# When Qualitative Research Meets Large Language Model: Exploring the Potential of QualiGPT as a Tool for Qualitative Coding

HE ZHANG, College of Information Sciences and Technology, Penn State University, USA

CHUHAO WU, College of Information Sciences and Technology, Penn State University, USA

JINGYI XIE, College of Information Sciences and Technology, Penn State University, USA

FIONA RUBINO, College of Engineering, Penn State University, USA

SYDNEY GRAVER, College of Engineering, Penn State University, USA

CHANMIN KIM, College of Education, Penn State University, USA

JOHN M. CARROLL, College of Information Sciences and Technology, Penn State University, USA

JIE CAI\*, College of Information Sciences and Technology, Penn State University, USA

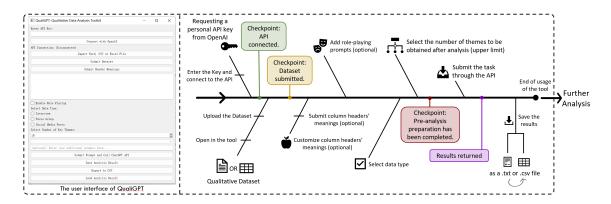


Fig. 1. Overview of the qualitative analysis toolkit, QualiGPT. The user interface of QualiGPT is displayed on the left. On the right side, the usage flow and design logic of QualiGPT are presented.

Qualitative research, renowned for its in-depth exploration of complex phenomena, often involves time-intensive analysis, particularly during the coding stage. Existing software for qualitative evaluation frequently lacks automatic coding capabilities, user-friendliness, and cost-effectiveness. The advent of Large Language Models (LLMs) like GPT-3 and its successors marks a transformative era for enhancing qualitative analysis. This paper introduces QualiGPT, a tool developed to address the challenges associated with using ChatGPT for qualitative analysis. Through a comparative analysis of traditional manual coding and QualiGPT's performance on both simulated and real datasets, incorporating both inductive and deductive coding approaches, we demonstrate that QualiGPT significantly improves the qualitative analysis process. Our findings show that QualiGPT enhances efficiency, transparency, and

Authors' addresses: He Zhang, hpz5211@psu.edu, College of Information Sciences and Technology, Penn State University, University Park, Pennsylvania, USA, 16802; Chuhao Wu, cjw6297@psu.edu, College of Information Sciences and Technology, Penn State University, University Park, Pennsylvania, USA, 16802; Jingyi Xie, jzx5099@psu.edu, College of Information Sciences and Technology, Penn State University, University Park, Pennsylvania, USA, 16802; Fiona Rubino, far5185@psu.edu, College of Engineering, Penn State University, University Park, Pennsylvania, USA, 16802; Sydney Graver, sjg6347@psu.edu, College of Engineering, Penn State University Park, Pennsylvania, USA, 16802; ChanMin Kim, cmk604@psu.edu, College of Education, Penn State University, University Park, Pennsylvania, USA, 16802; John M. Carroll, jmc56@psu.edu, College of Information Sciences and Technology, Penn State University, University Park, Pennsylvania, USA, 16802; Jie Cai, jpc6982@psu.edu, College of Information Sciences and Technology, Penn State University, University Park, Pennsylvania, USA, 16802.

<sup>\*</sup>Corresponding author.

accessibility in qualitative coding. The tool's performance was evaluated using inter-rater reliability (IRR) measures, with results indicating substantial agreement between human coders and QualiGPT in various coding scenarios. In addition, we also discuss the implications of integrating AI into qualitative research workflows and outline future directions for enhancing human-AI collaboration in this field.

CCS Concepts: • **Human-centered computing** → HCI design and evaluation methods; **Collaborative and social computing**; **Interactive systems and tools**; **Interaction techniques**.

Additional Key Words and Phrases: ChatGPT, toolkit design, large language models, prompt engineering, qualitative analysis, analytical evaluation, api application

#### 1 INTRODUCTION

Qualitative research provides a unique perspective into individuals' comprehension, attitudes, and insights regarding technology, phenomena, and specific topics. Over time, an increasing number of researchers have acknowledged the significance of qualitative methodologies across diverse fields. In the CSCW (Computer-Supported Cooperative Work) field, qualitative research methods are particularly important because they provide deep insights into how people collaborate with the support of technology. CSCW researchers often use qualitative methods such as interviews, observations, and content analysis to understand complex socio-technical systems and work practices. However, as the scale and complexity of data increase, CSCW researchers face the challenge of efficiently processing and analyzing large volumes of qualitative data, after all, analyzing qualitative data can be labor-intensive [40], especially with extensive and complex datasets. Moreover, the task of coding qualitative data not only demands significant effort but also poses challenges related to understanding context and ensuring consistency. Coding, arguably the most crucial task in qualitative analysis, is both a beloved and challenging aspect for analysts. Continuously optimizing methods for processing qualitative data remains a common goal among these professionals. As the production of qualitative data continues to surge, there is an escalating demand for innovative techniques to streamline and enhance the thematic analysis process [5].

To address these challenges, researchers have ventured into the development and utilization of qualitative analysis software. These tools employ computer-assisted collaborative efforts to simplify data management and enhance efficiency [18]. While such software has indeed streamlined the coding process and improved the quality of coding to some extent, existing platforms like Nvivo<sup>1</sup> and atlas.ti<sup>2</sup> still have limitations in terms of performance and operational complexity, failing to fully meet the needs of researchers [7, 34].

In addition to the high subscription costs, learning to use these software tools for qualitative data analysis is not straightforward. Early-career researchers or analysts often find themselves investing a significant amount of time in understanding how to accomplish their target tasks within these environments and navigating the multifaceted UI interfaces [27]. However, many of these features are designed to cater to specific needs. In other words, not all functionalities within the software are utilized frequently by analysts, leading to increased learning overheads. As described by Ragavan et al. [75], analysts not only have to be concerned about their primary tasks at hand but also bear the additional learning costs associated with the tools (software) they choose. Therefore, the development of a more user-friendly tool to reduce the workload of analysts in their primary workflows becomes especially crucial. Starting from 2022, with the emergence of GPT-3, researchers began to widely recognize the immense potential of Large Language Models (LLMs) in various domains. The subsequent releases of GPT-3.5, GPT-4 and GPT-40, were perceived

<sup>&</sup>lt;sup>1</sup>https://lumivero.com/products/nvivo/

<sup>&</sup>lt;sup>2</sup>https://atlasti.com/

When Qualitative Research Meets Large Language Model: Exploring the Potential of QualiGPT as a Tool for Qualitative Coding

by many as heralding a comprehensive technological revolution. The advent of large-scale language models seemed to offer a new avenue of possibilities. It was during this time that we were inspired to ponder whether it might be feasible to leverage LLMs to assist in qualitative analysis, aiming to enhance both efficiency and performance. To achieve this objective, we approached it from a practical standpoint, selecting one of the most popular LLM applications developed by OpenAI, ChatGPT and its API, as our research platform to bolster the universality of our research contributions. We drew inspiration from the recent works [24, 25, 93] on enhancing qualitative analysis using ChatGPT, emphasizing the importance of prompts and by extending some of the future work they had highlighted.

In summary, this study first categorizes and summarizes the typical issues encountered when using ChatGPT, identifying four major categories that encompass eight common types of erroneous ChatGPT responses. Concurrently, we compiled concerns from previous studies wherein analysts expressed reservations about employing ChatGPT for qualitative analysis tasks, as well as the challenges ChatGPT faces in such contexts. With these issues and challenges in mind, we introduced QualiGPT: a user-friendly integrated tool built on API and prompt design, specifically tailored for thematic analysis of qualitative data. The tool's performance was evaluated using inter-rater reliability (IRR) measures [51], with results indicating substantial agreement between human coders and QualiGPT in various coding scenarios.

We deployed QualiGPT [92] on both simulated and real datasets and compared its performance to manual coding. The results show that this tool effectively addresses the challenges inherent in the traditional qualitative data coding process. It streamlines the qualitative analysis workflow, reduces costs associated with processing qualitative data, and alleviates concerns regarding transparency and credibility in using ChatGPT for qualitative analysis. Additionally, due to its integrated design and API implementation, QualiGPT offers marked improvements in usability, user-friendliness, privacy protection, and performance over the web version of ChatGPT. When compared to conventional software, QualiGPT provides a more insightful user interface, significantly lowering the learning and usage costs for researchers.

# 2 RELATED WORK

Qualitative research is an important way to understand the human world. This method is based on experience and subjectivity, maintaining an open stance towards the meanings of the research subject through practical interactivity [64], and it is widely used by researchers across various disciplines. In the process of qualitative research, data analysis and criteria for rigor form an important component [72]. Researchers need to remain sensitive to the data and provide discussion and insights through continuous interpretive and critical thinking. Coding, as a method for analyzing qualitative data, is crucial for searching concepts, ideas, themes, and categories within the data to help researchers organize and interpret the data [29]. As Saldana [71] stated, no one can have the ultimate authority over the "best" method for qualitative coding due to the inherently high flexibility of qualitative analysis. Therefore, in this section, we will quickly review some of the processes and methods of qualitative coding and summarize some common coding processes. In addition, we will review a series of challenges that exist in qualitative coding. Based on our research objectives, we will examine both manual coding and coding conducted with technological assistance.

#### 2.1 Manual Qualitative Coding

Coding is the most crucial part of qualitative research, and the process is often labor-intensive [17, 40, 95]. Although coding enhances the understanding of data, and researchers can gain a series of new insights from the coding process, human coders or annotators still typically need to spend a considerable amount of time on the manual coding process and face a range of challenges [93]. To further address or reduce the negative impact of the challenges encountered

during the coding process, we first briefly review the inductive and deductive coding methods and their associated challenges.

2.1.1 Inductive Coding. Inductive coding is one of the commonly used methods in qualitative analysis [78]. This approach is based on the data itself, extracting and inferring themes through the analysis of the data [22]. In this section, we will briefly review the processes and roles of human coders and AI in inductive coding.

In qualitative inductive coding, human coders play a vital role. By reviewing qualitative data, human coders are tasked with identifying key concepts, themes, or patterns that emerge from the data, without relying on any pre-established codebook [20]. This approach shares similarities with the process of open coding, as both do not rely on pre-existing theories or assumptions and maintain flexibility throughout the coding process to adjust coding strategies based on the data [53]. This suggests that coders often require a certain level of coding experience or need to work under the guidance of experts [86].

Inductive coding can be particularly challenging for novice coders, as qualitative data is often more complex and messier compared to quantitative data [41]. Human coders must ensure coding consistency and identify implicit concepts and themes in the inductive coding process, which demands strong reflective abilities and skills, such as memo-writing [76].

Moreover, since inductive coding is typically an iterative process, coders are required to manage the data and the entire project [54]. When collaborating with others in the coding and analysis stages, coders also face challenges such as maintaining an open mindset, avoiding over-interpretation, and possessing high management skills.

2.1.2 Deductive Coding. Deductive coding is another primary method in qualitative coding [20]. In deductive coding, human annotators interpret the data using a pre-established coding framework, applying codes from a codebook [48] to the data [29]. Once a certain number of codes have been generated, deductive coding can be applied [42]. This approach requires human annotators to carefully review the qualitative data and the codebook, assigning codes to segments of the qualitative data that correspond to the predefined categories in the codebook [31]. This process demands a lower level of experience in exploring implicit patterns in the data compared to inductive coding. However, it necessitates a higher understanding of each coding category and its connotations [4, 32].

Consistency is equally crucial in deductive coding, and maintaining coding consistency becomes a significant challenge [56]. Annotators typically need to ensure the consistency of their coding decisions through meetings, discussions, agreements, and technical methods, promoting a shared understanding of the codes and the content being coded [13, 37, 90]. Multiple rounds of discussions among coders are particularly important for resolving issues encountered during the deductive coding process.

When human annotators encounter data segments that do not fit neatly into the existing coding framework, they must decide how to handle these "exceptions," either by creating new codes or adjusting the existing ones [59]. Multiple rounds of discussions among coders are particularly important for resolving issues encountered during the deductive coding process, which has been evidenced in several previous studies [9, 12, 45].

# 2.2 Technology-Assisted Qualitative Coding

Although researchers in qualitative research can choose different coding methods based on various data types, research purposes, and research processes, these coding methods often present a series of challenges for human coders, such as manual labor costs [17, 40, 95], consistency [56] and reliability issues [30], coding management [6], coding experience [39], and the learning curve for novices [60]. To address these challenges, qualitative researchers have been

continuously striving to improve the performance of qualitative coding and tackle the challenges encountered during the qualitative coding process through technological means.

With the increasing amount of qualitative data being generated, Computer-Assisted Qualitative Data Analysis (CAQDA) has been playing a critical role in qualitative research [10, 62, 84]. CAQDA software nowadays offer a wide range of functionalities such as processing data from multiple media channels (text, picture, audio, and video), visualizing the analysis results through automatic plotting of data, and quickly generating predefined and customized reports [61]. While the capabilities vary greatly among applications, more advanced functionalities often comes at the cost of a high subscription fee [67] that potentially deter researchers away. As a result, some free and open-source alternatives have been developed to support the growing need of qualitative research, such as Taguette [65] and RQDA [11], although their functionalities tend to be more basic than commercial products. Another problem with CAQDA software is their user experience and learnability. Paulus et al. [58] find that initial encounters can be intimidating for novices, yet with proper guidance, researchers can effectively integrate these tools. Still, studies on the interface design of CAQDA are rather limited and a comparison of both commercial and open source applications is necessary for designing better tools

The combination of AI and qualitative research has begun to redefine how researchers approach qualitative data and analysis [35, 88]. Technologies, especially AI algorithms, provide potential for improved efficiency in analyzing large datasets, a task that traditionally requires substantial time and resources when conducted by human analysts. In fact, in earlier years, researchers have been using computers or technologies to assist in qualitative studies [14, 79, 84].

AI can be used to gather and organize qualitative data from various sources, like social media platforms, online forums, and digital archives. This not only saves time and resources but can also uncover a wider range of data points that might be overlooked in manual collection [19]. Also, AI-powered transcription services can transcribe audio and video data into text format quickly and accurately. Typically, transcription and encoding in qualitative research present the biggest challenges for researchers, often consuming a lot of time. However, a good assistant tool allows researchers to focus more on analysis rather than on data preparation [49]. AI models can provide an initial analysis of textual data by summarizing content, identifying key themes, sentiments, or trends, and even insightful advice and generating questions that can help guide further research [15, 43, 46, 69, 73]. By comparing AI findings with human analysis, researchers can increase the validity and reliability of their findings [28]. With AI's ability to process data rapidly, researchers can conduct real-time analysis during data collection, helping them adjust their research approach as needed based on preliminary findings [57].

2.2.1 LLMs and Prompt Engineering. The advent of automated qualitative analysis techniques has enabled qualitative researchers to analyze volumes of data that would be difficult to analyze manually [85], and the rise of LLMs may further enhance the efficiency of analysis.

Prompt engineering is the deliberate design and optimization of instructions, or "prompts", aimed at enhancing the performance and accuracy of LLMs when generating outputs [68, 91]. This strategy is crucial as the type and specificity of prompts provided to LLMs can significantly shape their responses.

ChatGPT<sup>3</sup> by OpenAI<sup>4</sup>, developed within the Generative Pretrained Transformer (GPT) framework, underscores the importance of prompt engineering [21]. It is capable of understanding and generating human-like text (natural language), offering an interactive experience similar to that of human interactions. Renowned for its expertise in diverse

<sup>3</sup>https://openai.com/chatgpt

<sup>4</sup>https://openai.com/

language tasks, such as producing human-like text, content generation, sentence completion, and in-depth essay or report writing [2, 8, 44, 47]. However, ChatGPT is not immune to errors. It may yield outputs that seem nonsensical or incorrect, particularly when faced with unclear or ambiguous prompts [36, 74]. Hence, applying prompt engineering to enhance the capabilities of LLMs is a crucial method.

The value of prompt engineering gains further emphasis from studies revealing improved outcomes when LLMs like ChatGPT receive meticulously crafted prompts. Techniques such as few-shot learning [96], chain-of-thought methods [83], and role-playing scenarios [23] have demonstrated considerable efficacy. However, the performance of ChatGPT, even when paired with refined prompt engineering, can differ based on the domain in question. Mastery in domain-specific knowledge is pivotal for honing the model's efficacy [80, 82]. Thus, practitioners are encouraged to weigh the specific application context carefully during prompt engineering [38]. For areas like qualitative analysis, employing an iterative methodology—consistently adapting and evaluating diverse prompt engineering strategies—may be instrumental in harnessing the full potential of ChatGPT [24, 25, 93].

Despite the impressive capabilities of AI, machine learning, and LLMs, the complex nature of qualitative analysis presents unique challenges that these technologies are still learning to navigate [33].

#### 3 OVERALL MOTIVATION AND DESIGN CONSIDERATIONS OF QUALIGPT

By leveraging techniques proposed by researchers for using ChatGPT in qualitative task analysis, we tested and refined these techniques on the web version of ChatGPT. We integrated the solutions into our toolkit, serving as resources and prior knowledge for the development of QualiGPT. Specifically, the design considerations encompass two main parts: the first pertains to the common concerns of qualitative analysts about applying ChatGPT to qualitative analysis tasks, as introduced in Section 3.1. The second pertains to some of the current shortcomings of the web version of ChatGPT, as discussed in Section 3.2.

Our development of QualiGPT was informed by testing and refining techniques proposed by researchers for utilizing ChatGPT in qualitative task analysis. We conducted these tests on the web version of ChatGPT and integrated the resulting solutions into our toolkit, serving as resources and prior knowledge for QualiGPT's development. The design considerations encompass two main areas: first, addressing common concerns of qualitative analysts regarding the application of ChatGPT to qualitative analysis tasks, as introduced in Section 3.1; and second, addressing some of the current limitations of the web version of ChatGPT, as discussed in Section 3.2.

# 3.1 Common Challenges and Concerns of Qualitative Analysis and the Use of ChatGPT in the Qualitative Analysis Process

In previous researches [25, 93], scholars have identified several significant challenges associated with incorporating LLMs into qualitative data analysis. We revisited these challenges and further explored how to address them by designing an integrated tool. We revisited these challenges and further contemplated how to address them through the design of an integrated tool. In summary, our design aims to tackle the following challenges: (1) Lack of Transparency: ChatGPT's "black box" nature makes it difficult for researchers to understand how it processes data. Better prompt design can improve interpretability and transparency, (2) Consistency and Context Understanding Issues: Varied responses and difficulty in maintaining context in multi-turn dialogues are key challenges. Streamlined and precise prompts can enhance consistency, (3) Difficulty in Prompt Design: Creating effective prompts is time-consuming and lacks a standardized approach. QualiGPT simplifies this process with pre-designed prompts, (4) Challenges in Understanding ChatGPT's Responses: Using ChatGPT for qualitative analysis doesn't always save time compared

to traditional methods, especially if its outputs are lengthy and disorganized. This contradicts the goal of improving efficiency. To address this, we can design prompts to standardize ChatGPT's outputs for better readability. Additionally, implementing strategies to prioritize reading sequences can help users quickly identify key concepts, enhancing overall efficiency, and (5) **Data Privacy and Security**: In the digital era, data privacy is a major concern, especially when using large language models like ChatGPT for research. The risk of sensitive data leakage is significant, as shown by historical data breaches [1]. While efforts like GDPR [81] and encryption techniques [70] are made to protect data privacy, individual users also need to be aware of their online security practices. According to OpenAI's policies, user data in ChatGPT can be used to improve the model, while API service data isn't used for model training. This suggests that the API service offers better data security [50].

#### 3.2 ChatGPT Initial Performance Test

During the design process, we analyzed 1,000 data entries from a public Discord channel to assess ChatGPT's ability to provide accurate and standardized content. We also summarized typical errors, examples, and potential solutions encountered during the testing process. These primarily include: (1) network errors, (2) incorrect handling of data, (3) Violation of policy, and (4) Out of limits. It's worth noting that these errors are not exclusive to using ChatGPT for qualitative analysis tasks and have a certain degree of universality across various applications.

#### 4 DESIGN OF QUALIGPT

To further benefit qualitative researchers, address the challenges presented in Section 3, and overcome the limitations of using ChatGPT on the web interface, we introduce QualiGPT. It's a meticulously crafted, integrated qualitative analysis tool based on prompt engineering and API. This tool features a user-friendly visual interface and is designed to be easily used even by those with no programming experience. Fig. 1 presents the user interface and usage flow of QualiGPT. Fig. 2 is the user manual for QualiGPT. In the following sections of this chapter, we will delve into the functionalities, advantages, and design considerations of QualiGPT.

#### 4.1 Design Practice

QualiGPT is designed with a user-centric approach, addressing the challenges researchers face in qualitative analysis and simplifying interactions for novices with LLMs. It integrates OpenAI's API and data processing libraries, providing a simple, user-friendly interface that requires minimal technical expertise. Users only need to provide their API key to access the functionalities of different versions of GPT. Prioritizing user privacy, QualiGPT allows individuals to control their API connections, ensuring secure data exchange and automatic data deletion post-session. Understanding the diversity of qualitative data, QualiGPT is flexible in processing various text formats. It accommodates user-specific parameters like role labels, ensuring analyses align with the dataset's nature. Its dynamic prompt generation, based on research and literature, combines user input with qualitative research principles. QualiGPT's goal is to deliver insights that are comprehensible, shareable, and actionable. The tool presents results in a clear, transferable format, including themes, descriptions, quotes, and participant counts, capturing the essence of qualitative research. With export options, QualiGPT further emphasizes practicality and convenience. The functionalities are detailed in the subsequent sections.

# 4.2 Components and Architecture

*4.2.1 API Connection.* QualiGPT integrates the OpenAI API and Python libraries for advanced text data processing. Users provide their API key for GPT access via a user-friendly interface. This setup bypasses the token input limit of

traditional ChatGPT by segmenting data, allowing for more comprehensive analysis. The API's flexibility also enhances data privacy and security, giving users more control over their data.

- 4.2.2 User Input and Data Formatting. QualiGPT supports various textual data formats like Word, .txt, .csv, and .xlsx. Users submit datasets in these formats, receiving a system prompt confirming successful data submission. Optimal processing requires labeled data to differentiate participants in conversations. Users can define roles (e.g., interviewer, interviewee) and provide conversation overviews, which are integrated into a series of prompts to guide GPT in tailored data analysis.
- 4.2.3 Prompt Generation. Prompt generation in QualiGPT is rooted in previous researches [25, 93], focusing on four components: Task Background, Task Description, Processing Method, and Expected Output. The prompts are customizable, with options like role-playing for expert analysis and selection of data types (e.g., Interviews, Focus Groups). Users can control the depth of analysis by specifying the number of key themes and adding additional instructions. The system offers "fixed", "dynamic", and "user-choice-based" prompts, allowing diverse customization based on user needs. The process culminates in submitting these prompts to GPT for analysis.
- 4.2.4 Analysis Results. The prompts generated by QualiGPT guide GPT to execute a qualitative analysis on the submitted dataset, ensuring a rigorous and insightful analytic process. In addition, they guide GPT to present its results in a streamlined, coherent format, tailored for user-friendly interpretation and data exports. Specifically, QualiGPT organizes the results into a tabular format that encapsulates thematic findings. Each table features four columns for 'Themes' which represent the overarching patterns or topics within the data. Following that is the "Description", illustrating the nuances and depths of these themes. To give a clearer context, the table also includes "Quotes" linked to each theme, showcasing direct excerpts from the dataset that support or explain the theme. Moreover, a "Participant Count" associated with each theme is presented, offering a quantitative insight into the theme's prevalence or significance. QualiGPT provides users with a practical tool to export these findings in a csv file format. This facilitates further analysis, sharing, or integration with other tools or databases. Additionally, for those keen on preserving the entire analytical journey, from the raw dataset, the constructed prompts, to the derived findings, QualiGPT offers an option to encapsulate all these elements into a singular txt file, ensuring comprehensive documentation and easy recall.

# 5 ANALYSIS AND VERIFICATION

To demonstrate the performance of QualiGPT, we applied it to both simulated data and real datasets. By comparing the topics returned by QualiGPT with manually coded results, we showcased the powerful potential of QualiGPT in qualitative data coding tasks. In the LLM-assisted coding process, we designed prompts based on the recommendations from past researches [25, 93]. The prompts included descriptions of the task background, objectives, methods, outputs, and inputs, incorporating role-playing and acknowledgment expressions. Modular features, such as role-playing and acknowledgments, are included in the predefined prompts of QualiGPT, while more specific and customized information, like task background, can be input through text boxes.

## 5.1 Case Study One - Exploring Themes and The Advantages of Integrated Tools

In our case study one, we used ChatGPT to create a simulated focus group dataset on "transitioning to remote work". The dataset comprises 9,309 words, averaging about 27 words per feedback. It includes 6 medium-length responses (average 112 words) and 2 long responses (average 391 words). This dataset provides diverse insights into the experiences

of transitioning to remote work, highlighting both benefits like flexibility and challenges like work-life balance and technical issues. Researchers agreed that the dataset offers a broad thematic variety, reflecting a range of personal experiences and strategies related to remote work from individuals of various backgrounds and job roles.

5.1.1 Results and Evaluation. We submitted the data to both ChatGPT (web version) and QualiGPT. In QualiGPT, we selected the data type as "focus group" and enabled the "role-playing" feature. We also chose to obtain 20 potential topics. The final response results from QualiGPT and ChatGPT (web version) were similar. However, when using the web version of ChatGPT, we encountered several troubling issues that were resolved in QualiGPT:

- 1. Due to the data volume exceeding the token limit for a single submission, we had to manually split the dataset and input it into ChatGPT using the copy-paste method.
- 2. On the web version, we had to manually input prompts multiple times for debugging.
- 3. We had to manually organize the output results, such as transferring the results to a spreadsheet.

This made the work time and complexity on the web version of ChatGPT much greater than using QualiGPT. To highlight the efficiency advantage of QualiGPT, we repeated the same analysis process in QualiGPT three times and timed each run. From entering the API (starting checkpoint) to saving to a .csv file (ending checkpoint), the results showed that the average time to complete the analysis process in QualiGPT was 96.5 seconds. We provide the simulated dataset used for testing in the supplementary materials, and we welcome researchers to use this dataset for a quick test on QualiGPT to experience the efficiency improvement compared to the web version of ChatGPT or manual coding.

## 5.2 Case Study Two - Inductive Coing by using LLM

In Case Study 2, we used a real social media dataset collected in the past study and conducted inductive coding on 200 entries using both manual coding and LLM-assisted coding methods.

For manual inductive coding, we performed topic modeling on the raw dataset, which resulted in 8 distinct topics. We selected data from two of these topics and randomly extracted 100 entries from each. These entries were then independently subjected to inductive coding by two research assistants. Each entry was labeled with tags consisting of 2-5 words. Each entry in the dataset is a message from a user in that channel. The dataset does not contain any identifiable information. In the first round of manual coding, coding 200 entries took several hours of work, and the entire manual coding process lasted close to a week including discussions and negotiations on coding.

For inductive coding by using LLM, we removed the manually coded labels from the data and submitted it to QualiGPT for analysis, and we designed prompts that allow QualiGPT to explore the data that was processed in the manual inductive coding section and generate corresponding codes without prior knowledge. Specifically, we interact with QualiGPT through prompts. First, we employ role-playing to activate QualiGPT's capabilities, such as "You are now an excellent qualitative data analyst and qualitative research expert." Then, we inform it about the task it needs to assist with and provide the task background, for example, "You need to perform inductive coding on a dataset that was obtained from a public Discord server named "TwitchDev". This server is run by non-staff volunteers such as moderators and administrators. TPDs (Twitch Platform Developers) will get a developer role in this Discord by the administrators of this Discord server if they prove that their program developments are building for Twitch users or using the Twitch-provided tools. The community has thousands of TPDs to share their experiences. It has more than 2,000 active members daily, including Twitch official staff, TPDs, broadcasters, and viewers." We also inform it about the format of the input data, such as "Each row in the dataset represents a single data entry," and specify the output format, "Please help me determine a possible code for each data entry and return the results in a tabular format, with the first column being the data index and the

second column being the code." Since GPT can only input a limited number of prompts (tokens) in each interaction, we additionally employ prompts with acknowledgment to enhance the power of context in memory between each iterations and inputting new data, such as "Great job! Please continue analyzing the following data: [New Data]." If GPT provides an incorrect format or is not processing the task correctly, we will regenerate it.

5.2.1 Results and Evaluation. After the coding was completed, we aimed to verify the IRR between human and LLM in the inductive coding task. We re-evaluated the consistency between the coding results produced by the two research assistants, after discussion and negotiation, and those produced by GPT-4. This re-evaluation was necessary because the coding was done without a prior codebook, leading to potential variations in vocabulary usage. Specifically, we marked data entries as "1" for consistency if their meanings were identical or similar, such as "twitch merch" (human-coded label) and "Twitch merchandise" (LLM-coded label). Conversely, entries were marked as "0" for inconsistency if their meanings differed or if they represented different levels, such as "reputation of developer" (human-coded label) and "Twitch user behavior and reputation" (LLM-coded label). After this annotation process, we calculated Cohen's Kappa, which resulted in a value of 0.57.

Overall, the LLM exceeded expectations in completing the inductive coding task. Achieving high consistency among different coders is particularly challenging without extensive prior knowledge [20], as each individual may have their own interpretation of the data [71].

#### 5.3 Case Study Three - Deductive Coding by using LLM

5.3.1 Codebook development. Before conducting deductive coding, we first needed to prepare a codebook. This codebook was developed based on the inductive coding results from Case 2. Specifically, the two research assistants compared the labels for each data entry, either agreeing on one of the two labels or combining them to develop more suitable labels. The negotiated labels were then compiled into the initial codebook.

The senior researcher and the two research assistants reviewed this initial codebook, which contained 171 labels. They removed labels that were not relevant to the theme of social support. This refinement process resulted in a final codebook comprising 54 labels, where label 0 indicated that the message's topic was irrelevant, and label 53 indicated that the topic was relevant but not specified by the other labels. These labels became the codes that were going to be used to categorize in deductive coding process.

- 5.3.2 Deductive Coding Process. Subsequently, the research assistants used the created codebook to independently code another 200 randomly selected data entries. Simultaneously, we changed the prompt description of task to deductive coding and asked GPT-4 to code the same 200 data entries using the codes from the codebook. To minimize the impact of randomness in LLMs, we had GPT-4 perform three rounds of deductive coding on the 200 data entries, with each round being conducted independently. We also tested the deductive coding on the latest models (GPT-40 and Claude 3.5). The coding results and corresponding data indices from each round were also stored in a table for subsequent comparative analysis.
- 5.3.3 Results and Evaluation. Upon completing the coding process, we calculated the IRR between human coders and LLMs for the deductive coding task. The results indicated that the kappa value between human coders was approximately 0.73, reflecting substantial agreement. The Fleiss' Kappa value between human coders and GPT-4 ranged from 0.44 to 0.50, with an average of approximately 0.46, signifying moderate agreement. Among the three independent GPT-4 coding results, the Fleiss' Kappa value was approximately 0.87, demonstrating almost perfect agreement. Additionally,

we tested newer models, GPT-40 and Claude 3.5, which showed Fleiss' Kappa values with human coders ranging from 0.38 to 0.42, indicating moderate agreement. Additionally, we restarted the LLM process and randomly selected 50 data entries, along with their codes, which had been agreed upon by the researchers. These entries were used as prior knowledge and provided to the LLM (GPT-4) through prompting. The LLM then performed deductive coding on an additional 150 data entries. The results showed that the Fleiss' Kappa value between the LLM and the two human coders was approximately 0.46. However, because the kappa value between the two research assistants decreased to 0.64, the consistency between the LLM and the researchers slightly increased. Detailed results are presented in Tabl 1.

**Number of Coders** Kappa Valueα Index Type of Coding Coders **Inductive Coding** [Human coders], GPT-4 2 0.57 2 RA1, RA2 2 0.73 3 RA1, RA2, GPT-4 3 0.44 - 0.50**Deductive Coding** GPT-4 (s) 3 0.87 5 RA1, RA2, GPT-4o 3 0.38 6 RA1, RA2, Claude 3.5 3 0.42 Deductive Coding with prior knowledge RA1, RA2, GPT-4 3

0.42

Table 1. IRR for Inductive and Deductive Coding Across Various Coders

Overall, in the deductive coding phase, although the LLM did not surpass human researchers in consistency, the results still demonstrated significant potential. Notably, when we used different LLMs for coding, the IRR among the LLMs was very high. This indicates that LLMs perform more consistently than human coders when using a codebook for deductive coding.

#### 6 DISCUSSION

In recent times, the advent and evolution of LLMs such as GPT-3.5 Turbo and GPT-4 have opened up new avenues for automating tasks that were traditionally labor-intensive. One such task is the coding of qualitative data to derive thematic insights. Our tool, QualiGPT, leverages the capabilities of LLMs through prompt design and API calls to automate this coding process, offering a list of potential themes. This integrated tool significantly reduces the overhead associated with manual coding, addressing challenges encountered in traditional qualitative analysis and when using ChatGPT.

Specifically, QualiGPT employs prompts that have been validated in prior research [93], offering researchers an efficient means of categorizing themes in qualitative data. The prompts are highly structured, mitigating risks associated with using GPT for analysis, such as inconsistencies and lack of transparency. Compared to traditional qualitative analysis methods or software, the computational prowess of LLMs ensures that QualiGPT outperforms conventional software's auto-coding features in terms of performance. Furthermore, its coding speed far surpasses manual coding while maintaining a quality comparable to expert groups. This tool has the potential to revolutionize the paradigm of qualitative analysis in the future. In this section, we delve into the contributions and prospects of this tool, especially in terms of collaboration.

 $<sup>\</sup>alpha$  Round to two decimal places

## 6.1 Reflections on LLM-Assisted Qualitative Coding

Our study provides valuable insights into the potential of LLMs in qualitative research, particularly in the realm of inductive and deductive coding. The findings from our analysis and verification process reveal several key points worthy of discussion.

6.1.1 Human Consistency vs. LLM Potential. Our results demonstrate a high level of consistency between human coders, with Cohen's Kappa values reaching 0.73 in deductive coding tasks. This underscores the current gold standard in qualitative analysis. However, it's crucial to note the promising performance of LLMs, particularly GPT-4, in these same tasks. While not yet matching human-level consistency, GPT-4 achieved moderate agreement with human coders, with Fleiss' Kappa values ranging from 0.44 to 0.50. This performance is particularly noteworthy given the complexity of qualitative coding tasks. Typically, as the number of independent coders increases, consistency tends to decrease significantly in the beginning. Therefore, the current stage of IRR should be considered acceptable. Additionally, in multi-coder tasks, low consistency is not a severe issue because consensus can be reached through multiple rounds of group discussions and negotiations. During these discussions, different coders usually need to articulate their viewpoints, which helps foster mutual understanding and enhance the depth of analysis. In future work, if researchers use LLMs for qualitative coding, we recommend more interaction with the LLM. For example, researchers could ask the LLM to provide detailed considerations and explanations for the coded content and share their thought processes during coding with the LLM. Currently, such discussions with LLMs are largely constrained by the interaction method (through text prompts). However, in the future, this could be achieved through multimodal interactions [77, 89] and enhanced context memory and understanding in subsequent LLM versions. Additionally, as LLM continues to evolve rapidly, we anticipate significant improvements in their coding capabilities. The consistently high agreement between different LLM iterations (Fleiss' Kappa of 0.87) suggests that as these models become more sophisticated, they may approach or even surpass human-level consistency in qualitative coding tasks. In the future, we may need to pay more attention to the depth of analysis provided by LLMs in coding.

6.1.2 Practical Application of LLMs in Qualitative Analysis. Unlike previous studies that have explored high-level applications of LLMs in qualitative research (such as generating broad themes or categories), our work delves into the practical application of LLMs for specific coding tasks. We demonstrate the capability of LLMs to perform both inductive and deductive coding, core processes in qualitative analysis. In inductive coding, where no prior codebook exists, GPT-4 showed promising results with a Cohen's Kappa of 0.57 when compared to human coders. This suggests that LLMs can effectively identify emergent themes and patterns in qualitative data, a task traditionally requiring significant human expertise and intuition. For deductive coding, where a predefined codebook is used, GPT-4's performance was also encouraging. As we mentioned above, the moderate agreement with human coders indicates that LLMs can effectively apply predetermined codes to new data, a crucial skill in many qualitative research projects. Additionally, during the coding process, researchers need to repeatedly review the data and use methods such as note-taking to reinforce their familiarity with the data context, codebook, and specific coding content. This practice enhances consistency in the coding process but also poses a challenge to the researchers' expertise. This challenge is particularly pronounced when coding large datasets over extended periods (e.g., several weeks) and when multiple rounds of group coding are required, as well as when coding segmented data (e.g., interviews with Q&A or social media data). Variations in the data and new information emerging during the coding process may affect the consistency of the coders' perspectives. Although the effectiveness of LLMs in coding all data types is not yet clear, our work demonstrates the capability of LLMs to code more structured data. This capability allows researchers to process large amounts of data more quickly and has the potential to identify content that human coders may overlook.

# 6.2 QualiGPT as a Tool: Leveraging QualiGPT to Augment Efficiency in Qualitative Analysis

A primary concern among researchers regarding GPT-generated content stems from a lack of confidence in its accuracy [63]. There have been instances where GPT has been found to fabricate content, generating spurious information. Such behavior is unequivocally unacceptable in scientific research. However, when used as an auxiliary tool, these concerns can be significantly alleviated. In other words, when used as a tool, QualiGPT merely offers perspectives on the data, while the mechanism for human review remains intact. Under this modality, researchers can utilize QualiGPT for rapid coding. Specifically, they can select themes of interest from the generated responses and, aided by the justifications provided by QualiGPT (explanations and references to the original data), manually verify these themes. In this scenario, the researcher or user retains control over the accuracy of the results, with the final decision-making power remaining human-centric.

## 6.3 QualiGPT as a Collaborative Researcher

Qualitative analysis often carries a degree of subjectivity, which is typically viewed as an advantage [26, 66], allowing for unique insights to be gleaned from the data [52]. Concurrently, this subjectivity can lead to varied interpretations of the same qualitative data by different researchers. In traditional analysis workflows, discussions between co-researchers to reconcile coding results and reach a consensus are indispensable. Building on this procedural concept, we pondered the possibility of incorporating QualiGPT as an independent co-researcher in studies.

Under this new paradigm, both human researchers and LLM would analyze the qualitative data independently. Once the analyses are completed, the results from both the human researchers and LLM would be collated for a collective discussion, aiming to achieve consensus among all parties. Indeed, LLM appears to possess the potential to facilitate such a collaborative model, as it can generate several high-quality themes, providing genuine content references from the original text for each theme. Additionally, the advanced version of the LLM supports voice interaction, which introduces new opportunities for qualitative analysis in collaboration with AI. This enhancement makes the LLM function more like a research participant [16]. In this context, LLM should not merely be perceived as a tool assisting researchers but rather as an independent contributor, offering insights into the data and actively participating in discussions.

### 6.4 Challenges and Considerations

Despite the promising results, it's crucial to approach the use of LLMs in qualitative analysis with caution. Several challenges and considerations emerge from our findings: (1) **Consistency Across LLM Versions:** While we observed high consistency between different iterations of GPT-4, the performance varied when testing other models like GPT-40 and Claude 3.5. This highlights the importance of model selection and the need for researchers to validate results across different LLM versions. However, although our results show that newer versions of LLMs have lowered the IRR with human coders, this does not necessarily mean we should avoid using the updated models. This is because IRR is not the sole criterion for assessing the quality of analysis in qualitative research, and discrepancies in IRR can be addressed through discussion and iteration. Additionally, various test results [3, 55, 87] indicate that newer LLMs generally exhibit better reasoning abilities, speed, accuracy, and a larger knowledge base, supporting more diverse content (different languages, multimodal data, etc.). The issue of lower IRR with newer models might be due to LLMs producing more varied interpretations of the data, which can be further explained through discussions about inconsistent coding with the LLM.

Additionally, the prompts we used may not be perfect, potentially not fully activating the LLM's capabilities or affecting its judgment. To address this, we have retained a customizable prompt window in QualiGPT to support more tailored needs. (2) Codebook Development: Our results showed improved performance when using a predefined codebook for deductive coding. This underscores the continued importance of human expertise in developing and refining coding frameworks, even as LLMs take on more coding tasks. Also, when we attempted to provide more contextual information (prior knowledge) for GPT to learn from. However, while the results showed a slight improvement, it was not significant. This suggests that such learning may need to be built on more complex prompt-guided interactions, such as enabling the LLM to gain a more detailed understanding of the codebook development process. However, at present, this process may be complicated due to the challenge of (3) Human-AI Communication. Unlike human coding teams who can meet to discuss inconsistencies and reach consensus, current LLM implementations lack this interactive capability. Developing tools and methodologies for effective human-AI communication in the coding process remains a challenge. On one hand, the lack of more human-like interaction methods required for more complex interactions is a challenge. In human communication, people can exchange and understand large amounts of information in a short time through speaking, listening, reading, and writing. However, in current LLM interactions, typing (writing) prompts via a keyboard remains the primary mode of interaction, which significantly hinders the efficiency of information transfer. Hence, in the future, enabling LLMs to participate in discussions and negotiations through voice interaction could greatly assist in establishing prior knowledge for LLMs. On the other hand, understanding context is a crucial aspect of effective human communication. Currently, LLMs' understanding of context does not yet reach a human-like level [94], which can result in LLMs "forgetting" previous content during the communication process, thereby affecting their task performance. However, trends in LLM iterations show that newer models are progressively improving their memory of context. This is one of the key reasons why we remain open to using the latest versions of LLMs. (4) Ethical Considerations: Although QualiGPT was developed based on previous research findings and conceptualizations, and we believe it has achieved a high degree of usability and user-friendliness in addressing certain practical issues, it is not without flaws. This is particularly important to note given that LLM-assisted collaborative coding is still in its early stages. Our use of APIs is also aimed at strengthening ethical and policy considerations, which aligns with the consistent goals of CSCW. (5) Validation and Oversight: While LLMs show promise as coding assistants, human validation and oversight remain crucial. Researchers should view LLMs as tools to augment, not replace, human analysis in qualitative research. An innovative direction is to have LLMs critique their own content. This concept involves multiple LLMs process analyzing and debating perspectives, with a human-in-the-loop for final review and decision-making.

#### 7 CONCLUSION

The realm of qualitative research, while invaluable for its depth and nuance, has long grappled with the challenges of data analysis, particularly during the coding phase. Traditional qualitative analysis software, despite their merits, often fall short in addressing the complexities, costs, and performance demands of modern research. This study has illuminated a promising avenue for the future of qualitative analysis through the integration of LLMs, specifically GPT and its API, into the research workflow.

Our introduction of QualiGPT represents a significant stride forward in addressing the longstanding challenges in qualitative data analysis. By identifying and addressing the common issues associated with ChatGPT, we have not only enhanced the efficiency of the coding process but also bolstered the credibility and transparency of using LLMs in qualitative research. The comparative analysis between QualiGPT and manual coding underscores the tool's potential in streamlining the workflow, reducing processing costs, and ensuring a more transparent and credible analysis process.

Furthermore, the design considerations of QualiGPT, with its emphasis on usability and user-friendliness, mark a departure from the often cumbersome interfaces of traditional qualitative software. By offering a more intuitive interface, QualiGPT significantly diminishes the learning and usage overheads, making it an attractive option for both seasoned researchers and those in the early stages of their careers.

In light of our findings, it is evident that the integration of LLMs like ChatGPT into qualitative research holds substantial promise. As technology continues to evolve, it is imperative for the academic community to remain adaptive and open to such innovations, ensuring that research methodologies are not only rigorous but also efficient and user-centric. With tools like QualiGPT, we are one step closer to realizing this vision, ushering in a new era of qualitative research that marries depth with efficiency. Future work should continue to refine and expand upon these tools, ensuring they remain relevant and effective in the ever-evolving landscape of qualitative research.

#### **ACKNOWLEDGMENTS**

#### A MORE FIGURES

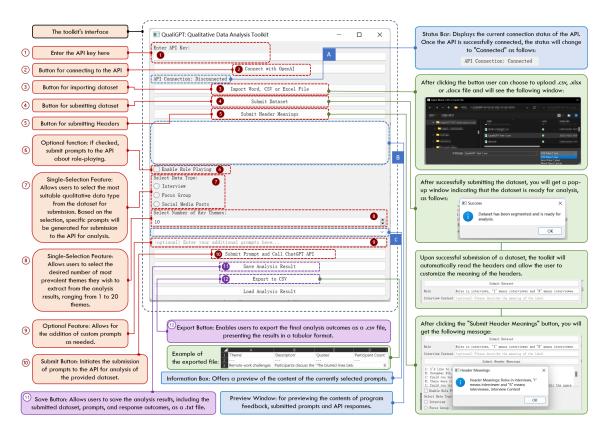


Fig. 2. User Manual for QualiGPT (A Qualitative Analysis Toolkit) - Interactive Features. QualiGPT offers a total of 13 interactive features that users can select, click, or input text into. The functionalities enclosed by the red boxes are related to invoking the API, while the interactive features shown in the purple boxes do not involve API calls.

#### **B** ONLINE RESOURCES

QualiGPT is open-sourced on Github (https://github.com/KindOPSTAR/QualiGPT).

#### **REFERENCES**

- Ida Madieha Abdul Ghani Azmi, Sonny Zulhuda, and Sigit Puspito Wigati Jarot. 2012. Data breach on the critical information infrastructures: Lessons from the Wikileaks. In Proceedings Title: 2012 International Conference on Cyber Security, Cyber Warfare and Digital Forensic (CyberSec). 306–311. https://doi.org/10.1109/CyberSec.2012.6246173
- [2] Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 15, 2 (2023), 1-4. https://doi.org/10.7759/cureus.35179
- [3] Anthropic. 2024. Claude 3.5: Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet Accessed: 2024-07-02.
- [4] Theophilus Azungah. 2018. Qualitative research: deductive and inductive approaches to data analysis. *Qualitative research journal* 18, 4 (2018), 383–400.
- [5] P. Bazeley. 2013. Qualitative Data Analysis: Practical Strategies. SAGE Publications. https://books.google.com/books?id=33BEAgAAQBAJ
- [6] Michael J Belotto. 2018. Data analysis methods for qualitative research: Managing the challenges of coding, interrater reliability, and thematic analysis. The qualitative report 23, 11 (2018), 2622–2633.
- [7] Michael Bergin. 2011. NVivo 8 and consistency in data analysis: Reflecting on the use of a qualitative data analysis program. Nurse researcher 18, 3 (2011). https://doi.org/10.7748/nr2011.04.18.3.6.c8457
- [8] Lea Bishop. 2023. A computer wrote this paper: What chatgpt means for education, research, and writing. Research, and Writing (January 26, 2023) (2023). https://doi.org/10.2139/ssrn.4338981
- [9] Jie Cai, Ya-Fang Lin, He Zhang, and John M. Carroll. 2024. Third-Party Developers and Tool Development For Community Management on Live Streaming Platform Twitch. In Proceedings of the CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 926, 18 pages. https://doi.org/10.1145/3613904.3642787
- [10] Yanto Chandra and Liang Shang. [n. d.]. