Bridging Human and Machine Perspectives: Integrating Large Language Models into Collaborative Coding and Peer Debriefing for Qualitative Inquiry*

Haoning Jiang[1,*], Seth Corrigan[1], Kylie Peppler[1] and Tisa Islam Erana[2]

[1] *University of California, Irvine, Irvine, CA 92697*
[2] *Florida International University, 11200 SW 8th St, Miami, FL 33199*

Abstract

Large Language Models (LLMs) offer promising opportunities to enhance qualitative research, particularly in collaborative coding and peer debriefing — practices that improve rigor, reflexivity, and coder agreement. LLMs can support these processes by clarifying evidence and warrants behind coders' claims, allowing for the refinement of reasoning and the surfacing of new perspectives. This study explores how LLMs can aid in collaborative coding and debriefing by suggesting alternative interpretations, highlighting ambiguities, and proposing overlooked warrants. Using a dataset of 104 transcribed interviews with artists, we integrated an LLM, GPT-4o, into the coding process. Findings show that GPT-4o, through being integrated into an iterative collaborative coding process, was able to surface ambiguities in key concepts during the coding process and effectively align with human coders (average F1 score across 3 trials: 0.83), illustrating the potential of LLMs as valuable tools in qualitative coding and research.

Keywords

LLM, collaborative coding, iterative coding, qualitative research

1. Introduction

Large Language Models (LLMs) offer novel opportunities to enhance qualitative research by supporting collaborative coding and peer debriefing, two key practices that improve rigor, reflexivity, and coder agreement. Collaborative coding and peer debriefing [1] [2] are processes that pool diverse perspectives to address disagreements and ambiguities, improve inter-rater agreement, challenge researcher assumptions, and introduce alternative interpretations.

The utility of LLMs in the collaborative coding process thus lies in the LLM's capacity to make explicit the evidence and warrants underpinning coders' claims, which allows researchers to refine and re-examine their reasoning. The integration of LLMs introduces a bidirectional influence: while prompts and LLM parameters are refined for better alignment with researchers, researchers' own views may also be transformed through the LLM's suggested codes, evidence, and reasoning.

This research investigates how LLMs can enhance these processes by suggesting alternative perspectives on evidence, surfacing ambiguities, and proposing potential warrants overlooked by human coders. Using a dataset of 104 transcribed semi-structured interviews with artists recounting outcomes of their arts participation and an existing hierarchical taxonomy of codes, we developed and deployed a pipeline incorporating a pre-trained LLM (GPT-4o) into an integrated human-machine collaborative coding and debriefing process. During this process, discussions between human coders and the LLM were recorded and analyzed to identify instances where the LLM's rationale — structured as claims, evidence, and warrants — led to modifications to prompts, adjustments to the researchers' codebook, or shifts in the researchers' perspectives.

**Joint Proceedings of LAK 2025 Workshops, co-located with the 15th International Conference on Learning Analytics and Knowledge (LAK 2025), Dublin, Ireland, March 03–07, 2025.*

[1*] Corresponding author.

haoningj@uci.edu (H. Jiang); scorrig1@uci.edu (S. Corrigan); kpeppler@uci.edu (K. Peppler) ; tisaislamerana@gmail.com (T. I. Erana)

This research investigates the following questions: 1) How can LLMs be effectively integrated into the collaborative coding process, such that sufficiently high model performance is achieved? and 2) How can the reasoning put forth by an LLM for its label be used to surface ambiguities? By investigating these questions, this study aims to deepen our understanding of the role LLMs can play in qualitative research methodologies and their potential to support human expertise in collaborative coding and debriefing processes.

