

Integrating Large Language Models into Qualitative Methods in Health Services Research: A Proof-of-Concept Study

Lia Chin-Purcell

liacp@stanford.edu

Stanford University School of Medicine

Elena Rosenberg-Carlson

Stanford University School of Medicine

Helene Chokron Garneau

Stanford University School of Medicine

Marzan Hamid

Stanford University School of Medicine

Mark McGovern

Stanford University School of Medicine

Research Article

Keywords: Large language models, natural language processing, qualitative analysis, deductive coding, qualitative methods, health services research, implementation science, substance use disorder, public health, rigor

Posted Date: May 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9726844/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

**Integrating Large Language Models into Qualitative Methods in Health Services Research:
A Proof-of-Concept Study**

Lia Chin-Purcell, MIMS, 0009-0005-3569-2259, Department of Psychiatry and Behavioral
Sciences, Stanford University School of Medicine, Palo Alto, California

Elena Rosenberg-Carlson, MPH, 0009-0008-5654-9728, Department of Psychiatry and
Behavioral Sciences, Stanford University School of Medicine, Palo Alto, California

Hélène Chokron Garneau, PhD, MPH, 0009-0000-8709-6057, Department of Psychiatry and
Behavioral Sciences, Stanford University School of Medicine, Palo Alto, California

Marzan Hamid, BHSc, Department of Psychiatry and Behavioral Sciences, Stanford University
School of Medicine, Palo Alto, California

Mark McGovern, PhD, 0000-0002-9736-573X, Department of Psychiatry and Behavioral
Sciences and Department of Medicine, Stanford University School of Medicine, Palo Alto,
California

Corresponding author:

Lia Chin-Purcell, MIMS

liacp@stanford.edu

Department of Psychiatry and Behavioral Sciences

Stanford University School of Medicine

Palo Alto, California, USA

ORCID iDs:

Lia Chin-Purcell: 0009-0005-3569-2259

Elena Rosenberg-Carlson: 0009-0008-5654-9728

Helene Chokron Garneau: 0009-0000-8709-6057

Marzan Hamid: 0000-0001-7669-8554

Mark McGovern: 0000-0002-9736-573X

ABSTRACT

Background

Qualitative methods are widely used in health services research to derive context-specific insights and depth of understanding. Manual coding, a foundational technique in rigorous qualitative analysis, is highly resource and time-intensive and difficult to scale. This is a particular challenge in health services research, where repeated rounds of interviews are common and rapid turnaround is often required. Natural Language Processing (NLP), specifically Large Language Models (LLMs), have shown potential to enhance efficiency in qualitative analysis. However, there is limited research providing guidance on how to integrate LLMs while maintaining rigor and trustworthiness. In this proof-of-concept study, we propose, apply, and evaluate an NLP-assisted coding method in a health services research setting.

Methods

We analyzed 22 interviews among public health officials, law enforcement, community organizers, and medical professionals at one California county to examine existing substance use service gaps and needs. A primarily deductive codebook was iteratively refined until two coders achieved an inter-coder reliability (ICR) > 0.95 and was applied to the transcripts using ATLAS.ti. We developed an NLP-assisted method that uses a semantic shift algorithm to segment transcripts which are then passed to GPT-4 for code assignment and explanation using the codebook and coding guidelines developed during the manual process. We evaluated the method with a quantitative assessment of agreement between human and NLP-assigned codes, a qualitative and quantitative soundness assessment by two reviewers, and a comparative efficiency analysis.

Results

The NLP-assisted method had moderate agreement with human coding (modified pooled Cohen's Kappa = 0.66), and 71.8% of codes were rated as sound by reviewers. Sound codes were more often observed for high-certainty and straightforward codes, and when text chunks were semantically well defined. The NLP-assisted method had more difficulty with non-linear conversation and entity-dependent codes. Coding time was reduced significantly from ~40 hours for the traditional method to ~1 hour for the NLP-assisted method.

Conclusions

These findings suggest that LLMs can be effectively incorporated into qualitative processes while maintaining rigor if humans are embedded into the process. By maintaining a human-in-the-loop workflow, our methodology allows for researchers to maintain familiarity with the data, define the research question(s) and codebook, and determine if there are results that are not sound. By incorporating LLMs into the coding stage of the process, key limitations of traditional qualitative methods in health services research can be addressed, such as scalability, and resource and time limitations.

KEYWORDS

Large language models, natural language processing, qualitative analysis, deductive coding, qualitative methods, health services research, implementation science, substance use disorder, public health, rigor

BACKGROUND

Qualitative methods are widely used in health services research. These approaches provide key insights about human experiences and contextual factors not otherwise or easily obtained with quantitative methods. In rigorous qualitative methods, coding is a foundational technique. It consists of in-depth data review and the assignment of labels to segments of text to categorize and summarize information relevant to the research question [1]. Codes act as a critical, initial connection between qualitative data and their explanation of meaning [1]. Manual coding is highly resource and time-intensive, often requiring several days to even months to complete, depending on the specific methods used and the volume and complexity of the data. Additionally, this process requires researchers skilled in the qualitative analysis process, both to lead analysis and to apply codes.

This poses particular challenges in health services research because rapid turnaround is often necessary when implementing new programs or treatment to address contextual factors or adapt interventions [2]. Additionally, qualitative methods may be used in health services research where there are rapid changes that existing quantitative datasets do not capture, and time-sensitive information is needed [3]. This is particularly important in implementation science, where insights from key informant interviews often need to be provided to participating organizations quickly to support intervention implementation efforts. Repeated rounds of interviews are common for healthcare organizations, often used to assess needs or barriers to service delivery throughout a program or study. This requires consistent, efficient methods to garner accurate and actionable information from each round of data collection [4]. Building sufficient capacity of skilled qualitative researchers creates an additional challenge, particularly for smaller and resource-limited organizations.

98 Researchers have been exploring the use of Natural Language Processing (NLP), computational
99 techniques for analyzing human language, to expedite and enhance qualitative analyses [5–12].
100 NLP has been shown to be a potentially useful tool to analyze text data and enhance efficiency at
101 various stages of the qualitative analysis process. While earlier NLP techniques showed promise,
102 supervised approaches require large amounts of human-annotated data, and unsupervised
103 approaches rarely captured the specific categories researchers intend to target [9,10,13–19]. The
104 recent advancements of Large Language Models (LLMs), like Generative Pre-trained
105 Transformer (GPT), Claude, and Gemini, offer increased possibilities for deriving robust insights
106 from qualitative data quickly. Because they have a high capacity for reasoning and are flexible to
107 various applications, they show particular promise where other NLP methods may fall short
108 [6,20,21].

109 Recently, there has been an emergence of the use of LLMs to extract insights from qualitative
110 data, and to label data given a codebook to ease the burden of time and cost [5–
111 8,12,21,22]. Researchers have explored how LLMs can be used to generate preliminary
112 emerging themes from qualitative data to inform an inductive coding approach, and how those
113 themes compare to ones generated by humans [5,11,23–26]. Researchers have also explored
114 utilizing LLMs with a deductive approach to assign labels to qualitative data based on a
115 predefined codebook, both with pre-defined excerpts of text [27,28], and using LLMs to identify
116 the sequence of text to which the code applies [5,7,8,29]. Aside from assigning codes,
117 researchers have also suggested that LLMs may be used to identify ambiguous text and resolve
118 disagreements between coders [11,26], or as research assistants [30].

119 Despite this growing body of literature, few studies have explored applications of LLMs in
120 qualitative health services research, a context that presents unique considerations for the

application of LLMs. First, deductive coding approaches are common because research questions are typically tied to established theoretical frameworks or aimed at answering specific questions relevant to service delivery. Second, health services interviews are often greater than 30 minutes in length, producing transcripts too long to pass to an LLM in a single input. Taken together, these elements necessitate a workflow that allows text to be processed in smaller, meaningful segments and then consistently applies a deductively derived codebook.

Existing literature does not adequately address these needs. The majority of aforementioned studies using LLMs for qualitative analysis did not use interview data at all, instead applying LLMs to structured text such as social media posts, forum comments, historical passages, or survey responses that require no segmentation [5–8,11]. Social media posts, journalistic text, and other public-facing content appear particularly well represented in LLM training data, meaning LLMs may perform better on these data compared to real-world interview transcripts. Among studies that did use interview transcripts, the text was either passed to the LLM as a whole or divided based on context window constraints [12,22,24,26,29]. Some studies were conceptual and did not apply LLMs to empirical data [22,25,30]. Also, few studies have evaluated an LLM-assisted coding workflow and output using qualitative rigor criteria [31,32]. The primary goal of this study is to help address these gaps by proposing and evaluating a new method for the application of LLMs in qualitative health services research.

The use of LLMs in qualitative research also raises important questions about rigor and trustworthiness. In qualitative research, the human researcher is traditionally considered the analytic instrument. The ontological and epistemological principles of traditional qualitative research as well as several techniques to establish rigor and trustworthiness in qualitative research, such as prolonged engagement with the data and memoing, are predicated on a human

analytic process [33]. How LLMs can be used in qualitative analysis is very early in its research developmental stage, and there is no standard of practice for how LLMs are best used in the analysis process. Because of this, guidance and methodology are needed to determine how best to integrate LLMs while maintaining rigor and trustworthiness of results. Additionally, methodologies are needed in which researchers are engaged with the text and can review results for credibility. Therefore, a secondary goal of this study is to demonstrate how qualitative rigor and trustworthiness can be preserved when LLMs are integrated into the analytic process.

METHODS

Dataset

The dataset comprises 22 interviews among public health officials, law enforcement, community organizers, and medical professionals in one county in the State of California. The interviews are semi-structured and were conducted with the goal of gaining insight into the existing substance use disorder (SUD) service network and gaps and needs within the county. Participants were recruited using a snowball sampling method where a small sample of initial participants was identified, and additional participants were identified by asking who else might have information on SUD services in the county. The interviews were conducted via videoconference over the course of 10 weeks in 2023 and were recorded with the consent of the interviewees. The interviews were transcribed by a professional transcription service which also removed the names of the interviewees and assigned participant IDs. The recordings and transcriptions were securely stored on the Stanford Medicine Box. The interviews are on average 8,301.6 tokens, 12.8 transcribed pages (12pt font single spaced), and about an hour in length.

Traditional Coding Method

Our traditional coding method involved 2 human coders, 1 of whom is the first author (LCP), and 1 arbiter, an experienced qualitative researcher who is the second author of this paper (ERC). A primarily deductive coding method was used, with an initial codebook and coding guidelines developed by the arbiter based on the project objectives and interview guide and then refined in collaboration with the 2 coders via an iterative process. The interview guide was developed for this study and is provided as Additional File 1 – Interview Guide. The codebook and accompanying coding guidelines were independently applied to selected interview transcripts using ATLAS.ti (version 25.0.2; ATLAS.ti Scientific Software Development GmbH) [34]. The codebook and coding guidelines were revised through multiple rounds of discussion and reliability testing until high intercoder agreement was achieved (Cohen’s kappa > 0.95) [35]. After intercoder reliability was established, the 2 coders each independently coded 11 transcripts using the finalized codebook and coding guidelines.

NLP Methods Tested

Several NLP methods were identified and evaluated for viability prior to developing the final workflow. Details of this initial evaluation are provided in Additional File 2 – NLP Methods Tested. Ultimately, a semantic shift segmentation algorithm was developed for segmenting long interview transcripts into meaningful units, as it identifies topic changes using embedding-based similarity rather than relying on paragraph boundaries or fixed character windows [36].

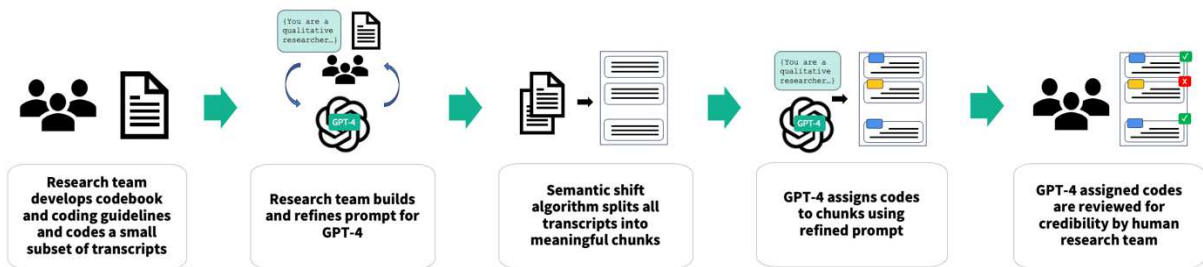
Finalized NLP-Assisted Method Workflow

Our approach outlined here incorporates the use of LLMs in a way that is iterative and collaborative with researchers. We aim to preserve human engagement in and leadership of the

coding process while utilizing LLMs to enhance efficiency and make the process more scalable for larger amounts of data. Also, we consider the limitations of LLMs, such as hallucinations, context window issues, and non-specialization, and provide a workflow where these limitations are mitigated. The workflow is primarily model agnostic, with the hope that it continues to be useful as others consider how LLMs can be incorporated into their coding process.

We demonstrate our approach in a proof-of-concept format where we show its application to our dataset of 22 interview transcripts from one qualitative study. We present a developed “NLP-assisted” method which incorporates LLMs into the workflow as well as other NLP methods including SentenceTransformer embeddings. The overall structure of the proposed method is to split the transcripts up into meaningful “chunks” using a semantic shift algorithm, which are then passed to GPT-4 for coding, along with the codebook and coding guidelines (see Figure 1). We explain the details of the proposed method below.

