Generative AI for Qualitative Analysis in a Maternal Health Study: Coding In-depth Interviews using Large Language Models (LLMs)

Shan Qiao[1], Xingyu Fang[2,3,4], Camryn Garrett[1], Ran Zhang[1], Xiaoming Li[1], Yuhao Kang[2,3,*]

[1] Department of Health Promotion, Education, and Behavior, Arnold School of Public Health, University of South Carolina

[2] GISense Lab, Department of Geography and the Environment, The University of Texas at Austin

[3] Department of Geography, University of South Carolina

[4] School of Civil, Environmental and Geomatic Engineering, University College London, London, United Kingdom

[*] Corresponding author: yuhao.kang@austin.utexas.edu

Study Objectives: In-depth interviews are one of the most widely used approaches for qualitative studies in public health. The coding of transcripts is a critical step for information extraction and preliminary analysis. However, manual coding is often labor-intensive and time-consuming. The emergence of generative artificial intelligence (GenAI), supported by Large Language Models (LLMs), presents new opportunities to understand human languages, which may significantly facilitate the coding process. This study aims to build a computational coding framework that uses GenAI to automatically detect and extract themes from in-depth interview transcripts.

Methods: We conducted an experiment using transcripts of in-depth interviews with maternity care providers in South Carolina. We leveraged ChatGPT to perform two tasks automatically: (1) deductive coding, which involves applying a predefined set of codes to dialogues; and (2) inductive coding, which can generate codes from dialogues without any preconceptions or assumptions. We fine-tuned ChatGPT to understand the content of the interview transcripts, enabling it to detect and summarize codes. We then evaluated the performance of the proposed approach by comparing the codes generated by ChatGPT with those generated manually by human coders, involving human-in-the-loop evaluation.

Results: The results demonstrated the potential of GenAI in detecting and summarizing codes from in-depth interview transcripts. ChatGPT could be utilized for both deductive and inductive coding processes. The overall accuracy of GenAI is higher than 80% and the codes it generated showed high positive associations with those generated manually. More impressively, GenAI reduced the time required for coding by 81%, demonstrating its efficiency compared to traditional methods.

Discussion: GenAI models like ChatGPT show high generalizability, scalability and efficiency in handling large datasets, and are proficient in multi-level semantic structure identification. They demonstrate promising results in qualitative coding, making it a valuable tool for supporting people in public health research. However, challenges such as inaccuracy, systematic biases, and privacy concerns must be addressed when using them in practice. GenAI-based coding results should be handled with caution and reviewed by human coders to ensure accuracy and reliability.

1. Introduction

Qualitative study usually involves collecting and analyzing non-numerical data (e.g., text, video, or audio) to understand concepts and perceptions, attitudes and opinions, and lived experiences in specific contexts [1], [2]. Qualitative studies have been widely used in social sciences and public health research [3], [4]. Researchers employ qualitative methods to gather in-depth insights into problems, discover hidden patterns, and generate new research ideas [5]. In mixed-methods research, qualitative data and analysis can also be used to provide detailed explanations and interpretations of quantitative results [6], [7].

The in-depth interview is a primary qualitative data collection strategy in which trained interviewers ask participants questions based on interview guides that cover key topics and allow for probing into additional points [8]. In-depth interviews are thus characterized by their flexible structure, interactive process, and deep conversation via asking follow-up questions [9]. The process of analyzing in-depth interview data typically requires researchers to transcribe interview audios verbatim, review the transcripts, code the data, and identify emerging themes based on the categorization of codes and synthesis of their relations. Coding is a critical step in qualitative analysis not only as a fundamental step in data analysis but also as a key process in understanding and interpreting the deeper meaning of qualitative data.

Coding is a process of assigning labels or categories to data extractions (e.g., words, phrases, sentences, or paragraphs) that describe their content or meaning [10], [11]. There are two main types of coding: deductive and inductive coding [12]. Deductive coding is often theory-driven, consisting of a predefined coding scheme (e.g., based on theory or literature review) that sorts data into preset categories [13]. Deductive coding relies on existing theories, concepts, or frameworks from previous research (e.g., codebook or coding framework) to guide data coding [12], [13]. Inductive coding is data-driven and focuses on extracting patterns, themes, and categories from detailed data without preconceptions [14]. Inductive coding is a process of generating codes and themes grounded in the dataset [12].

Coding is a time-intensive process given that it involves in-depth reading, understanding, and labeling texts line by line [15]. In recent years, with the development of computer software, many qualitative data analysis software have begun to provide coding functions, helping researchers efficiently manage and organize qualitative datasets [16]. However, the cost of expensive software and the opaque data processing workflow may be challenges for individual researchers or small research teams with limited budgets.

The emergence of Generative Artificial Intelligence (GenAI), with the development of several new tools and methods, offers promising opportunities for qualitative research [17]. The field of Natural Language Processing (NLP) has made significant advancements in enabling computers to understand human languages and manipulate texts [18]. A notable example of this advancement refers to ChatGPT, a cutting-edge NLP model developed by OpenAI [19]. Built on the Generative Pre-trained Transformer (GPT) architecture, ChatGPT is designed to generate human-like conversations [20]. By training on vast amounts of text data, ChatGPT can generate coherent, contextually relevant responses to a wide range of human language inputs. More importantly, the superior ability of ChatGPT to understand language allows it to engage in dynamic and natural conversations. This deep understanding enables ChatGPT to respond in ways that are contextually appropriate, nuanced, and reflective of human-like reasoning, creating opportunities across a wide range of domains. Researchers are increasingly exploring the integration of ChatGPT into various

fields, recognizing its potential to revolutionize the handling of qualitative data. For instance, researchers in psychology and public health have begun investigating how ChatGPT can be leveraged to facilitate their studies [21], [22], [23]. Therefore, we aim to examine the potential of ChatGPT in assisting qualitative data analysis in health studies. Using in-depth interview data collected from maternity care providers in South Carolina, the current study aims to assess 1) the feasibility of using ChatGPT in both deductive coding and inductive coding; as well as 2) the performance of ChatGPT in coding, in terms of accuracy and validity, by comparing its coding results with those coded manually.

2. Data

2.1. Interview Raw Data

The qualitative data used in this study were derived from a previous qualitative study regarding racial disparities in maternal healthcare services and outcomes during the COVID-19 pandemic in South Carolina. In that study, we conducted in-depth interviews among 39 women who gave birth from March 2020 to July 2021 and 9 maternity care providers (i.e., physicians, nurses, and case managers) from clinics serving communities with a high proportion of Black and/or Hispanic populations. In the current study, only the data from the maternity care providers was used. More details about the dataset and study protocol can be referenced in Zhang et al. 2024 [24]. The audio recordings of the interviews were transcribed with the assistance of Otter.ai [25], and the transcripts were manually reviewed and refined by two research assistants to ensure accuracy. The questions covered in the interview guide are listed in Table 1. The reuse of the transcript data was approved as a study amendment by the IRB committee at the University of South Carolina (#Pro00115169).

Table 1. Questions of the in-depth interview guide

#	Questions
1	What are the challenges in providing maternal care?
2	How have you dealt with those challenges?
3	Are there any differences in labor and delivery in your facilities during COVID-19 compared with pre-pandemic?
4	How did COVID-19 change the policies and practices in the department/unit where you work?
5	Do you think the changes have affected the quality of care for women? How?
6	Do you think the changes have affected maternal health outcomes? How?
7	Are there any changes in practice or procedure that you think should be continued beyond COVID-19? If so, which?
8	Based on your observations, what are the specific challenges that your patients have experienced in the pandemic?
9	What has made them stressed and anxious? Would you like to share any examples?
10	How would you describe your clients' psychological conditions during the pregnancy in general?
11	Did they complain about any problems related to mental health?
12	What about their psychological conditions after giving birth?
13	What do you think about the health disparity in maternal health outcomes in South Carolina?
14	In your opinion, what are the main factors that contribute to the disparities?
15	How did the COVID-19 pandemic affect these factors?
16	Do you have any suggestions to reduce health disparity in maternal health in South Carolina?
17	If there was an intervention, or a program, dedicated to addressing the maternal health of Black/African American or Latino women throughout the Covid-19 pandemic, what would you like to see from that program?
18	Do you have any suggestions for how that program could best support Black/African American or Latino women?
19	What questions do you have for me? Is there anything else you would like to add?

2.2. Interview Transcript and Structure

Given the structure of an interview, the raw interview transcript typically follows a question-and-answer structure that allows researchers to consider a variety of responses, between participants, to the same question, thereby identifying diverse opinions and lived experiences. To facilitate GenAI's coding, we first reformatted the text into structured, analyzable data by matching each question with its corresponding answer. Each question-answer pair (hereafter referred to as a *dialogue*) was assigned a unique identifier (ID#), making the data computer-readable and facilitating efficient data management and query. **Figure *1*** illustrates the reformatting process from raw text into structured data, where interviewers' questions were aligned with participants' responses. This structured data was saved in .CSV files, with each participant's responses organized accordingly.

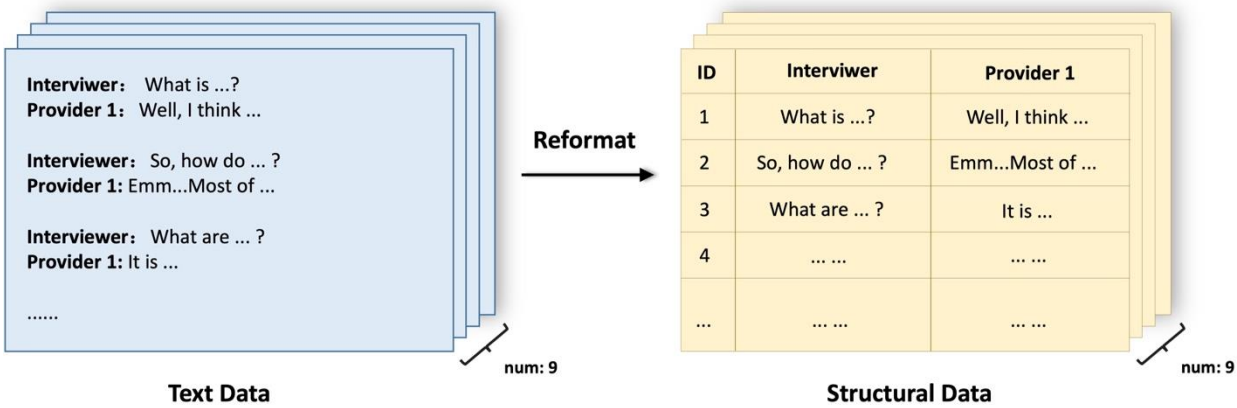

Figure 1. Reformatting raw text data into structured computer-readable data. Each interviewers' question is matched with its corresponding participants' answer as a dialogue.

2.3. Codebook

The codebook was developed using a deductive coding approach based on the pre-listed questions in the interview guide. In qualitative research, a *codebook* is an important tool used to guide the coding process [26]. It documents the codes (or "tags", "labels") used to analyze qualitative data, along with their corresponding definitions, application guidelines, and examples. Moreover, the codebook provides a common framework for the research team, ensuring that all members follow the same standards and rules when coding the data, which improves coding consistency and reliability. In deductive coding, which requires the researcher to classify the data into preset codes, the codebook plays a crucial role in providing clear code definitions and coding guidelines. The preexisting codebook (Table 2) from our previous qualitative study [24] was also used in the present study to guide coding processes. In addition, this preexisting codebook is utilized to aid in the GenAI models' performance assessment for coding in-depth interviews, ensuring that the assessments were aligned with established coding standards.

Table 2. Examples of the codebook for maternity care in-depth interview

Questions	Themes
What are the challenges providing maternal care	Vaccination hesitancy
	.....

	Policies turned women away from care
How have you dealt with those challenges?	Telemedicine
	.....
	Policies
.....	
How did COVID-19 change the policies and practices in the department/unit where you work?	Visitation policy
	.....
	Telemedicine

3. Coding of In-depth Interview with GenAI

In this study, we utilized ChatGPT 4, one of the most cutting-edge GenAI models empowered by Large Language Models (LLMs), to code in-depth interview data. Through carefully designed instructions and question prompts, we aimed to investigate the potential of using ChatGPT to automatically classify, summarize, and interpret interview transcripts. Notably, we tested the feasibility and efficiency of ChatGPT in both deductive and inductive coding processes. This method was feasible as it is particularly suitable for processing large amounts of textual data, often found in qualitative studies.

3.1. Deductive Coding

Deductive coding involves applying predefined codes from a codebook to the data [27]. This process requires a thorough understanding of participants' responses in order to accurately assign the appropriate codes based on the codebook's definitions and guidelines. Traditionally, the deductive coding process follows five general steps [13]:

(1) Develop the codebook: Create a comprehensive codebook with predefined codes, definitions, and application guidelines;
(2) Familiarize with data and codebook: Human coders review the interview transcripts to understand the context and details of the codebook;
(3) Apply codes: Assign appropriate codes to the data based on the predefined codebook, matching codes to the responses;
(4) Review and refine: Continually review and refine the codebook and coding process through pilot testing and iterative revisions;
(5) Generate coded data: Produce a dataset with coded responses, ready for further analysis.

In this study, we utilized ChatGPT to automate these steps. ChatGPT was employed to understand the dialogues in the interview transcripts, interpret the providers' answers, and apply the predefined codes from the codebook (Section 2.3) to the entire interview dataset. Since the codebook was provided from the prior study, ChatGPT was able to be utilized to simulate steps (2) through (5). This process involved two key steps for deductive coding: *dialogue filtering* and *encoding*. **Figure *2*** demonstrates the process of using ChatGPT for deductive coding.

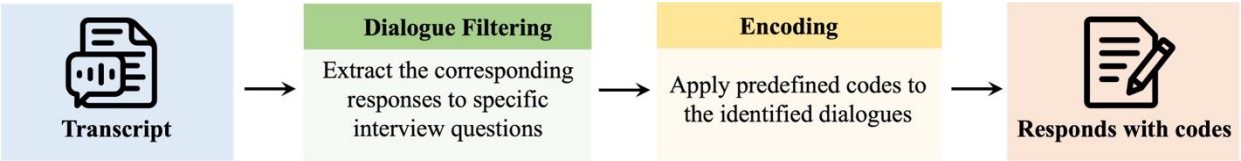

Figure 2. The workflow of using ChatGPT for deductive coding.

3.1.1. Dialogue Filtering

In-depth interviews often feature a deep, conversational nature where participants might address supporting points for one question while responding to another, either unintentionally or to reinforce their viewpoints. As a result, the same perspective might be repeated across different responses and intersect with multiple interview questions. Given this complexity, it is impractical to code every part of the dialogue, as responses may contain multiple relevant codes for different questions. To address this issue, the first step was *dialogue filtering*. Using predefined codes, ChatGPT was employed to identify specific interview questions within the dialogue and locate the corresponding responses. This step ensured that only relevant sections of the dialogue are further coded, enhancing the accuracy and efficiency of the analysis. The prompt template used for dialogue filtering was illustrated in **Figure *3***. The prompt instructed ChatGPT to identify if a specific interview question is present within the dialogue and, if so, to return the ID of the relevant dialogue. If no relevant content was found, it returned a 0. By doing so, for each question, we identified all relevant dialogue from each participant for each question.

Prompt: Identify if there is a dialogue asking the question: '{question}' (if yes tell me the number the dialogue, no then say 0, just give me the number): '{dialogue_string}'

Parameter:
{question}: Specific interview question
{dialogue_string}: All dialogues in an interview transcript

Figure 3. The prompt template used for dialogue filtering. If the given question is asked in the interview transcript, returns the row number of the dialogue, else then returns 0.

3.1.2. Encoding

In this step, we applied predefined codes to the dialogues identified in the previous step, for each question. Instructions were sent to ChatGPT to determine if there was a matching code and to encode the answer accordingly. **Figure *4*** illustrates the prompt template used for encoding. The prompt contained the question and predefined codes. An additional code '0' was designed to indicate "No Code" for answers that did not fit any codes in the codebook. ChatGPT then assessed if the answer belonged to one of these predefined codes.

Prompt: Can you identify which is the class of the content:'{answer}' among the class: '{predefined codes}'

Parameter:
{answer}: Answer from the corresponding dialogue
{predefined codes}: Predefined codes by experts

Figure 4. The prompt template used for encoding. The output was ChatGPT's assessment of if the answer belonged to one of the predefined codes.

Throughout the two steps of dialogue filtering and encoding, ChatGPT could extract and classify the participants' responses based on the predefined codes, providing coded interview data. Specifically, ChatGPT simulated the deductive coding process by processing the interviewers' questions regarding various aspects of maternal health services. This method ultimately generates

a structured dataset for further analysis and has the potential to mirror the traditional human deductive coding approach.

It is worth noting that ChatGPT may produce inaccurate and misleading results, or what are known as "hallucinations," where the AI generates incorrect or nonsensical text. Therefore, it is necessary to review the codes generated by GenAI. Unlike human deductive coding where human coders can review and refine the coding process iteratively, researchers must manually verify the applicability of the codes generated by ChatGPT. This review process ensures that the codes and codebooks used for deductive coding are accurate and reliable. More details on this review process are discussed in Section 3.3.

3.2. Inductive Coding

Inductive coding refers to creating codes based on the data without any preconceptions or assumptions [28]. Researchers performed a thematic analysis, to generate a set of codes, by identifying and summarizing key points mentioned by each participant. After deriving codes from the data, a codebook was developed for further qualitative analysis. Braun and Clarke's six-phase coding framework is widely adopted for thematic analysis of inductive coding and was employed in the present study [29]. The framework involves the following steps:

(1) Familiarizing: Read the data thoroughly, transcribe if necessary, and jot down initial ideas.
(2) Generating initial codes: Systematically code features across the dataset, collating relevant data.
(3) Searching for themes: Organize codes into potential themes, gathering relevant data for each theme.
(4) Reviewing themes: Check the application of themes, in relation to the coded extracts and the entire dataset, generating a thematic "map" of the analysis.
(5) Defining and naming themes: Refine themes and the overall narrative, generating clear definitions and names.
(6) Producing the report: Conduct the final analysis, select compelling examples, relate the analysis to the research question and literature, and produce a scholarly report.

After performing the thematic analysis to generate the codebook, researchers often then apply the codebook to the data, in another round of coding that resembles the methodology for deductive coding.

In this study, we used ChatGPT to assist researchers in performing the six steps of thematic analysis. ChatGPT's advanced text understanding capabilities allowed it to effectively simulate the thematic analysis procedure, starting from the second step. The output of using ChatGPT for inductive coding is a codebook derived from participants' responses. The specific methods included *dialogue filtering*, *code generation*, and *code aggregation*, which are illustrated in Figure 5.

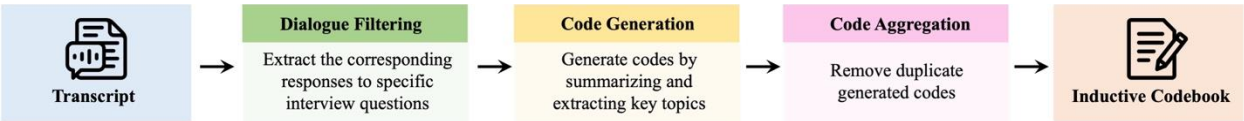

Figure 5. The workflow of using ChatGPT for inductive coding.

3.2.1. Dialogue Filtering

We performed dialogue filtering using a strategy similar to that illustrated in Section 3.1.1. To locate participants' responses to specific questions, we first instructed ChatGPT to determine whether a specific question was asked during the interview. When the question was identified, ChatGPT provided the dialogue ID, allowing researchers to observe participants' responses to the question. We utilized the same prompt as shown in **Figure *3***.

3.2.2. Code Generation

After identifying the dialogues for each question, we instructed ChatGPT to perform *code generation* by summarizing and extracting key topics from the dialogues. Figure 6 illustrates the prompt template used for generating codes. In this step, answers to specific questions, from all interview transcripts, were collected and ChatGPT was tasked with extracting and generating codes from the collected answers. This process enabled the identification of themes and topics across participants' responses which were further utilized to create the codebook.

Prompt: Can you extract the topics from the answers about the question '{question}': '{collected_answers}'

Parameter:
{question}: Specific interview question
{collected_answers}: All answers responding to the specific interview question in interview transcripts

Figure 6. The prompt template used for code generation.

3.2.3. Code Aggregation

After GenAI generated a set of codes reflecting the key topics in response to each question, we further aggregated these codes into categories and developed themes. The first task was to merge codes with similar semantic meanings. While ChatGPT generated the most fitting code for each participant's response, slight variations in expressions resulted in many repetitive codes that were useless for analysis and required manual aggregation. The second task was to synthesize codes by their conceptual meanings. The codes were combined and grouped into categories based on their definitions and scope in the conceptual system. The process of *code aggregation* ensured that the extracted codes were concise and useful for identifying themes in further qualitative data analysis.

3.2.4. Manual Review and Deductive Coding

Manual review of the generated and aggregated codes was crucial to ensure the applicability of the codebook. Although ChatGPT is often efficient at identifying general topics present in dialogue, it may lack domain-specific knowledge. Therefore, experts were consulted to review and refine the codebook by selecting the most relevant and meaningful codes. The generated codes were then matched to the participants' responses according to the codebook, in a process similar to deductive coding. More details are introduced in Section 3.1.2.

After finalizing the codebook, we further assigned codes to dialogues, which resembles conducting a round of deductive coding. In particular, we matched participants' responses to the corresponding questions, in accordance with the codebook generated by the inductive coding process.

3.3. Evaluation

Two types of evaluation strategies were performed to assess the validity of the generated codes and encoding results: a human-centered assessment and a machine-based assessment. Human-centered assessment involved incorporating human-in-the-loop processes to guide and validate the correctness of the coding. We invited experts in maternal health to assess both the accuracy of the coding results and the generated codebook. By comparing differences between the manually created and the GenAI-generated codebooks, we gained a comprehensive understanding of the applicability of ChatGPT for qualitative coding. In addition, a machine-based metric was used to supplement the human-centered assessment by quantifying the accuracy of GenAI in coding dialogue and generating codes. Combining human expertise with machine-based analysis, we provided a comprehensive evaluation of the effectiveness of GenAI in qualitative coding.

3.3.1. Human-centered Assessment

Existing studies suggest that ChatGPT may overlook domain knowledge. Therefore, in the human-centered assessment, we engaged domain experts in maternal health, with relevant background knowledge, for a thorough evaluation of the qualitative data. One faculty member and one doctoral student in public health, both with extensive experiences in qualitative studies, evaluated both the deductive and inductive coding processes. Each of them independently reviewed the codes. Any disagreements were sufficiently discussed and resolved before finalizing the assessment report.

In the deductive coding evaluation, the experts first assessed whether the coded dialogues were accurately classified with appropriate codes. They then determined whether the applied codes comprehensively aligned with the codes in the pre-existing codebook. In the inductive coding evaluation, one faculty member and two doctoral students reviewed the codebook generated by GenAI and marked the codes as “accepted” or “unaccepted”.

3.3.2. Machine-based Metric

Due to the limitations of human-centered assessment (e.g., time-consuming, labor-intensive, subject to biases and errors), we created a machine-based metric that can assess the accuracy and consistency of generated code, reducing the potential impacts of personal bias and subjectivity.

To accomplish this, we evaluated and quantified the differences between the manually defined preexisting codes and those generated using ChatGPT, by comparing their semantic similarity. In particular, we utilized a cutting-edge text embedding method, BERT, to convert the codes into high-dimensional vectors [30]. No preprocessing was conducted since the codes (whether generated by humans or ChatGPT) consisted of only a few words or paraphrases. These high-dimensional feature vectors captured the semantic meanings of the input codes. We calculated the cosine similarity between the two high-dimensional vectors using the following formula.

$$S(V_G, V_H) = \frac{V_G \cdot V_H}{\|V_G\| \times \|V_H\|}$$

V_G represents the high-dimensional vector of a GenAI-generated code, V_H represents the high-dimensional vector of a human-defined code, $S(V_G, V_H)$ represents the cosine similarity of V_G and V_H. $S(V_G, V_H)$ ranges from 0 to 1, and the higher the $S(V_G, V_H)$, the more similar the two codes.

4. Results

In this section, we report evaluation results regarding the performance of GenAI for both deductive and inductive coding tasks. In addition, we compare the time efficiency of using GenAI vs. human coders throughout the qualitative coding process.

4.1 Deductive Coding

In this study, we used ChatGPT to simulate the human coding process for interview responses. The process involved filtering dialogues within the transcripts for each question and then encoding by matching the identified response dialogue with the corresponding codes from the codebook (i.e., human defined codes from a previous qualitative study). After filtering and encoding, we obtained both the responses to the interview questions and the thematic codes discussed in those dialogues.

To assess the accuracy of the deductive coding conducted by ChatGPT, we calculated the "acceptance rate" and "consistency rate". Specifically, two research staff reviewed the dialogues and their corresponding codes, assigned by ChatGPT, to determine whether each code was "accepted" or "unaccepted". We then calculated the acceptance rate by dividing the number of "accepted" codes by the total number of the codes. Figure 7 illustrates the overall average acceptance rate and the acceptance rate of each question. ChatGPT demonstrated high acceptance rate in deductive coding, exceeding 90% in 13 out of the 18 questions. This indicates that ChatGPT had a strong understanding of the in-depth interview topics and was effective in applying predefined codes to dialogues, accurately.

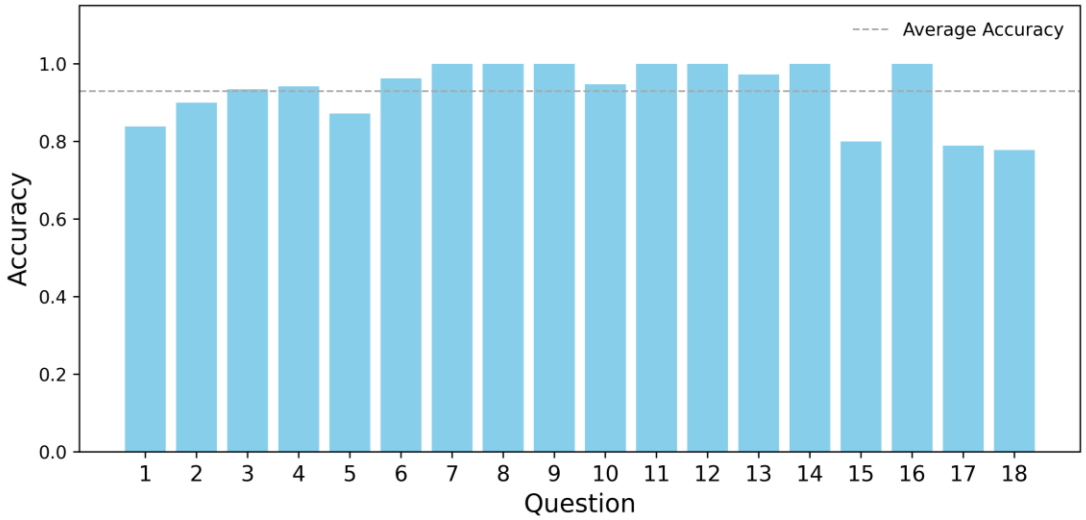

Figure 7. Accuracy of ChatGPT's deductive codes as determined by domain experts based on "acceptance rate". The x-axis represents the question number, and the y-axis represents the accuracy.

The consistency rate was based on a comparison between the codes assigned by ChatGPT and those assigned by human coders. If both human coders and ChatGPT applied the same code to the same dialogue, it was considered to have "consistent coding". Consistency rate was the proportion of "consistent codes" among the total codes. It should be noted that, a code assigned by ChatGPT may be accepted by reviewers as an "accepted code" but it may not be consistent with the code assigned by human coders in the preexisting codebook.

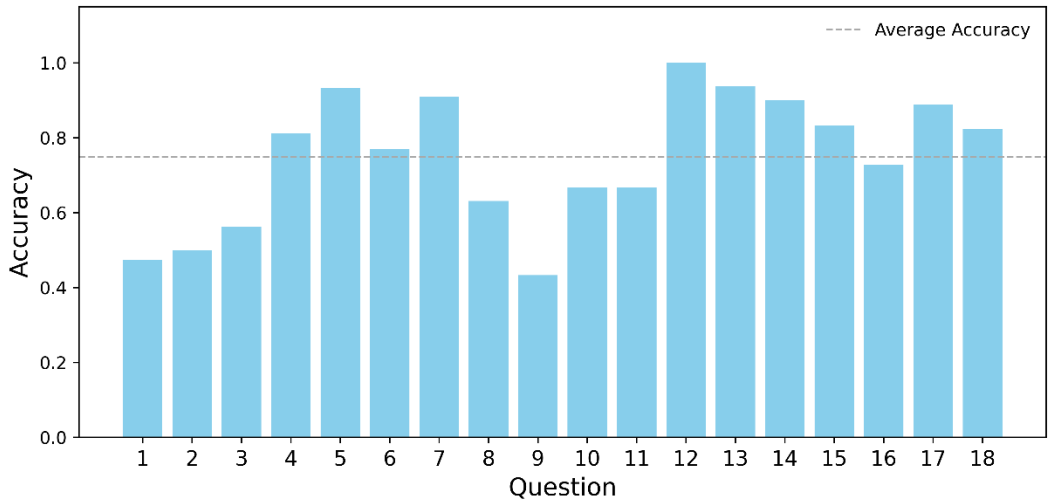

Figure 8 illustrates the average consistency rate and the consistency rate for each question. The results indicate that the codes generated by ChatGPT did not always align with those originally assigned by humans. Specifically, the highest consistency rate was achieved in Question 12, reaching 100%, while the lowest consistency rate was in Question 9, at 47.5%. The discrepancy between the consistency rate and the acceptance rate suggests that ChatGPT may supplement traditional coding methods by generating insights that human coders might miss.