2. Background and Related Works

2.1. Claim, Grounds, and Warrant

The Toulmin model of argumentation presents the following framework for making a persuasive argument: the claim, which is the key argument the speaker wishes to make; the grounds, which is the concrete evidence that backs up the speaker's claim; and the warrant, which is the reasoning that connects the grounds to the claim. The warrant essentially explains how and why the grounds is relevant to the claim and how the grounds specifically implies that the claim is true. [3]

2.2. Collaborative Coding and Repair

The classification of qualitative data — such as interview transcripts, texts, field notes, and so on — involves a level of subjectivity, since concepts must be defined and there exist boundary cases whose classifications are unclear. Thus, the coding process (i.e., the process in which human labels are assigned to the data) should be consistent across time and across different coders, and the codebook each coder uses (ie. the definitions of key concepts and labels used to determine when a label should be assigned) should also be consistent across coders. The processes of collaborative coding and peer debrief allow coders to align on key definitions and develop a shared consensus and understanding, which in turn improves rigor, reflexivity, and coder agreement. [1] [2]

Qualitative labeling is traditionally performed in an iterative manner, with researchers labeling one or more segments of text, discussing areas of disagreement, refining their labeling rules, and then continuing on with additional segments of the data, ultimately labeling with fewer discussions and improved alignment between the labellers, measured through one or more estimates of inter rater reliability [4] [5] [6]. Alignment between human coders and an LLM can be accomplished by integrating the LLM into this same iterative process. This iterative process is similar to what has been described as other-initiated other-repair (OIOR) [7] [8] — with the coders initiating repair of each other's understandings of the coding task and key concepts involved.

We posit that, by having each coder use the Toulmin model of argumentation [3] to make explicit what evidence and reasoning they are employing to back up their label, discrepancies between coder understandings and weaknesses in existing conceptual constructs in the codebook can be more easily exposed: whether through comparing different pieces of evidence used to argue different claims, or through comparing different lines of reasoning employed to analyze the same piece of evidence in different directions.

2.3. Large language model (LLM)

A Large Language Model (LLM) is a kind of artificial intelligence program that uses deep learning to perform natural language tasks, such as language generation and text analysis. Examples of LLMs include BERT and OpenAI's ChatGPT. Unlike many other artificial intelligence programs, LLMs are typically pre-trained on massive sets of data and then fine-tuned for specific tasks; thus, for the average user, training the LLM itself is unnecessary [9].

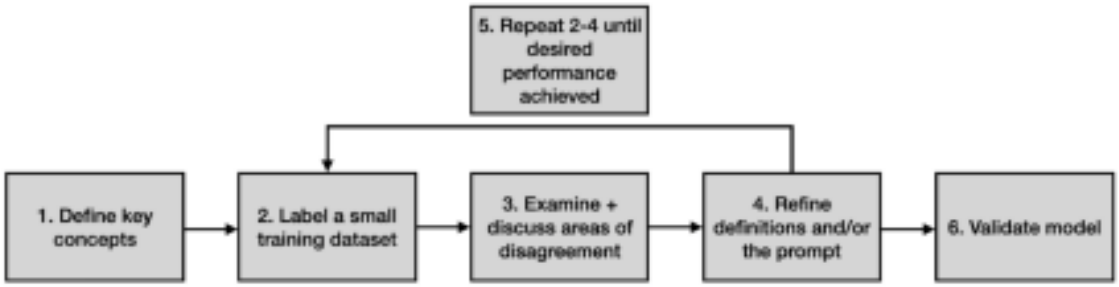

Figure 1: Sample pipeline involving an LLM in a text classification process.

Large Language Models (LLMs) have rapidly emerged as powerful tools for processing and reasoning about large volumes of data. Their ability to perform a wide variety of natural language processing tasks, such as coding or labeling text based on user-provided instructions, makes them useful for qualitative research.

2.4. Iterative Prompt Refinement with LLMs

While LLMs can be highly useful for qualitative coding, qualitative text classification tasks still involve a level of subjectivity. Furthermore, LLMs can be biased or unreliable, misunderstand directions, and produce unexpected results [10]. Therefore, the coding process (i.e., the process in which human labels are assigned to the data) should be consistent across time and across different coders—human and model based coders. As a means to bring about alignment between human-based and model-based coders, the LLM should be included in a systematic coding process alongside researchers in ways that afford humans insight into a) whether the LLM is applying the prompt instructions as intended, and b) whether the LLM's resulting labels are well aligned with the human members of the coding team.

Therefore, following Törnberg (2024) [11], we present the following systematic labeling procedure for the integration of an LLM into the collaborative coding process. This iterative process requires the LLM to provide an explanation and rationale for each of its labels. The LLM's explanation and rationale provides a "window" of sorts into the LLMs handling of the given content and can provide information relevant for improved control over the LLM's performance and ultimately its alignment with the human coders. Just as alignment between different human coders can be achieved by using a collaborative, iterative, coding process, alignment between human coders and an LLM can also be accomplished by integrating the LLM into such a collaborative process.

Consider the example of a systematic coding pipeline integrating an LLM provided in Figure 1 [11]:

1. Define key concepts. Provide both the researchers and the LLM with high-quality and clear definitions of all involved concepts, as well as clear instructions as to how the data should be classified.
2. Label a small training dataset. Have your researchers assign labels to a small subset of the data. Then have your LLM assign labels to the same data.
3. Examine and discuss areas of disagreement. Examine the areas in this small dataset in which the labels assigned by the researchers and the labels assigned by the LLM disagree. Ask your researchers to explain their reasoning; in your prompt to the LLM, ask the LLM to explain its reasoning as well. Based on the provided justifications, decide which label is correct and/or which concepts are being misinterpreted, and by whom.
4. Refine definitions and/or the prompt. Based on your analysis of the areas of disagreement above, make adjustments to your concept definitions and prompts to either the researchers and/or the LLM.

5. Repeat steps 2-4. Continue to refine the prompt with this process until the desired output quality is achieved on the small training dataset.
6. C Validate your model. Have your researchers annotate all of the remaining data, then have your LLM annotate all of the remaining data as well. Then assess the performance of the LLM, compared to the labels assigned by the researchers, with a useful measure such as the F1 score.

As presented here, our pipeline operationalizes the notion of "repair" in conversation analysis [7]. In our pipeline, the LLM is required to provide an explanation and rationale for its labels, or codes, so that whenever an area of disagreement is identified, the explanation and rationale for the LLM's labels can be reviewed and considered by the researcher(s). The researchers can then make joint decisions to either adopt the LLM's approach and modify their own coding rules, or to modify the LLM's approach by modifying one or more aspects of the prompt—either the terms used, the instruction as a whole and/or the examples provided. Such refinements are intended to "repair" misalignments between researcher(s) and the LLM, and ensure each labeller, human and LLM alike, maintains sufficiently similar definitions, examples and understandings.

3. Methodology

We apply this iterative coding process to a classification task: the binary classification of existing segments of interview transcripts as to whether a specific given topic is discussed in that interview segment or not. We first deploy the iterative section of the above coding process on a sample dataset of 10 interview segments. We then have the human researchers and the LLM code the larger dataset of 100 interview segments separately. Finally, we calculate similarity metrics between the human labels and the LLM's labels to assess the LLM's performance, and analyze the areas of disagreement between the human labels and the LLM labels.

3.1. Dataset: Interview Segments on Experiences in Arts Education

For our data, we used a corpus of transcribed interviews with n = 104 adults reporting lasting impacts of their arts participation. Each interview is divided into segments of 30 lines total. Using the corpus of interviews, we aimed to draw conclusions about the long-term impacts of arts participation. For the current paper, we worked to detect when the topic of "meaningful relationships formed and/or deepened through shared interests in the arts, creative pursuits, and/or arts mentoring" was discussed.

3.2. Task Setup

We aimed to have OpenAI's ChatGPT API perform the following binary classification task: classify each interview transcript as to whether the target topic of friendship through shared interests in the arts was discussed or not. We began by operationalizing our target topic and by writing a prompt directing the LLM to identify instances of the target topic in the corpus. We then pulled a subset of 10 interview segments from our dataset to jointly code, in order to refine our human understandings and refine our prompt to the LLM.

3.3. Labeling Pipeline

We began with a description of the target discussion topic (ie. youth arts programs facilitating friendship through shared interests) and wrote a prompt asking the LLM to return 1 if the prompt was discussed in the "Coding" section and 0 otherwise. We improved our prompt using the general guidelines given above by structuring the prompt properly, by listing all of the classification options (1, 0, or "I don't know), and by providing some few-shot examples of positive and negative utterances.

To iterate our prompt, we selected a subset of 10 interview segments from our dataset. We first had two researchers individually label each interview segment as either 1 (the target topic was

discussed) or 0 (the target topic was not discussed), and also ran our prompt on each interview segment with the LLM. We then examined areas of disagreement between each researcher and between the researchers and the LLM; the researchers aligned with each other what we meant to include and exclude in our target topic after weighing the LLM's reasoning as well, and we then updated our human definitions and our LLM prompt to reflect this alignment. All together, we iterated over the subset 3 times before we achieved total agreement with each other and with the LLM on the sample subset.

Table 1 lists the adjustments we made to both the definitions used by our researchers and to the prompt given to the LLM during this alignment process.

Table 1
Iterative refinement and alignment of key definitions and prompts using sample subset of interview segments.

Iteration	Updates to Human Definitions	Updates to LLM Prompt
0 (start)	Beginning definition of target topic: mean ingful connections formed through shared interests in the arts in youth arts programs.	Provided this definition in the prompt.
1	Clarified scope of target topic: includes bonds formed and deepened by shared in terests in all arts, not just bonds formed between peers in youth arts programs. Includes existing friendships being deep ened; do not limit to new friendships only. Added wording about ongoing relation ships to the section of prompt about relationships. Deleted wording that implied positive responses should be limited to new relationships only.	Added wording about ongoing relation ships to the section of prompt about re lationships. Deleted wording that implied positive responses should be limited to new relationships only.
2	Include shared interest in arts mentoring. Do not limit to peer-to-peer friendships; include meaningful relationships between mentors and mentees over mutual arts pas sion as well.	Added wording pointing to arts mentoring to the prompt.
3	None	Added positive example utterance involv ing arts mentoring facilitating a meaning ful relationship.

We began with a broad formulation of the target topic, merely stating that the LLM should include all discussions of friendships formed through shared interests in the arts.

In our first iteration, the LLM identified and labeled as positive evidence that one piece of researchers both human in a certain had missed: segment, the interview discussed a interviewee with their friendship janitor as school's shared facilitated by interest. poetry human Though the initially researchers to whether disagreed as this should

count, we eventually concluded that the LLM's reasoning held water, and thus agreed that all bonds, not just bonds involving two peers from the same youth arts program, should count under the target topic. We also clarified that the target topic should include additional cases where existing relationships were deepened by shared interest in the arts, and not be limited to just the formation of new relationships.

In our second iteration, the LLM presented a different line of reasoning on a certain piece of evidence than the human researchers: the LLM excluded and labeled as negative a case where two peer mentors bonded over their shared passion for arts mentoring. While both human researchers had included this as an example of the target topic, the LLM instead reasoned that passion for arts mentoring was not the same as passion for the arts itself. After discussion, we decided that meaningful connections based upon shared interest in arts mentoring should also be included. Thus, in our third iteration, in order to help add this to the LLM's understanding of the target topic, we identified and provided an example of an utterance demonstrating this idea.

After the third iteration, the labels on the sample subset generated by each researcher, as well as the labels on the sample subset generated by the LLM under our prompt, were completely in agreement, so we ended the iterative alignment process here.

3.4. Validation

Next, we selected a new subset of 100 interview segments from our dataset. We ensured there was no overlap between this set of 100 interview segments and the previously used subset of 10 interview segments. Our two researchers individually labeled each interview segment as either 1 (the target topic was discussed) or 0 (the target topic was not discussed) based on the human definitions from the final iteration above. Then we created a final set of labels for these 100 interview segments: '1' if any labeller had labeled an interview segment as '1', and '0' otherwise. Our set of 100 interview segments ended up having 49 interview segments human-labeled as '0' (the target topic was not discussed), and 51 interview segments human-labeled as '1' (meaning the target topic was discussed).

Finally, to validate our model, we ran the final iteration of our prompt on each of these 100 interview segments through the LLM and recorded the LLM's answers and reasoning. For our LLM, we used OpenAI's ChatGPT API, with the GPT-4 model. We then compared the LLM's labels to the human labels with a variety of performance metrics. To check the consistency of our model, we ran the same final prompt on each of the same 100 interview segments three times, in three separate trials referred to as Trial 1, Trial 2, and Trial 3

4. Results

Overall, as seen in Table 2, the performance of the LLM as directed by our prompt yielded an F1 score, which measures the balance between the precision (the proportion of true positives to positive predictions) and recall (the proportion of true positives to actual positive instances), averaging at around 0.82 (with a score of 1.0 being perfect). That the F1 score is relatively close to 1.0 suggests a good trade-off between precision and recall.

Across all 3 trials, we had more false positives than false negatives, which is why our precision in each trial is lower than our recall in each trial. For the task of identifying whether a target topic was discussed in a given interview segment or not, a model that produces more false positives than false negatives may in fact be preferable: a qualitative researcher is most likely to give such a task to a model because they wish to further investigate the interview segments in which the target topic was identified, and in such a case, it may be preferable to investigate a few interview segments that do not actually contain the target topic than it would be to miss a few interview segments that do in fact contain the target topic.

Table 2

Validation Performance Metrics. The precision, recall, and F1 score for each of the three trials are listed above. Each trial uses the exact same final prompt on the exact same model and the exact same 104 interview segments.

Trial	Precision	Recall	F1 Score
Trial 1	0.763	0.882	0.818
Trial 2	0.789	0.882	0.833
Trial 3	0.776	0.882	0.826

In sum, this suggests that, using the above iterative labeling pipeline and prompting techniques, acceptable performance on a text classification task can be achieved by an LLM.

5. Discussion

After calculating the performance metrics above, we analyzed the cases in which the LLM label and the human labels did not match. Here, we summarize the team's analysis of these areas of disagreement: in which cases we concluded the LLM was wrong, and in which cases we concluded that the LLM had effectively identified an area of ambiguity because its reasoning, though divergent, was as justified and sound as our own.

False positives are instances where the LLM indicated the presence of the construct but the researchers did not. False negatives are instances where the LLM indicated the construct was not present when the researchers had indicated it was present.

5.1. False Positives

All of the false positive results involved interview segments that included discussion of either passion for the arts or meaningful connections. All of the false positive results involved the LLM being more generous in its definitions of "shared interests in the arts" or "meaningful relationships": either the LLM was more willing than the researchers to assume that a meaningful relationship existed, or the LLM was more willing than the researchers to declare that a meaningful relationship had been built around shared interests in the arts. For some of these false positives, we conclude that the LLM was merely overzealous, in that it was either directly incorrect or made a decent but less persuasive argument than the human researchers; for other false positives, however, we instead conclude that the LLM has identified an area of ambiguity in our codebook definitions.

In a few cases, the LLM was simply incorrect. For example, one segment had the speaker discussing bonding with another student during a summer soccer camp. The LLM labeled this as 1 because the speaker and their friend shared a passion; however, the LLM should have labeled this as 0 because said passion was not in the arts specifically.

In a few other cases, the LLM reasoned decently but less effectively than the human researchers as to what was implied by an interview segment, and thus came to different conclusions. For example, in one interview segment, the interviewee discussed how she and many other alumni from a dance program put on a joint dance performance at a memorial for the dance program's founder. Since nothing the interview said indicated that she had any personal connection to the other alumni besides working on a project together, both researchers labeled this segment as 0. However, the LLM instead reasoned that the alumni working together implied that they had a meaningful relationship founded upon shared passion for dance.

In this case, we concluded that the LLM was overzealous in its labeling. The human researchers and the LLM disagreed on their reasoning for the same piece of evidence: while the human researchers said that the experience of working on a project with other program alumni was not

enough to imply a personal connection, the LLM had a different, excessively generous interpretation of the same evidence, and thus reached a different conclusion.

