trauma-informed care
Systemic and Structural Barriers	Systemic challenges such as healthcare access, transportation, and external disruptions.	- Challenges in accessing healthcare - Transportation barriers in rural areas - Impact of external factors (e.g., COVID-19, natural disasters) - Cultural norms influencing care
Individualized Support and Collaboration	The importance of personalized support and collaboration among stakeholders.	- Importance of individualized education plans (IEPs) - Collaboration with community organizations - Role of social workers in connecting families to resources - Importance of clear communication between stakeholders
Awareness and Stigma	Societal stigma and lack of awareness about neurodiversity.	- Stigma and lack of awareness about neurodiversity - Need for public awareness campaigns
Intersectionality and Equity	The intersectional challenges faced by neurodiverse individuals and the need for equitable support.	- Intersectionality (e.g., race, gender, socioeconomic status) - Challenges in balancing neurodiverse and neurotypical student needs
Environmental and Sensory Needs	The need for sensory-friendly environments to support neurodiverse individuals	- Need for sensory-friendly environments

the embedding and clustering process when aggregating the dataset. Multiple codes considered salient in the survey-only or the interview-only codebook failed to appear in the final aggregated structure. This suggests that the dimensionality reduction and clustering steps may inadvertently filter out less frequent yet meaningful concepts, potentially overlooking nuances that emerge in smaller subsets of the data. This limitation highlights the need for careful and human-in-the-loop evaluation of how AI-assisted workflows handle thematic integration across diverse qualitative datasets.

Overall, the GATOS workflow successfully captured the breadth of topics discussed in the qualitative data; however, some nuances are lacking. Many codes are framed in a standard structure as "need for something," "lack of something," or "importance of something," which does not fully convey how these codes interact with each other to portray interconnected themes depicting lived experiences. Future iterations of this approach could enhance the depth of analysis by incorporating relational coding strategies that better illustrate the interplay between themes.

We used the freely available DeepSeek model in this study, and it is important to acknowledge that differing LLMs might have produced different results. As AI tools continue to evolve, their role in qualitative research will likely grow, offering new ways to assist with analysis and interpretation. Further research is needed to refine AI-assisted qualitative coding methods to ensure that they not only identify key themes but also capture the complexity of human experiences with greater fidelity and richness.

5. CONCLUSION

This study demonstrates how AI-assisted workflows like GATOS can augment qualitative research by efficiently and systematically synthesizing diverse perspectives in an authentic context. The modified approach provided in this paper shows how two different types of qualitative data, such as open-ended survey responses and interview transcripts, can be fused and analyzed using data mining techniques and adequate LLM prompting. As AI continues to evolve, its role in qualitative research will likely expand, offering new possibilities for analysis and interpretation.

6. ACKNOWLEDGMENTS

We would like to thank all the interview participants for generously sharing their time and insights, which significantly contributed to the depth and quality of our research.

7. REFERENCES

[1] Chew, R., Bollenbacher, J., Wenger, M., Speer, J., and Kim, A. 2023. LLM-assisted content analysis: Using large language models to support deductive coding. *arXiv preprint arXiv:2306.14924.*

[2] 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. *Soc. Sci. Comput. Rev.* 42, 4, 997–1019.

[3] DiCicco-Bloom, B. and Crabtree, B. F. 2006. The qualitative research interview. *Med. Educ.* 40, 4, 314-321.

[4] Gao, J., Guo, Y., Lim, G., Zhang, T., Zhang, Z., Li, T. J.-J., and Perrault, S. T. 2024. CollabCoder: A lower-barrier,

rigorous workflow for inductive collaborative qualitative analysis with large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–29.

[5] Katz, A., Fleming, G. C., and Main, J. 2024. Thematic Analysis with Open-Source Generative AI and Machine Learning: A New Method for Inductive Qualitative Codebook Development. *arXiv preprint arXiv:2410.03721*.

[6] Katz, A., Shakir, U., and Chambers, B. 2023. The utility of large language models and generative AI for education research. *arXiv preprint arXiv:2305.18125*.

[7] Lincoln, Y. S. and Guba, E. G. 1986. But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. *New Dir. Program Eval.* 30, 73-84.

[8] Nguyen, H., Ahn, J., Belgrave, A., Lee, J., Cawelti, L., Kim, H. E., et al. 2021. Establishing trustworthiness through algorithmic approaches to qualitative research. In *Advances in Quantitative Ethnography: Second International Conference, ICQE 2020* (Malibu, CA, USA, February 1-3, 2021). Springer International Publishing, 47-61.

[9] Renz, S. M., Carrington, J. M., and Badger, T. A. 2018. Two strategies for qualitative content analysis: An intramethod approach to triangulation. *Qual. Health Res.* 28, 5, 824-831.

[10] Tschisgale, P., Wulff, P., and Kubsch, M. 2023. Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory. *Phys. Rev. Phys. Educ. Res.* 19, 2, 020123.