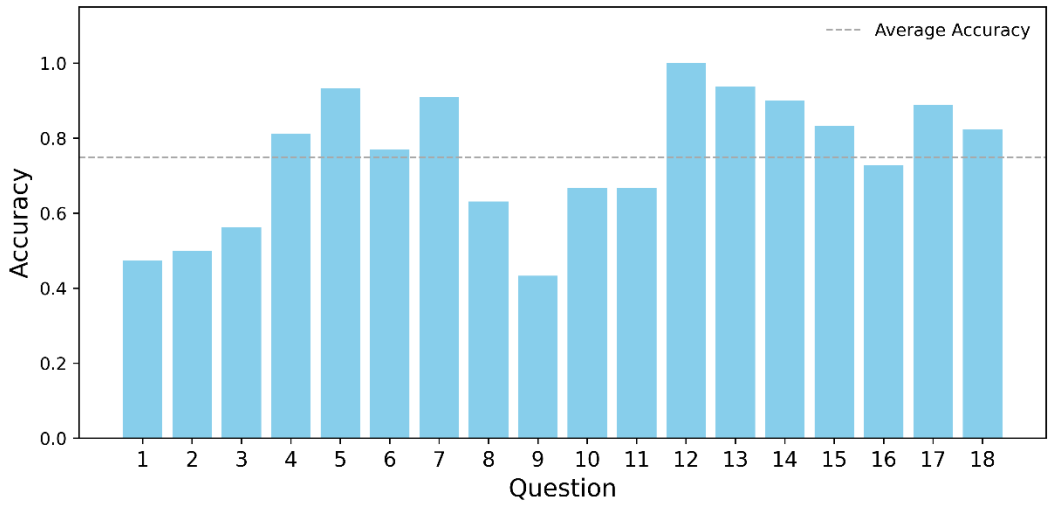

Figure 8. Accuracy of ChatGPT's deductive codes based on “consistency rate”. The x-axis represents the question number, and the y-axis represents the accuracy.

The results also show different patterns between the two approaches. ChatGPT identified interview responses related to questions based solely on the subsequent answer context, ensuring consistency with the interviewer’s question. However, it may overlook relevant information from

other parts of the interviewee's responses. In comparison, human coding is inherently more subjective. Within a human-generated codebook, codes might originate from different sections of the transcript with inconsistencies. Furthermore, we observe that different codes for the same question in the human-generated codebooks may have semantic overlap or convey similar meanings. This can pose a challenge for ChatGPT, as it may be unable to match dialogues to codes that, while semantically similar, are not identical in the codebook.

4.2 Inductive Coding

When conducting inductive coding, ChatGPT generated corresponding codes for each question by summarizing and extracting key themes from the in-depth interview transcripts. Table 2 lists the generated codes for three example questions. Each question was coded based on the participant's direct responses to the questions.

Table 2. Three example questions and their corresponding codes created by ChatGPT.

Question	Code
Example A	
What are the challenges providing maternal care?	1 Visitor restrictions, 2 Telehealth limitations, 3 Inconsistent care standards, 4 Partner attendance issues, 5 Missed appointments, 6 Vaccination hesitancy, 7 Difficulty managing COVID patients, 8 Socio-economic disparities, 9 Technology accessibility, 10 Pandemic safety concerns
Example B	
How did COVID-19 change the policies and practices in the department / unit where you work?	1 Knee jerk reactions to changes, 2 Communication issues, 3 Visitation policy alterations, 4 COVID testing for patients, 5 Push on vaccinations, 6 Labor and delivery support changes, 7 Personalized care challenges, 8 Clothing policy changes, 9 Donning and doffing PPE, 10 Masking rules enforcement
Example C	
Do you think the changes have affected the quality of care for women? How?	1 Reduced support for women, 2 Less tolerance for birth plans, 3 Nursing staff burnout, 4 Early postpartum discharge, 5 Fear of COVID in hospitals, 6 Increased depression and anxiety, 7 Missed appointments, 8 Halted preventive screenings, 9 Reduced physical touch in care, 10 Implicit bias and fear

Overall, the codes created from the text, comprising the codebook, are relevant and demonstrate alignment, indicating that ChatGPT has provided relatively accurate and meaningful results. For instance, the codes generated for the question about challenges in providing maternal care (Example A) cover a broad range of relevant issues including objective (e.g., visitor restrictions), personal (e.g., vaccination hesitancy), and social factors (e.g., socio-economic disparities). These results indicate that ChatGPT has successfully identified and categorized key themes from the participants' responses, providing a relevant codebook for each question.

The performance of ChatGPT's performance of inductive coding was quantitatively evaluated by three research staff. They reviewed each code and determined whether it was "accepted" or

"unaccepted". Figure 9 presents the number of codes generated by GenAI for each question, the number of "accepted" codes after manual review, and the corresponding accuracy rates. The overall accuracy rate was observed to be 85.15% across all questions. For individual questions, the accuracy rates ranged from a maximum of 100% to a minimum of 66.67%. Notably, 11 questions had coding accuracy rates equal to or greater than 90%, representing more than half of the total number of questions, with 8 questions achieving a coding accuracy rate of 100%. Only three questions had coding accuracy rates less than 80%. After reviewing these three questions, we found that the responses are highly descriptive and ChatGPT tended to extract more generalized information from these detailed responses. For example, in coding Question 6, "Do you think the changes have affected maternal health outcomes? How?", ChatGPT provided the broad summary code of "Changes affecting maternal health." For Question 10, which asks, "How would you like to describe your clients' psychological conditions during the pregnancy in general?", ChatGPT provided summarized codes like "Patients' health and baby's health concerns", while human-generated codes were more specific, such as "Psych," "Anxiety", and "Anger". This finding indicated that a small portion of questions did not achieve high accuracy in coding. Overall, the inductive coding results are reliable and ChatGPT demonstrated promise in providing support for qualitative coding.

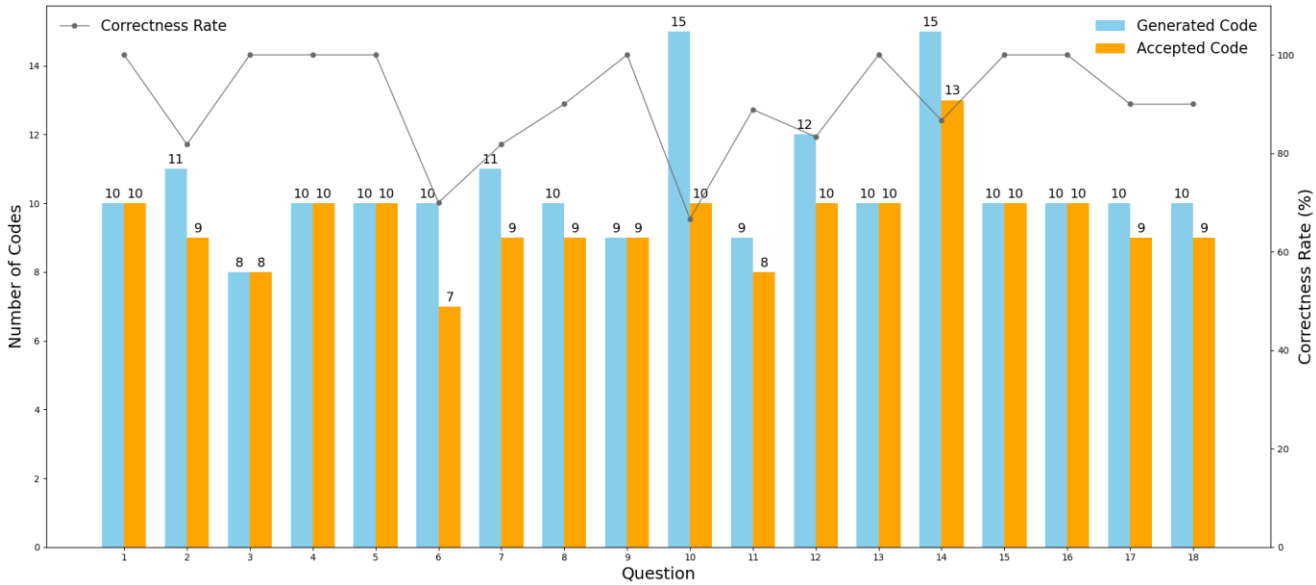

Figure 9. The number of generated codes, the number of "accepted" codes after manual review, and the accuracy rate for each question.

Moreover, we evaluated the inductive codes generated by ChatGPT, for each question, by comparing them to the codes in the preexisting codebook (i.e., previously developed for deductive coding) using a machine-based metric. We calculated the text semantic similarity of the inductive codes generated by ChatGPT with the codes in the codebook. Figure 10 shows the cosine similarity between the high-dimensional vectors of the two sets of texts. The highest similarity score we obtained was 0.879, with an overall similarity of 0.795. Three sample codes are illustrated in Table 3 with high, medium, and low semantic similarity scores. This indicates that the semantic meaning of the codes generated by ChatGPT are very similar to those in the preexisting codebook by human coders, illustrating the proficiency of using ChatGPT for the inductive coding of in-depth interview texts.

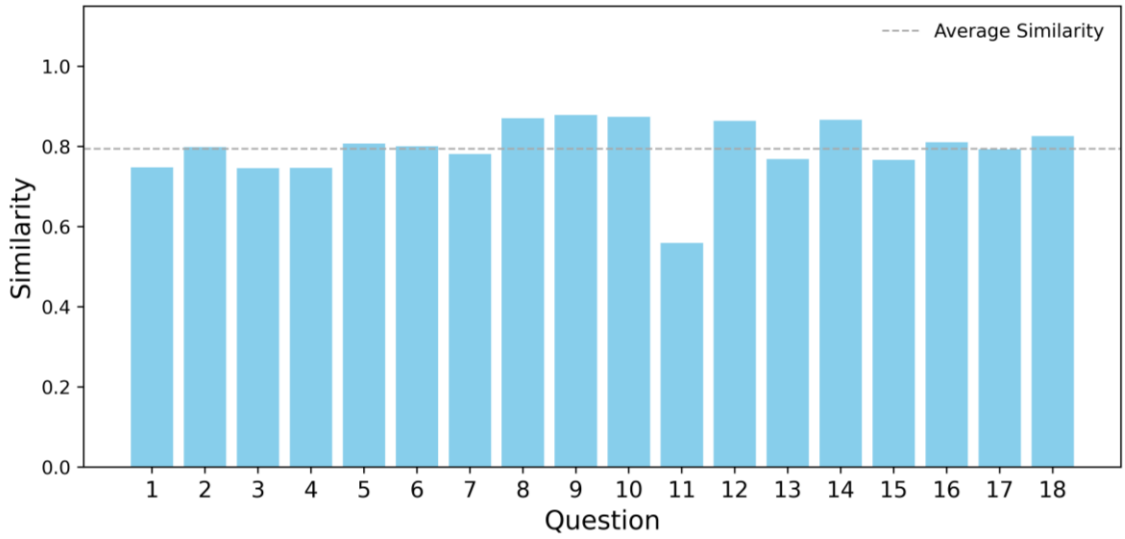

Figure 10. Text similarity between the inductive codes generated by ChatGPT for each question and the codes in the codebook. The x-axis represents text similarity, and the y-axis represents the question number.

Table 3. Similarity of inductive codes generated by human-coders and ChatGPT

Inductive Code generated by human	Inductive Code generated by ChatGPT	Similarity
Example A		
Covid vaccine, Job security/ work, Domestic situation, Risk for COVID, Mask wearing, Childcare, Hard jobs, Vaccine hesitancy coupled with a high risk work environment, Childcare/ work/ balance, Pregnancy stress coupled with COVID stress, Vaccine and mask discord, Financial, transportation issues, Unsafe work environment, Having to homeschool or manage virtual learning for children, Anger and Anxiety, Concerns about pregnancy during the pandemic among better-off patients, Financial concerns, Isolation, Fear of COVID itself, Vaccine decisions, Domestic violence maybe but screening wasn't optimal, Scared to come in to appointment, Family members dying from COVID, Vaccine hesitancy	Jobs and employment, Family safety concerns, Difficulty breathing in masks, Pregnancy mask concerns, Deciding on vaccination, Isolation from loved ones, Financial worries, Fear of the virus, Vaccine anxiety	0.879
Example B		
Visitation restrictions, Differences in care for COVID positive patients, Less triages, Policy changes affected patient experience, PPE and infection control measures, Telemedicine, Center shut down, Nothing stopped at their clinic	Visitor restrictions, Barriers to care, Reduced one-on-one support, PPE requirements, Changes in room availability, Staffing and nursing shortages, Patient experiences affected, Supply chain issues	0.746
Example C		

Depression, General life stressors, COVID fear, Anxiety, Mood, Sleep issues	No more mental health complaints; Pregnancy-related insomnia; Increased domestic violence; Emotional, physical, sexual abuse; Overwhelmed mothers; Small children at home; Challenges of work-from-home; Unemployment effects; Pandemic-related fatigue	0.559

4.3 Time efficiency

In addition to demonstrating accuracy, the GenAI-based coding approach has also shown significant time efficiency compared to traditional methods. The total estimated time spent by human coders for deductive coding was approximately 30.5 hours, of which 4.5 hours were spent getting familiar with the 9 transcripts and 11.25 hours were spent assigning codes to the dialogues. The inductive coding process required a similar amount of time spent for coding, with an additional 14.75 hours needed to generate the codebook.

The usage of GenAI was found to significantly reduce the time required to complete both deductive and inductive coding. The time expense of GenAI-based coding primarily reflects two activities: *setup* and *coding*.

The first is the time spent setting up the environment, including data preparation for GenAI coding, designing appropriate prompts and reformatting the transcript files. The second is the time required for the actual coding, by GenAI. The first part, setting up the environment, requires most of the time in the work pipeline, around 5 hours. In particular, a human coder takes 3 hours to reformat the original transcript into Excel to prepare the materials for use in the GenAI coding. Then, it

takes around 2 hours to test different prompts and reformat the transcript files to be fed into GenAI. Once the setup is complete, the actual coding process by GenAI is much faster, taking only around 20 minutes to encode themes following the deductive approach, and 20 minutes to create the codebook following the inductive approach.

It should be noted that additional human-centered tasks are still necessary with current GenAI techniques. For example, these methods require domain experts to review the coding results and assess the quality of coding, which results in additional time costs. However, with the development of more robust and advanced GenAI models in the future, these review steps may be simplified, and the process may be shortened. Consequently, the total time required for GenAI-based coding might be reduced to just 19% of the time required for traditional coding methods.

5 Discussion

5.1 ChatGPT for coding in maternal health

These results of using ChatGPT for deductive and inductive coding highlight the promise of GenAI in qualitative studies. ChatGPT achieved relatively high performance in the encoding process for deductive coding, and in the coding and subsequent creation of the codebook for inductive coding. Traditionally, coding interview transcripts manually is time-consuming and subject to human biases. Using GenAI in coding can greatly shorten this process and maintaining relatively high accuracy. Thus, it could enhance the efficiency and objectivity in qualitative research.

5.2 Opportunities brought by GenAI for qualitative study

The emergence of GenAI technology provides unprecedented opportunities to enhance qualitative research methods, particularly in the coding of qualitative data. We summarize the potential of using GenAI in qualitative studies based on its three prominent advantages.

High generalizability. GenAI has high generalizability, making it easily adaptable to various domains without requiring extensive prior knowledge or domain-specific input. The advanced natural language understanding ability makes GenAI automatically screen interview transcripts and identify specific questions or topics related to a given question. When designing prompts in ChatGPT, there is no need to input a large training dataset. Even when using a small number of sample dialogues, ChatGPT could still achieve high accuracy in identifying and categorizing topics in the data. In our study, although the main research topic is related to maternal health, ChatGPT successfully identified discussions related to cannabis, opioid, and alcohol use from the participants' responses, showing its flexibility across different public health areas. This suggests GenAI can be applied to qualitative studies in many domains beyond maternal health.

Multi-level semantic structure identification. GenAI demonstrates an ability to identify multi-level semantic structures when coding, which is well represented in the code aggregation process. ChatGPT can generate fine-resolution codes, and merge those with similar semantic meanings into broader categories. This granularity of analysis provides a more detailed understanding of qualitative data, enabling researchers to capture the complexity and diversity of participants' perspectives. Traditionally, qualitative researchers must manually merge and aggregate codes at different levels, which may lack standardization and heavily rely on human experience. With GenAI, subtopics can be automatically grouped under their main topics, enhancing the overall coherence of the analysis.

Scalability and efficiency in large datasets. GenAI excels at handling large datasets, providing a scalable solution for qualitative research. Analyzing large datasets is often characterized by time

and human resource limitations. Additionally, human coders may introduce potential biases into the coding process, due to variations in text interpretation and working conditions (e.g., fatigue). Comparatively, GenAI can strictly follow consistent coding criteria across the entire dataset, reducing potential biases and errors. In addition, once the environment setup is completed, GenAI can process transcripts in parallel by leveraging its Application Programming Interfaces (APIs), which can dramatically improve coding efficiency. Through this function, researchers could performing scripts to use GenAI functions in parallel. Therefore, GenAI can serve as a tool, completing most of the initial coding and codebook generation tasks, thereby allowing human coders and experts to expend their efforts on the careful review and discussion of outputs as well as the applicability, interpretation, and dissemination of the findings. Thus, collaboration between GenAI and human coders can potentially shift the time required for qualitative data analysis while preserving the quality of the data analysis.

In summary, GenAI can significantly improve coding efficiency in qualitative studies. With the continuous evolution of cutting-edge GenAI technologies beyond ChatGPT, there is an increasing potential to enhance the efficiency and quality of GenAI-based coding. We expect a continued benefit for qualitative research when carefully integrating advanced artificial intelligence into data analysis.

5.3 Challenges and Concerns When Using GenAI

Despite the promising potential of GenAI for coding in qualitative studies, several challenges and concerns also deserve our attention, including inaccuracy, systematic biases, and privacy issues.

The inaccuracy issues in GenAI-based coding stems from two aspects. First, we observed that ChatGPT tends to produce more general codes than human coders and lacks the domain-specific terminology and professional knowledge unique to experts. For example, for Question 10, “How would you like to describe your clients’ psychological conditions during the pregnancy in general?”, one of the codes generated by ChatGPT was “Post-traumatic stress from having COVID,” while the human-generated code was “Some form of PTSD symptoms in patients who were really sick with COVID,” which used more professional medical terminology and provided a more detailed description of the psychological issues described. This discrepancy may be attributed to ChatGPT being trained on a general corpus with limited exposure to public health knowledge, especially maternal health knowledge. Second, the “hallucination” phenomena in commonly observed in current GenAI models. Given that current GenAI models, such as ChatGPT, are trained on a general corpus, rather than specific domain knowledge, GenAI can produce inaccurate content and false information. These hallucinations further demonstrate the necessity of human coders and experts to carefully review all outputs generated by GenAI during the coding process.

Additionally, systematic biases are rooted in representation issues of the training datasets. As the behind-the-scenes generation process of GenAI is inaccessible to prompt engineers, it is difficult to quantify errors and debug them. The datasets used to train GenAI are not readily available to users and may underrepresent certain vulnerable population groups, thereby overlooking their lived experiences and opinions. Based on these datasets, GenAI may then replicate or amplify existing biases, leading to misleading or harmful conclusions. This lack of transparency highlights the importance of rigorously evaluating GenAI’s output to avoid inaccurate or misleading results.

Further, using GenAI systems, such as ChatGPT, poses privacy concerns, particularly regarding the handling of sensitive or personal interview data. Current GenAI systems may use input data

for further model training, which could lead to misuse or sharing of private information. In our study, we removed all personal and identifiable data from the transcripts, to protect participants' privacy and confidentiality, before conducting data analysis, following our standardized qualitative study protocol [24]. In addition, we applied for and obtained the approval of an IRB amendment for the reuse of transcripts collected for a traditional qualitative study for use in the present context. Considering the potential ethical issue and privacy protection, we chose not include the transcripts or data from ethnic minority women in this exploratory study. Researchers must sufficiently assess potential ethical issues before entering sensitive information into GenAI models, follow relevant IRB policies, and implement rigid data protection strategies.

5.4 Limitations and Future Directions

There remains room for improvement in this study. First, this study is limited to in-depth interviews with maternity care providers during the COVID-19 pandemic, which may not provide generalizable results in different settings. Future studies should apply the computational workflow across a wider range of datasets and research contexts to showcase its adaptability and potential in qualitative research. Second, the inherent subjectivity and potential biases in the manual coding process may impede the power of human-centered assessment, making the performance assessment of GenAI coding complex. Third, in the current study, we tested only one GenAI model (i.e., ChatGPT). Future studies may explore the potential to incorporate more advanced GenAI approaches such as agent-based GenAI methods, or different GenAI models (e.g., commercial vs. open-sourced models).

6 Conclusion

In summary, we proposed a computational framework for coding qualitative data by leveraging advanced GenAI methods. Our findings demonstrated the potential of GenAI (i.e., ChatGPT) in automating both deductive and inductive coding processes, thereby enhancing efficiency and scalability in qualitative studies. In our case study on maternal health, ChatGPT could assign appropriate codes for dialogues in deductive coding and generate meaningful codes from the data to uncover key themes and topics in inductive coding. Our evaluation showed a relatively high overall accuracy rate of GenAI-generated codes and a significant reduction in the time spent on coding processes. The findings suggest that GenAI can effectively support qualitative coding, provide reliable results, and offer substantial benefits for various public health studies. Despite its promise, we have also noted the challenges in applying GenAI in qualitative research, including inaccuracies, systematic biases, and privacy issues. Therefore, it is crucial to have domain experts carefully review the results of GenAI-based coding given the current technologies. With advancement in AI technologies, we believe qualitative research will continue to benefit from these cutting-edge methods.

References

[1] U. Kuckartz, *Qualitative Text Analysis: A Guide to Methods, Practice and Using Software*. SAGE, 2014.

[2] G. Mey, "Qualitative Methodology," in *International Handbook of Psychology Learning and Teaching*, J. Zumbach, D. A. Bernstein, S. Narciss, and G. Marsico, Eds., Cham: Springer International Publishing, 2023, pp. 453–478. doi: 10.1007/978-3-030-28745-0_22.

[3] H. K. Mohajan, “Qualitative research methodology in social sciences and related subjects,” *J. Econ. Dev. Environ. People*, vol. 7, no. 1, pp. 23–48, 2018.

[4] M. Lichtman, *Qualitative Research for the Social Sciences*. SAGE Publications, 2013.

[5] S. J. Tracy, *Qualitative Research Methods: Collecting Evidence, Crafting Analysis, Communicating Impact*. John Wiley & Sons, 2024.

[6] P. M. Nardi, *Doing Survey Research: A Guide to Quantitative Methods*, 4th ed. New York: Routledge, 2018. doi: 10.4324/9781315172231.

[7] A. Moser and I. Korstjens, “Series: Practical guidance to qualitative research. Part 3: Sampling, data collection and analysis,” *Eur. J. Gen. Pract.*, vol. 24, no. 1, pp. 9–18, Jan. 2018, doi: 10.1080/13814788.2017.1375091.

[8] B. DiCicco-Bloom and B. F. Crabtree, “The qualitative research interview,” *Med. Educ.*, vol. 40, no. 4, pp. 314–321, 2006, doi: 10.1111/j.1365-2929.2006.02418.x.

[9] E. Knott, A. H. Rao, K. Summers, and C. Teeger, “Interviews in the social sciences,” *Nat. Rev. Methods Primer*, vol. 2, no. 1, pp. 1–15, Sep. 2022, doi: 10.1038/s43586-022-00150-6.

[10] B. B. Kawulich, “The BERA/SAGE Handbook of Educational Research,” pp. 1–1170, 2016.

[11] W. E. Ruona, “Analyzing qualitative data,” in *Research in organizations: Foundations and methods of inquiry*, Berrett-Koehler, 2005, pp. 223–263.

[12] T. Azungah, “Qualitative research: deductive and inductive approaches to data analysis,” *Qual. Res. J.*, vol. 18, no. 4, pp. 383–400, Jan. 2018, doi: 10.1108/QRJ-D-18-00035.

[13] A. J. Bingham and P. Witkowsky, “Deductive and inductive approaches to qualitative data analysis,” in *Analyzing and interpreting qualitative research: After the interview*, SAGE, 2021, pp. 133–146.

[14] M. Naeem, W. Ozuem, K. Howell, and S. Ranfagni, “A Step-by-Step Process of Thematic Analysis to Develop a Conceptual Model in Qualitative Research,” *Int. J. Qual. Methods*, vol. 22, p. 16094069231205789, Mar. 2023, doi: 10.1177/16094069231205789.

[15] V. Elliott, “Thinking about the coding process in qualitative data analysis,” *Qual. Rep.*, vol. 23, no. 11, 2018, Accessed: Sep. 05, 2024. [Online]. Available: https://ora.ox.ac.uk/objects/uuid:5304bf7f-6214-4939-9f1b-b64415d4fac1

[16] T. M. Paulus, “Using Qualitative Data Analysis Software to Support Digital Research Workflows,” *Hum. Resour. Dev. Rev.*, vol. 22, no. 1, pp. 139–148, Mar. 2023, doi: 10.1177/15344843221138381.

[17] C. Stokel-Walker and R. Van Noorden, “What ChatGPT and generative AI mean for science,” *Nature*, vol. 614, no. 7947, pp. 214–216, Feb. 2023, doi: 10.1038/d41586-023-00340-6.

[18] K. R. Chowdhary, “Natural Language Processing,” in *Fundamentals of Artificial Intelligence*, New Delhi: Springer India, 2020, pp. 603–649. doi: 10.1007/978-81-322-3972-7_19.

[19] OpenAI *et al.*, “GPT-4 Technical Report,” Mar. 04, 2024, *arXiv*: arXiv:2303.08774. Accessed: Sep. 05, 2024. [Online]. Available: http://arxiv.org/abs/2303.08774

[20] G. Yenduri *et al.*, “Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions,” arXiv.org. Accessed: Sep. 05, 2024. [Online]. Available: https://arxiv.org/abs/2305.10435v2

[21] D. Demszky *et al.*, “Using large language models in psychology,” *Nat. Rev. Psychol.*, Oct. 2023, doi: 10.1038/s44159-023-00241-5.

[22] S. S. Biswas, “Role of Chat GPT in Public Health,” *Ann. Biomed. Eng.*, vol. 51, no. 5, pp. 868–869, May 2023, doi: 10.1007/s10439-023-03172-7.

[23] K. M. Jang *et al.*, “Place identity: a generative AI’s perspective,” *Humanit. Soc. Sci. Commun.*, vol. 11, no. 1, pp. 1–16, Sep. 2024, doi: 10.1057/s41599-024-03645-7.

[24] Zhang, T. Byrd, S. Qiao, M. E. Torres, X. Li, and J. Liu, “Maternal care utilization and provision during the COVID-19 pandemic: Voices from minoritized pregnant and postpartum women and maternal care providers in Deep South | PLOS ONE.” Accessed: Sep. 05, 2024. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0300424

[25] “Otter.ai - Voice Meeting Notes & Real-time Transcription.” Accessed: Jan. 08, 2023. [Online]. Available: https://otter.ai/

[26] K. Roberts, A. Dowell, and J.-B. Nie, “Attempting rigour and replicability in thematic analysis of qualitative research data; a case study of codebook development,” *BMC Med. Res. Methodol.*, vol. 19, no. 1, p. 66, Mar. 2019, doi: 10.1186/s12874-019-0707-y.

[27] “Accelerating Deductive Coding of Qualitative Data: An Experimental Study on the Applicability of Crowdsourcing | Proceedings of Mensch und Computer 2021.” Accessed: Sep. 06, 2024. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3473856.3473873

[28] L. Kuczynski and K. Daly, “Qualitative Methods for Inductive (Theory-Generating) Research: Psychological and Sociological Approaches,” in *Handbook of Dynamics in Parent-Child Relations*, 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc., 2003, pp. 373–392. doi: 10.4135/9781452229645.n18.

[29] V. Braun and V. Clarke, “Using thematic analysis in psychology,” *Qual. Res. Psychol.*, vol. 3, no. 2, pp. 77–101, Jan. 2006, doi: 10.1191/1478088706qp063oa.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” May 24, 2019, *arXiv*: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.