In other cases, however, the LLM's reasoning was deemed sound. The LLM's justification could be considered to be equally or even more reasonable than the researchers' justifications, such that we instead concluded that the LLM had found an area of ambiguity. For example, in another interview segment, the speaker discusses their interactions with their English teacher over creative writing: the English teacher recognized that the speaker was not fitting in at school and encouraged the speaker to express themselves through writing; the speaker also mentions the teacher creating a creative writing program. Both human researchers felt that this was more of a case of a teacher being a good mentor to a student than it was a meaningful friendship specifically between the teacher and the student. However, the LLM instead reasoned that the teacher's encouragement of the speaker to express themselves through writing led to mutual engagement in the creative writing process, and that the teacher's creation of the creative writing program created a community in shared creative interests — and that, because of these factors, the relationship between the speaker and their teacher was closer than that of an average student and their teacher, and therefore counted as a meaningful relationship.

Several other false positives also fell into this category, in that the speaker described a relationship with a teacher or mentor figure, the human researchers coded those segments as 0, but the LLM coded them as 1 instead, reasoning that those mentorship relationships were both meaningful and based in shared arts passion. In these cases, the LLM has surfaced an ambiguity in our understanding of key concepts, which is the question of when the relationship between a teacher and a student stops being that of an ordinary instructor and student, and starts becoming a meaningful connection instead. Students can form incredibly close interpersonal bonds with their mentors and consider their mentors to be lifelong friends; however, it is also possible for a student to attend a class or a series of lectures and effectively learn all of its content, while barely interacting with the teacher or lecturer at all. Therefore, where this deciding line should be drawn is an area of ambiguity, one that the LLM effectively surfaced.

5.2. False Negatives

Overall, there were fewer false negatives than false positives. Furthermore, all of the false negative cases were the result of areas of ambiguity, in that for every case, we concluded that the LLM's arguments for its labels were reasonable and sound. For these cases, we conclude that the LLM had found areas of ambiguity because its reasoning, though divergent from our own, still effectively utilized evidence and reasoned coherently about said evidence in order to form a believable argument for its label; thus, the LLM's reasoning could not be said to be inferior to that of the human labellers.

For example, in one interview segment, the interviewee discusses working with three other individuals on hip-hop music. Both researchers thought that the reported history of creative collaboration between these individuals implied a meaningful relationship between them. But the LLM instead reasoned that, because the speaker's language was more focused on professional development and was not especially warm, the relationship between these individuals was most likely purely professional and therefore did not count as a meaningful relationship founded or deepened by shared arts passion.

In another interview segment, the interviewee discussed how an in-person interaction with a famous culinary chef created the feeling of a personal connection, which then inspired the interviewee in their own culinary crafts. Both researchers counted culinary art as an art form and counted the interaction with the famous chef as a meaningful connection. However, the LLM instead reasoned that this was a purely professional connection and not a connection based on shared interests in artistic passion. In this case, the LLM's reasoning that a professional meetup does not count as a meaningful connection is entirely understandable.

For both of these above examples, the area of ambiguity the LLM had surfaced was the question of when a professional connection should also count as a meaningful personal relationship, versus

when it should count as merely a professional connection without the personal significance factor. Professional connections can be highly beneficial and therefore significant, in that they provide personal career development, new opportunities, inspiration, and so on, but that does not necessarily make them always equivalent to the sort of personal, friendship-based relationship we had in mind for our target topic. Therefore, the question of where and how the deciding line should be drawn is an area of ambiguity.

Since almost all of the false positives and false negatives in which an area of ambiguity was surfaced had to do with what constituted a ”meaningful” relationship, it can be said that the biggest ambiguity that the LLM had surfaced was the question of what entailed a ”meaningful” connection itself, as well as the question as to what degree such a connection should be founded upon shared interest in order to count. This then implies that the ambiguity was inherent to the task itself, as the question of what ”meaningful relationships” entail is subjective and varies even from researcher to researcher.

The word ’meaningful’ itself is already vague and highly dependent on the labeller’s own experiences. For example, one of our researchers was willing to code the dance interview segment discussed above as 0 because they had ample experience working with others on shared projects and not forming any lasting bonds with other project members; however, a different labeller who did form long-lasting and intimate bonds with others after working on a shared project together might feel differently. Perhaps these ambiguities could have been clarified by replacing individual words such as “meaningful” with more specific triples, such as personally significant, intimate, and lasting.

6. Conclusion

We have incorporated an LLM, GPT-4o, into an iterative collaborative coding process, in which we aligned the understandings of human researchers and the LLM of the target topic and key definitions. Our iterative labeling process was based upon the notion of repair in discourse analysis: in each iteration, we identified areas of disagreement that indicated problems in understanding or communication, shared our reasoning with each other and with the LLM to better understand how each labeller was coming to their conclusions, and then repaired those areas of misalignment by clarifying to each other and in our prompt to the LLM what we meant. Through this process, the LLM effectively surfaced ambiguities and areas that needed clarification in our existing codebook.

Our findings are significant in demonstrating how LLMs can improve inter-rater agreement, contribute to the deepening of understandings of the core ideas behind constructs, and potentially support the development of data-driven theories. Incorporating the LLM in collaborative coding and debriefing process resulted in high precision (0.87), accuracy (0.84), and acceptable recall (0.79) when the LLM was tasked to identify interview segments where artists described developing close relationships with friends and mentors as outcomes of their arts participation. The overall F1 score for the classification task was 0.83, reflecting a balanced performance. Additionally, during peer debriefing simulations, the LLM effectively surfaced ambiguities in prompts and code definitions, as well as expanded the scope of evidence and warrants considered by researchers.

These findings underscore the potential of LLMs to enrich qualitative research by informing the iterative coding process, challenging assumptions, and enhancing the depth of analyses. While not a replacement for human interpretation, LLMs serve as supports that complement human expertise. This study contributes to the growing body of literature on integrating AI technologies into qualitative methodologies, offering practical insights for researchers aiming to harness LLMs as partners in coding and debriefing processes.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4o to support qualitative coding by classifying interview segments and providing rationale for each label. After using this tool, the

authors reviewed and edited all outputs as needed and take full responsibility for the publication's content.

References

[1] K. A. R. Richards, M. A. Hemphill, A practical guide to collaborative qualitative data analysis, Journal of Teaching in Physical Education 37 (2018) 225–231. doi:10.1123/jtpe.2017- 0084.

[2] J. P. Barber, K. K. Walczak, Conscience and critic: Peer debriefing strategies in grounded theory research, Qualitative Research in Psychology 6 (2009) 257–274. doi:10.1080/14780880902973832.

[3] S. Toulmin, The uses of argument (2nd ed.), Cambridge University Press, 2003.

[4] W. A. Stock, Systematic coding for research synthesis, in: H. Cooper, L. V. Hedges (Eds.), The handbook of research synthesis, Russell Sage Foundation, 1994, pp. 125–138.

[5] Y. Chun Tie, M. Birks, K. Francis, Grounded theory research: A design framework for novice researchers, SAGE Open Medicine 7 (2019) doi:10.1177/205031211882292.

[6] B. G. Glaser, A. L. Strauss, The discovery of grounded theory: Strategies for qualitative research, Aldine, 2009. doi:10.4324/9780203793206.

[7] E. A. Schegloff, G. Jefferson, H. Sacks, The preference for self-correction in the organization of repair in conversation, Language 53 (1977) 361–382. doi:10.2307/413107.

[8] E. A. Schegloff, Sequence organization in interaction: A primer in conversation analysis, Cambridge University Press, 2007. doi:10.1017/CBO9780511791208.

[9] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned language models are zero-shot learners, arXiv (2022). URL: https://arxiv.org/abs/2109.01652.

[10] E. Ollion, R. Shen, A. Macanovic, et al., Chatgpt for text annotation? mind the hype!, SocArXiv (2023). URL: https://osf.io/preprints/socarxiv/x58kn.

[11] P. Törnberg, Best practices for text annotation with large language models, arXiv (2024). URL: https://arxiv.org/abs/2402.05129.