Figure 1. NLP-Assisted Method Workflow



Step 1. Research team develops codebook and coding guidelines and codes a small subset of transcripts

The first step of our method involved a human research team developing a codebook and coding guidelines to code a subset of randomly selected transcripts. In this step, the human research

team first generated a codebook and coding guidelines based on the interview guide and research question. Then, they iterated on the codebook and coding guidelines until consensus is reached amongst the coders.

By coding a subset of transcripts, the human research team preserved familiarity and engagement with the data, which are key aspects of qualitative research [1]. They were also better equipped to review the codes applied by the model for credibility.

Step 2. Research team builds and refines prompt for GPT-4

In the next step, the codebook and coding guidelines were used to develop a prompt for the LLM. When designing the prompt, researchers have found that crafting a useful prompt to the LLM is key to obtaining good results, specifically assigning the model a role and specifying what the output should look like [26]. We began with human-identified text excerpts rather than automated segmented text, allowing us to iterate on prompt wording without the added complexity of chunking issues.

In our case, our prompt was structured by first stating the task for the LLM: “*You are a qualitative researcher coding interview transcripts. You need to classify the following text segment.*” Next, we proceeded with the codebook, the current chunk of text to code, the previous chunk of text (if available) for context, and the coding guidelines developed in the previous step (see Figure 2). The LLM is asked to provide its response in structured, machine-readable format (JSON) format, and for each code assigned we asked for a certainty indicator (high, medium, low) and an explanation for why the code was assigned.

228 Figure 2. Prompt Structure Overview

```
229 You are a qualitative researcher coding interview transcripts. You need to classify the following text segment
according to this codebook. Please analyze this text carefully and assign a code or codes according to the codebook
descriptions.

CODEBOOK:
{codebook text}

TEXT TO CODE:
{current chunk of text}

Previous Context (if available):
{previous chunk of text}

{coding guidelines}

Multiple Coding:
- You may apply multiple codes to the text segment if it is highly aligned with the codebook description for multiple
codes.

Please provide your response in JSON format with these fields:
{{
  "codes": [
    {{
      "code": "string", # The code from the codebook
      "explanation": "string", # Explanation for why this code was selected
      "certainty": "high|medium|low" # Confidence level: "high" if very confident, "medium" if somewhat confident,
      "low" if uncertain
    }}
  ]
}}
```

230

231 The prompt was applied to each chunk of text identified by the human coding. The results of this
232 coding are then compared to the codes applied by the humans for the purpose of iterating on the
233 prompt to enhance the LLM's ability to apply codes as intended by the research team. For
234 example, we initially observed an under-coding of parent codes compared to the manual codes
235 applied. To address this, we included the term "(Parent Code)" in the code name for those codes,
236 which resulted in the LLM more frequently making appropriate use of the parent codes. We
237 repeated this process until satisfactory agreement was reached between the LLM coding and the
238 human coding, and the human research team had sufficient confidence in the LLM's ability to
239 appropriately apply the codes.

240 **Prompt Iteration with Semantic Shift Chunks**

241 The LLM prompt was developed iteratively across 5 transcripts (22% of the dataset), with
242 refinements including reducing emphasis on interviewer questions, improving guidance for
243 context-specific codes, and determining the optimal number of codes to request per chunk. Full
244 details of the prompt iteration process are provided in Additional File 3 – Prompt Iteration, and
245 the finalized codebook and coding guidelines in the prompt are provided in Additional File 4 –
246 Codebook.

247 **Step 3. Semantic shift algorithm splits all transcripts into meaningful chunks**

248 Once the prompt was finalized, the semantic shift algorithm was applied to all transcripts to split
249 them into meaningful chunks. In our implementation, we used the SentenceTransformer model
250 all-mpnet-base-v2 to generate vector embeddings of each paragraph [37]. Our transcripts were
251 split into paragraphs by our transcription service that were on average 3.5 sentences in length.
252 We added two question-detection rules to ensure that interviewer questions and their responses
253 were kept together in the same chunk, and that new questions initiated a new chunk.

254 We made 5 additional rounds of iteration on the prompt at this stage, before making the
255 determination that the method displayed sufficient ability to code the chunks as intended by the
256 research team and proceeding with the whole corpus of transcripts.

257 **Step 4. GPT-4 assigns codes to chunks using refined prompt**

258 Once transcripts were segmented, the refined prompt was applied to each chunk to assign codes.
259 For transcripts that contain Protected Health Information (PHI) or other sensitive data, this step
260 should be completed using a secure or local instance of an LLM to preserve privacy and security.

An API was also used within a Python script to automate calls to the LLM across all chunks, rather than manually submitting them, to create a streamlined and efficient process.

In our case, we used a secure instance of GPT-4 hosted by an academic medical center for all our LLM calls. This academic medical center has deployed a secure, private instance of GPT-4 referred to as Secure GPT. Unlike public-facing LLMs like ChatGPT, Secure GPT is configured to handle Protected Health Information (PHI) safely and securely using a private infrastructure and closed network access [38]. The result was each chunk receiving 2–3 codes on average, along with an explanation and certainty indicator for each code (see Figure 3). Table 1 shows the distribution of chunks by number of codes assigned across all transcripts. Because the output was stored in JSON, coded chunks could then be organized by code or parent code for downstream analysis.

Figure 3. Example coded chunk

Chunk 15

Interviewer: Can you identify any key connectors or bridge organizations that facilitate communication between the organizations that you work closely with, like you mentioned? Is it mostly direct one-to-one?

Interviewee: It's mostly direct, or it's the organizations I just referenced. I should include what was the [redacted]. It recently was rebranded or retitled the [redacted]. Again, a post-disposition county government agency that provides services to clients who are returning to our community. It's often we're referring them to other government agencies, and then they're connecting them either with direct providers that are internal in their organization or through contracts they have.

I would say, for us, the other government agencies like the probation department, behavioral health and the [redacted]. Our direct point of contact, and then they know the providers. They either have the providers in-house or they have the providers through contract.

Similarity Score:

0.443

Classification (2 Codes Applied):

Code 1: Organization Collaboration: Collaborative Relationships

Code 2: Network Characteristics: Network Connectors

Code 1: Organization Collaboration: Collaborative Relationships

Explanation: The interviewee discusses specific entities (e.g., [redacted], probation department, behavioral health) with which their organization collaborates, including the roles these entities play in connecting clients to services. This aligns closely with the description of 'Collaborative Relationships' as it details the operational collaboration between their organization and these external agencies.

Certainty: HIGH

Code 2: Network Characteristics: Network Connectors

Explanation: The interviewee identifies the ([redacted] (recently rebranded as the [redacted])) as a facilitator or key connector between organizations. This aligns with the 'Network Connectors' code, which captures institutions that act as bridges within the County's service network.

Certainty: HIGH

Table 1. Distribution of Chunks by Number of Codes

Number of Codes	Chunks, n (%)
1	110 (13.8%)
2	399 (49.9%)
3	218 (27.3%)
4	59 (7.4%)
5	10 (1.3%)
6	3 (0.4%)

Step 5. GPT-4 assigned codes are reviewed for credibility by human research team

As a final step, the research team reviews the final codes applied by the LLM for credibility. The goal of this step is similar to an arbiter’s review process to ensure that codes are consistently applied appropriately based on alignment between the chosen code(s) for a given chunk and the coding frame and guidelines. During this step, the research team may correct certain codes and note patterns or concerns. This step is crucial because the LLM may occasionally hallucinate, and the human research team would be equipped to audit the codes applied due to their experience with applying the codebook without the use of an LLM. Researchers doing this step benefit from having coded a subset of transcripts in Step 1, as that familiarity with the codebook helps them to assess the model's output. In our case, the research team reviewed a subset of transcripts LLM-assigned codes for evaluation purposes detailed below.

Evaluation of the NLP-Assisted Method

To critically evaluate our proposed method, we conducted both a quantitative assessment of agreement between human coders and the NLP-assisted method, and a qualitative assessment of the soundness of the codes assigned by the NLP-assisted method. We conducted both assessments to have a direct comparison against human coders while also recognizing multiple sound perspectives may exist regarding which code(s) to apply, consistent with the constructivist lens underpinning most traditional qualitative research. An efficiency analysis was also conducted to compare time required to apply codes using the NLP-assisted method versus the manual method.

Quantitative Evaluation of Coding Agreement

To assess quantitative agreement between the human coders and the NLP-assisted method, we computed a modified version of Cohen's kappa. The semantic shift algorithm segments the entire transcript and GPT-4 assigns a code to every chunk. In contrast, human coders were instructed to code only relevant content, exercising judgment about whether to include interviewer questions based on context. As a result, the NLP-assisted method inevitably produced sections with LLM-assigned codes that received no corresponding human code. Therefore, we computed a modified kappa which measures recall-oriented agreement, that is, how well the LLM captures what humans coded, rather than general inter-coder reliability. We also report the total proportion of rows and characters that were LLM-only coded across transcripts. To compute the modified kappa, we identified all start and end character positions from both LLM and human codes. Each row in the dataset corresponded to a section with start and end positions, and the corresponding semicolon-separated LLM codes and human codes. Rows where there were only LLM codes

were removed. Post-processing filters were applied to remove duplicate rows and very short sections.

Because GPT-4 code assignment could contain multiple codes per section, GPT-4 codes were split into their individual codes and directly compared to the human code. If the human code matched any of the GPT-assigned codes, the section was treated as an agreement. If the human code did not match any LLM code, the section was treated as a disagreement. 2 variants of the modified Cohen's kappa were computed for each transcript, as well as a pooled Cohen's kappa. The first variant of Cohen's kappa retained all codes, and the second had the "Miscellaneous" code removed, a codebook category used to capture content deemed relevant but not fitting any other code. The pooled Cohen's kappa is computed by aggregating all sections across all transcripts and computing the metrics for all chunks together. Evaluation was performed on all transcripts. Credibility of the NLP-assisted method findings was further addressed through the soundness assessment reported below.

Qualitative and Quantitative Evaluation of Coding Soundness

In addition to the quantitative assessment of agreement, a quantitative and qualitative assessment of reasonableness or soundness of the GPT-4 codes assigned to segments was conducted. This assessment accounts for the possibility that the codes applied by GPT-4 may be different than the ones applied by humans, while still being reasonable given the codebook, coding guidelines, and provided explanation.

A subset of 5 transcripts was selected to ensure representation of different interviewee roles across the county SUD service network. The qualitative assessment was completed by 2 members of the research team who are also the first and second author of the paper. Both

reviewers assessed the first 3 codes applied to each chunk of the 5 transcripts and evaluated the code for 1) soundness with respect to the codebook, 2) soundness with respect to the coding guidelines, and 3) soundness of the provided explanation. A binary rating of “sound” or “not sound” was applied to each code based on these requirements, as well as a short 1-2 sentence justification. After both reviewers independently rated each code applied by the NLP-assisted method, we reviewed our ratings together to come to a consensus and discussed patterns observed. We then calculated overall soundness as the percentage of sound codes among all reviewed codes and computed this metric separately for high certainty and medium certainty codes.

Comparative Efficiency Analysis

We compared the amount of time taken for the NLP-assisted method to complete its coding of the 22 interviews compared to the amount of time it took for a human research team to code the interviews. We measured the time spent applying the codebook to the interviews rather than the time spent developing the codebook and coding guidelines because these outputs are also used in the NLP-assisted approach. Since coding is typically the most time-intensive step in qualitative analysis, we focused on this phase to have the greatest potential for increase of efficiency while preserving key aspects of the traditional coding process.

A multi-threaded approach with back-off was implemented to make API calls to Secure GPT to increase efficiency by handling multiple requests while also avoiding rate-limiting errors. Total processing time for the manual method and the NLP-assisted method were recorded and compared.

RESULTS

Quantitative Evaluation of Coding Agreement

Agreement between human coders and the NLP-assisted method was assessed using 2 variants of the modified Cohen's kappa. The modified kappa reflects how well the NLP-assisted method captured what humans coded, rather than a conventional inter-coder reliability metric. Generally, recall-oriented agreement between human coders and the NLP-assisted method was moderate to substantial. With all codes included, the modified Cohen's kappa ranged from 0.45 to 0.85 (mean = 0.63; median = 0.63, IQR = 0.57–0.69) with a pooled agreement across transcripts of $\kappa = 0.63$. When the miscellaneous code was excluded, agreement improved slightly with kappa values ranging from 0.54 to 0.85 (mean = 0.67; median = 0.64, IQR = 0.61–0.70) with a pooled agreement of $\kappa = 0.66$. Across all transcripts, LLM-only coded rows represented 46.8% of total rows and 24.2% of total characters, reflecting the method's design of coding all content.

Qualitative and Quantitative Evaluation of Coding Soundness

Across the 5 transcripts reviewed, 71.8% (295 / 411) of the codes applied by the NLP-assisted method were rated as sound based on the codebook, coding guidelines, and provided explanation. Codes that were rated sound accurately reflected the definitions in the codebook and appropriately cued into the entity being discussed in a given chunk. The explanation provided by GPT-4 was often grounded in the text and aligned with what the reviewers would have used to assign a code. Additionally, explanations often acknowledged interpretation or implications for codes assigned that were less explicit, which aided in review. If a code was high certainty, it was more likely to be sound (82%, 244/296); however, there were medium certainty codes that were

also rated as sound (45%, 51/113), and not all high certainty codes were rated as sound. The one low certainty code reviewed was not sound.

Some codes assigned that were more consistently sound were more straightforward and discrete codes (formatted as Parent Code: Code) such as “Organization Characteristics: Organization Resources” (90.91%, 10/11) and “Organization Characteristics: Populations of Focus” (87.50%, 21/24). On the other hand, codes with lower soundness percentages often were more complex and relational, for example “Organization Collaboration: Collaboration Facilitators” (54.5%, 6/11) and “Network Characteristics: Communication Patterns” (55.6%, 5/9). Generally, when the chunks of text were semantically well-defined, included a question from the interview guide that mapped directly onto a code, and the interviewee response was directly related to that question, the codes applied by the NLP-assisted method were found to be sound.

One common reason codes were rated as not sound was because another code unequivocally was more appropriate for the given chunk. For example, a chunk included the interviewer asking about funding a resource allocation. The response included discussion of budgets and allocation at the county and discussed opioid settlement funds used for opioid and substance use-related services. GPT-4 applied both the “Organization Characteristics: Organization Resources” code and the “Organization Characteristics: Opioid Use Services” code, and in this case, we reasoned that “Organization Characteristics: Organization Resources” was unequivocally more appropriate. Therefore, the additional application of “Organization Characteristics: Opioid Use Services” was not sound. Relatedly, we found that generally GPT-4 applied more codes than humans overall, and the “Opioid Use Services” code was particularly over-coded, as well as “Community Needs.”

418 Some of the coding assignments by GPT-4 that were not sound required an understanding of the
419 specific entity being discussed. For example, there were instances in which a vague pronoun
420 reference such as “we” was used to refer to the interviewee’s organization, and GPT-4 assigned a
421 code as if they were referring to the service network broadly.

422 Another reason codes were rated as not sound was that given the explanation, GPT-4 seemed to
423 be referencing the previous chunk too heavily and assigning a code that would be appropriate for
424 the previous chunk, rather than using the previous chunk only for context. We also observed that
425 sometimes, due to irregular conversation, there were chunks that were not semantically well-
426 defined such as a back-and-forth end of conversation, or only including the interviewer question.
427 Here, “Miscellaneous” might be the most appropriate code, however GPT-4 might assign a code
428 based on the previous chunk or assign a code based on the question asked. Overall, we observed
429 that the “Miscellaneous” code was under-coded.

430 Occasionally, the model hallucinated a code or slightly changed the code name. For example,
431 “Network Characteristics: Network Challenges” became “Network Characteristics: Network
432 Gaps,” “Network Characteristics: Network Resources” was made up (our codebook included
433 “Organization Characteristics: Organization Resources”, but not “Network Characteristics:
434 Network Resources”), “Organization Collaboration” and “Organization Characteristics” lost their
435 parent code indicators, and sometimes the model mixed up parent and child code relations. Once,
436 the model did not assign a code. There were also a few more benign errors, like making
437 populations in “Organization Characteristics: Populations of Focus” not plural or removing the
438 parent code designation (“Organization Characteristics (Parent Code)” becoming “Organization
439 Characteristics”). For these more obvious errors, we conducted post-processing to align the code
440 names.

Comparative Efficiency Analysis

The total time required for the NLP-assisted method was 57 minutes and 40 seconds. The minimum time it took to code a transcript was 1 minute and 34 seconds, and the maximum time was 4 minutes and 15 seconds. In contrast, human coders took 39 hours and 15 minutes to code the same interviews. On average, the NLP-assisted method took 2 minutes and 30 seconds per transcript, compared with 107 minutes and 2 seconds per transcript for humans.

DISCUSSION

Overall, our findings demonstrated moderate-to-substantial recall-oriented agreement between human coding and the NLP-assisted method ($\kappa = 0.66$), and that the NLP-assisted method assigned codes that were deemed “sound” a high portion of the time (71.8%). However, we also observed some limitations of the NLP-assisted coding method, namely the semantic shift algorithm struggling to segment the transcript into meaningful segments when conversational flow was very non-linear and unstructured, and the GPT-4 tending to over-code some codes and occasionally hallucinating (providing a code or explanation that was not logically related to the context provided).

Strengths of the NLP-assisted Method

Efficiency and Consistency

The NLP-assisted approach demonstrated a significant reduction in coding time compared to the human coders, indicating a gain in efficiency. Methods utilizing LLMs can code many transcripts quickly, which is especially helpful for projects with large amounts of data. Scaling up to even hundreds of transcripts might be feasible for a small research team when it otherwise would not be possible due to time and capacity constraints. In addition to efficiency gains, the NLP-assisted

approach applied codes relatively consistently across all transcripts, reducing the variability that can arise between human coders or over time.

It is important to note that there are time and expertise requirements to implement the NLP-assisted approach. Once an NLP-assisted workflow has been established, it would be relatively easy to adapt it to new research questions and new codebooks, enhancing flexibility and efficiency in the process. While this workflow requires an initial investment in infrastructure setup, subsequent rounds of interviews can be analyzed more efficiently, making it well-suited to health services organizations that conduct repeated interviews (see Table 2).

Table 2. Setup Activities vs. Repeated Activities

One-time setup	Per-round of qualitative interviews
Develop codebook and coding guidelines	Run automated semantic shift
Humans code small subset of transcripts	Run automated coding
Develop prompts for LLMs through iteration	Humans review automated coding
Adapt semantic shift method for transcripts	

Sound Coding

Based on our quantitative soundness assessment, we found that there was a relatively high proportion, 71.8% (295/411), of codes that were applied reasonably given the codebook and coding guidelines. These findings are broadly consistent with other studies utilizing LLMs for deductive coding, which have reported fair to substantial kappa values [5,6,8] and accuracy rates ranging from approximately 66–90% [5,7], though performance varies by data type and task complexity. This shows promise for GPT-4’s ability to appropriately apply codes when the prompt is adapted for a particular coding task. While oversight of coding is still required with an

NLP-assisted method, the output provides a structured set of codes that researchers can efficiently review and use for further analysis.

Challenges of the NLP-assisted Method

Semantic Shift Algorithm Limitations

The semantic shift algorithm struggled to segment meaningfully when the text was less structured. We found that including a series of rules for detecting questions and answers, clarifications, short responses, and using a more robust Sentence Transformer model for embeddings improved performance. However, these rules are hard coded into the algorithm and ultimately are subject to the nuances of natural language and the non-linear nature of human conversation.

When the algorithm struggled to segment meaningfully and provide appropriate context, GPT-4 tended to struggle to assign appropriate code(s). Many incorrect coding assignments resulted from not cueing into the appropriate entity. In some cases, to make a sound code application it was essential to understand context that may have been embedded throughout the interview. In other cases, the appropriate context was included, but understanding the entity being discussed was difficult or interpretive, which humans also struggled with. This suggests that algorithms used to segment transcripts may work better for structured interviews. If interviews are more free-form, obtaining segments during transcription may be more efficient and yield better results.

Over-coding

Based on our qualitative review of the GPT-4 codes, we found that GPT-4 tended to apply too many codes to the chunks as reflected in the 46.8% of rows that were only coded by the LLM. Despite instructions in the prompt to only apply secondary codes if they were highly aligned

with the codebook, GPT-4 sometimes assigned lower certainty second or third codes to chunks where reviewers determined the first code was unequivocally more appropriate. This tendency may show an over confidence in GPT-4's responses, which has been demonstrated for LLMs in the literature [26].

The over-coding of the "Organization Characteristics: Opioid Use Services" and "Community Needs" codes may be due to the specificity of the topic; the goal of the project was to explore the landscape of substance use services, so words related to substance and opioid use are common throughout the interviews. GPT-4 seemed to code the "Organization Characteristics: Opioid Use Services" code when substance use related words appeared, but it was aimed at capturing a specific part of the interviews where the interviewees discuss the services they provide. Similarly, the "Community Needs" code was aimed at capturing information about communities that face barriers to accessing services, which was a specific question in the interview guide. GPT-4 seemed to apply this code whenever a population or community was mentioned (where "Organization Characteristics: Populations of Focus" may have been more appropriate), or an area where there may be a gap in services (where "Network Characteristics: Network Challenges" may have been more appropriate). This reliance on keywords rather than contextual meaning is consistent with the finding that LLMs perform better on context-independent than context-dependent coding tasks [8]. This is also an example of a potential shortcoming of using a general model such as an out of the box large language model like GPT-4 for a relatively specific area of focus. One way that this could be addressed is by fine-tuning the model with material from the specific context before assigning codes.

The Role of LLMs in Qualitative Research

Many qualitative researchers argue that the interpretive, meaning-making process of qualitative research is and should remain a distinctly and exclusively human practice [39]. We agree that humans have an essential role to play in qualitative analysis, and do not propose the uncritical replacement of traditionally human analytic activities with LLMs. We likewise believe that traditional, reflexive “Big Q” Qualitative analysis approaches are uniquely valuable in their ability to produce in-depth, nuanced understanding of people and social processes. We do not suggest that our methodology should be considered as equivalent, nor that an automated process should be applied uncritically in circumstances where this is the goal. However, in fields such as health services research where obtaining results quickly is a high priority, NLP-assisted approaches may be useful. Given the accelerating use of LLMs in qualitative research, developing methodologies that preserve trustworthiness is both timely and necessary.

Limitations

While this proof-of-concept allowed us to understand what some of the strengths and challenges are to using an NLP-assisted coding method, it should be noted that the dataset used is comprised of a relatively small set (22) of interviews, in a relatively specific subject area. The unique dataset may limit the generalizability of the proposed method’s results to other qualitative research projects.

Additionally, we report a modified Cohen's kappa to provide a quantitative assessment of the agreement between human coders and the NLP-assisted coding approach. Because this metric is modified, it may not be directly comparable to standard measures of agreement, for example between two human coders. We modified Cohen's kappa to account for the differences in coding application: the NLP-assisted method codes all segments of the transcript and can apply multiple codes per section, while human coders selectively code text and are more likely to double code

overlapping sections rather than assign multiple codes to the same section. Our modified version of Cohen's kappa captures agreement and disagreement between human and the NLP-assisted coding, but also avoids large penalties for expected differences, such as the NLP-assisted method coding more text than the humans, or humans not applying all codes to a particular part of text that the NLP-assisted method has chosen.

CONCLUSIONS

Our results suggest that LLMs may be able to be effectively utilized for qualitative coding to increase efficiency of qualitative analysis processes while maintaining trustworthiness. The importance of maintaining human oversight and engagement in LLM-assisted coding has been emphasized in the broader literature, with researchers recommending iterative prompt refinement, human oversight, and review of outputs for credibility [6,26,31,32]. Low-resourced settings where there may not be enough analysts or time to code and analyze qualitative data could benefit particularly from this workflow, whereby using this methodology, they could ask research questions that they wouldn't have otherwise been able to pursue. Additionally, in health services research settings, where repeated rounds of interviews are common, a rigorous workflow that can be efficiently repeated is particularly valuable.

Future research should test the method with other types of data and larger pools of data to assess applicability for other settings and research questions. It would also be useful to conduct usability testing of the method itself and how researchers find its ease of use and how it can be improved. Also, there may be several ways that NLP and LLMs specifically can arrive at themes and insights using more of an inductive approach, or may deviate from the traditional qualitative process more, for example not applying codes to transcripts per se but rather retrieving relevant information or summarization. For example, Retrieval Augmented Generation (RAG), which

grounds LLM outputs in retrieved passages from the dataset, has been explored to assist with the analysis of interview transcripts. This method moves away from code application toward retrieving relevant information from the data and generating summaries [40].

New LLMs and associated methodologies are being developed quickly, and best practices and strategies change frequently. As new models are released, new methodologies may work better or worse depending on the strengths of the model. Currently, using LLMs within Computer-Assisted Qualitative Data Analysis Software (CAQDAS) such as ATLAS.ti often lacks transparency and can be difficult to modify or tune to researchers' needs. However, these processes could be improved to enable smoother integration of LLMs into traditional coding workflows. As the field continues to develop methodology using LLMs to assist with qualitative research, this proof-of-concept offers a human-centered workflow grounded in qualitative rigor. This methodology has the potential to expand research possibilities for qualitative health services researchers under time and resource constraints.

LIST OF ABBREVIATIONS

NLP: Natural Language Processing

LLM: Large Language Model

GPT: Generative Pre-trained Transformer

SUD: Substance Use Disorder

IRB: Institutional Review Board

JSON: JavaScript Object Notation

PHI: Protected Health Information

592 API: Application Programming Interface

593 IQR: Interquartile Range

594 RAG: Retrieval Augmented Generation

595 CAQDAS: Computer-Assisted Qualitative Data Analysis Software

596 **DECLARATIONS**

597 **Ethics Approval and Consent to Participate**

598 This study was reviewed and approved by the Stanford University Institutional Review Board
599 (IRB) under Protocol #72775, titled "Landscape of Services for Individuals Who Use Opioids in
600 Santa Clara County" (Approval Period: 04/25/2025 – ongoing; Review Type: Exempt Human
601 Subjects Research). This study was conducted in accordance with the Helsinki Declaration. All
602 participants provided verbal informed consent prior to participating in interviews, which were
603 conducted via videoconference. Participation was voluntary, and participants were informed of
604 their right to withdraw at any time without consequence. No protected health information (PHI)
605 was collected. Interview recordings were stored on an encrypted, access-controlled platform
606 (Stanford Medicine Box) and deleted upon completion of transcription. All transcripts were de-
607 identified by removing participant names and assigning sequential participant IDs prior to
608 analysis.

609 **Consent for publication**

610 Not applicable. This article does not contain data from any individual person in any identifiable
611 form.

612 **Availability of data and materials**

613 Raw interview data are not publicly available to protect the privacy of participants and in
614 accordance with IRB restrictions. The code used to implement the NLP-assisted method
615 described in this article is available in the following repository [41]:

616 Project name: NLP-Assisted Qualitative Coding Method

617 Project home page: https://github.com/lisasop/nlp_assisted_method

618 Archived version: <https://doi.org/10.5281/zenodo.19896562>

619 Operating system(s): Platform independent

620 Programming language: Python

621 License: MIT

622 Any restrictions to use by non-academics: None

623 **Competing interests**

624 The authors declare that they have no competing interests.

625 **Funding**

626 This work was supported by the National Institute on Drug Abuse (NIDA), National Institutes of
627 Health (NIH), under Award Number P50DA054072 (Center for Dissemination &
628 Implementation At Stanford [C-DIAS]; PI: McGovern). The content is solely the responsibility
629 of the authors and does not represent the official views of NIDA or NIH.

630 **Authors' contributions**

631 LCP: Conceptualization, Methodology, Data curation, Formal analysis, Software, Writing —
632 original draft, Writing — review and editing. ERC: Conceptualization, Methodology, Formal

analysis, Writing — review and editing. HCG: Conceptualization, Supervision, Writing — review and editing. MM: Supervision, Writing — review and editing. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank the participants of the interviews that comprised the dataset used in this project. The authors also thank Briana Patrick and Shubhi Sharma for their contributions to the manual coding process. Additionally, the authors thank members of the Center for Dissemination & Implementation At Stanford (C-DIAS) Research Project 1 (PI: Hendricks Brown) for their support of the qualitative study.

REFERENCES

1. Charmaz K. *Constructing Grounded Theory*. 2nd ed. Sage Publications; 2014.
2. Kowalski CP, Nevedal AL, Finley EP, et al. Planning for and Assessing Rigor in Rapid Qualitative Analysis (PARRQA): a consensus-based framework for designing, conducting, and reporting. *Implement Sci IS*. 2024;19:71. doi:10.1186/s13012-024-01397-1
3. Weiner BJ, Amick HR, Lund JL, Lee SYD, Hoff TJ. Use of qualitative methods in published health services and management research: a 10-year review. *Med Care Res Rev MCRR*. 2011;68(1):3-33. doi:10.1177/1077558710372810
4. Calman L, Brunton L, Molassiotis A. Developing longitudinal qualitative designs: lessons learned and recommendations for health services research. *BMC Med Res Methodol*. 2013;13:14. doi:10.1186/1471-2288-13-14

- 653 5. Bijker R, Merkouris SS, Dowling NA, Rodda SN. ChatGPT for Automated Qualitative
654 Research: Content Analysis. *J Med Internet Res*. 2024;26(1):e59050. doi:10.2196/59050
- 655 6. Dunivin ZO. Scaling hermeneutics: a guide to qualitative coding with LLMs for reflexive
656 content analysis. *EPJ Data Sci*. 2025;14(1):1. doi:10.1140/epjds/s13688-025-00548-8
- 657 7. Katz A, Wei S, Nanda G, Brinton C, Ohland M. Exploring the Efficacy of ChatGPT in
658 Analyzing Student Teamwork Feedback with an Existing Taxonomy. *arXiv*. Preprint posted
659 online May 9, 2023:arXiv:2305.11882. doi:10.48550/arXiv.2305.11882
- 660 8. Xiao Z, Yuan X, Liao QV, Abdelghani R, Oudeyer PY. Supporting Qualitative Analysis
661 with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In:
662 *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*.
663 IUI '23 Companion. Association for Computing Machinery; 2023:75-78.
664 doi:10.1145/3581754.3584136
- 665 9. McHugh D, Shaw S, Moore TR, Ye LZ, Romero-Masters P, Halverson R. Uncovering
666 themes in personalized learning: Using natural language processing to analyze school
667 interviews. *J Res Technol Educ*. 2020;52(3):391-402. doi:10.1080/15391523.2020.1752337
- 668 10. Badal VD, Graham SA, Depp CA, et al. Prediction of Loneliness in Older Adults Using
669 Natural Language Processing: Exploring Sex Differences in Speech. *Am J Geriatr*
670 *Psychiatry Off J Am Assoc Geriatr Psychiatry*. 2021;29(8):853-866.
671 doi:10.1016/j.jagp.2020.09.009

- 672 11. Chen NC, Drouhard M, Kocielnik R, Suh J, Aragon CR. Using Machine Learning to Support
673 Qualitative Coding in Social Science: Shifting the Focus to Ambiguity. *ACM Trans Interact*
674 *Intell Syst.* 2018;8(2):9:1-9:20. doi:10.1145/3185515
- 675 12. Ashwin J, Chhabra A, Rao V. Using Large Language Models for Qualitative Analysis can
676 Introduce Serious Bias. *Sociol Methods Res.* Published online May 27,
677 2025:00491241251338246. doi:10.1177/00491241251338246
- 678 13. Leeson W, Resnick A, Alexander D, Rovers J. Natural Language Processing (NLP) in
679 Qualitative Public Health Research: A Proof of Concept Study. *Int J Qual Methods.*
680 2019;18:160940691988702. doi:10.1177/1609406919887021
- 681 14. Bastiaansen JAJ, Veldhuizen EE, De Schepper K, Scheepers FE. Experiences of Siblings of
682 Children With Neurodevelopmental Disorders: Comparing Qualitative Analysis and
683 Machine Learning to Study Narratives. *Front Psychiatry.* 2022;13:719598.
684 doi:10.3389/fpsyt.2022.719598
- 685 15. Guetterman TC, Chang T, DeJonckheere M, Basu T, Scruggs E, Vydiswaran VGV.
686 Augmenting Qualitative Text Analysis with Natural Language Processing: Methodological
687 Study. *J Med Internet Res.* 2018;20(6):e231. doi:10.2196/jmir.9702
- 688 16. Lennon RP, Fraleigh R, Van Scoy LJ, et al. Developing and testing an automated qualitative
689 assistant (AQUA) to support qualitative analysis. *Fam Med Community Health.*
690 2021;9(Suppl 1):e001287. doi:10.1136/fmch-2021-001287

- 691 17. Odden TOB, Tyseng H, Mjaaland JT, Kreutzer MF, Malthe-Sørenssen A. Using text
692 embeddings for deductive qualitative research at scale in physics education. *Phys Rev Phys*
693 *Educ Res.* 2024;20(2):020151. doi:10.1103/PhysRevPhysEducRes.20.020151
- 694 18. Palacios G, Noreña A, Londero A. Assessing the Heterogeneity of Complaints Related to
695 Tinnitus and Hyperacusis from an Unsupervised Machine Learning Approach: An
696 Exploratory Study. *Audiol Neurotol.* 2020;25(4):174-189. doi:10.1159/000504741
- 697 19. Criss S, Nguyen TT, Michaels EK, et al. Solidarity and strife after the Atlanta spa shootings:
698 A mixed methods study characterizing Twitter discussions by qualitative analysis and
699 machine learning. *Front Public Health.* 2023;11:952069. doi:10.3389/fpubh.2023.952069
- 700 20. Turobov A, Coyle D, Harding V. Using ChatGPT for Thematic Analysis. *arXiv.* Preprint
701 posted online May 13, 2024:arXiv:2405.08828. doi:10.48550/arXiv.2405.08828
- 702 21. Wei J, Tay Y, Bommasani R, et al. Emergent Abilities of Large Language Models. *Trans*
703 *Mach Learn Res.* Published online June 26, 2022. Accessed April 29, 2026.
704 <https://openreview.net/forum?id=yzkSU5zdwD>
- 705 22. Anis S, French JA. Efficient, Explicatory, and Equitable: Why Qualitative Researchers
706 Should Embrace AI, but Cautiously. *Bus Soc.* 2023;62(6):1139-1144.
707 doi:10.1177/00076503231163286
- 708 23. Li KD, Fernandez AM, Schwartz R, et al. Comparing GPT-4 and Human Researchers in
709 Health Care Data Analysis: Qualitative Description Study. *J Med Internet Res.*
710 2024;26(1):e56500. doi:10.2196/56500

24. 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. *Soc Sci Comput Rev.* 2024;42:997-1019. doi:10.1177/08944393231220483
25. Hitch D. Artificial Intelligence Augmented Qualitative Analysis: The Way of the Future? - Danielle Hitch, 2024. *Qual Health Res.* Published online December 8, 2023. Accessed February 5, 2026. <https://journals.sagepub.com/doi/10.1177/10497323231217392>
26. Zhang H, Wu C, Xie J, Lyu Y, Cai J, Carroll JM. Redefining Qualitative Analysis in the AI Era: Utilizing ChatGPT for Efficient Thematic Analysis. *arXiv.* Preprint posted online September 19, 2023:arXiv:2309.10771. doi:10.48550/arXiv.2309.10771
27. Şen M, Sen S, Şahin T. A New Era for Data Analysis in Qualitative Research: ChatGPT! *Shanlax Int J Educ.* 2023;11:1-15. doi:10.34293/education.v11iS1-Oct.6683
28. Siiman LA, Rannastu-Avalos M, Pöysä-Tarhonen J, Häkkinen P, Pedaste M. Opportunities and Challenges for AI-Assisted Qualitative Data Analysis: An Example from Collaborative Problem-Solving Discourse Data. In: Huang YM, Rocha T, eds. *Innovative Technologies and Learning.* Lecture Notes in Computer Science. Springer Nature Switzerland; 2023:87-96. doi:10.1007/978-3-031-40113-8_9
29. Lee VV, Lubbe SCC van der, Goh LH, Valderas JM. Harnessing ChatGPT for Thematic Analysis: Are We Ready? *J Med Internet Res.* 2024;26(1):e54974. doi:10.2196/54974
30. Nguyen-Trung K. ChatGPT in thematic analysis: Can AI become a research assistant in qualitative research? *Qual Quant.* 2025;59:4945-4978. doi:10.1007/s11135-025-02165-z

- 731 31. Gao J, Shu Z, Yeo SY, et al. Efficiency with Rigor! A Trustworthy LLM-powered Workflow
732 for Qualitative Data Analysis. *arXiv*. Preprint posted online October 26,
733 2025:arXiv:2501.00775. doi:10.48550/arXiv.2501.00775
- 734 32. Mehta SD, Paul S, Awiti E, et al. Evaluation of large language models within GenAI in
735 qualitative research. *Sci Rep*. 2025;15(1):34993. doi:10.1038/s41598-025-18969-w
- 736 33. Duke MR. Artificial intelligence and the end(s) of qualitative data analysis. *Hum Organ*.
737 2025;0(0):1-3. doi:10.1080/00187259.2025.2546439
- 738 34. Atlas.ti. Published online 2025. <https://atlasti.com>
- 739 35. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J*
740 *Mach Learn Res*. 2011;12:2825-2830.
- 741 36. Solbiati A, Heffernan K, Damaskinos G, Poddar S, Modi S, Cali J. Unsupervised Topic
742 Segmentation of Meetings with BERT Embeddings. *arXiv*. Preprint posted online June 24,
743 2021:arXiv:2106.12978. doi:10.48550/arXiv.2106.12978
- 744 37. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-
745 Networks. In: Inui K, Jiang J, Ng V, Wan X, eds. *Proceedings of the 2019 Conference on*
746 *Empirical Methods in Natural Language Processing and the 9th International Joint*
747 *Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for
748 Computational Linguistics; 2019:3982-3992. doi:10.18653/v1/D19-1410
- 749 38. Ng MY, Helzer J, Pfeffer MA, Seto T, Hernandez-Boussard T. Development of secure
750 infrastructure for advancing generative artificial intelligence research in healthcare at an

751 academic medical center. *J Am Med Inform Assoc JAMIA*. 2025;32(3):586-588.
752 doi:10.1093/jamia/ocaf005

753 39. Jowsey T, Braun V, Clarke V, Lupton D, Fine M. We Reject the Use of Generative Artificial
754 Intelligence for Reflexive Qualitative Research. *Qual Inq*. Published online December 17,
755 2025:10778004251401851. doi:10.1177/10778004251401851

756 40. Bhaduri S, Kapoor S, Gil A, Mittal A, Mulkar R. Reconciling Methodological Paradigms:
757 Employing Large Language Models as Novice Qualitative Research Assistants in Talent
758 Management Research. *arXiv*. Preprint posted online August 20, 2024:arXiv:2408.11043.
759 doi:10.48550/arXiv.2408.11043

760 41. Chin-Purcell L. NLP-Assisted Qualitative Coding Method. Published online 2026.
761 doi:10.5281/zenodo.19896562

762

763

764

765

766

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AdditionalFile1InterviewGuide.docx](#)
- [AdditionalFile2NLPMETHODSTESTED.docx](#)
- [AdditionalFile3PromptIteration.docx](#)
- [AdditionalFile4Codebook.docx](#)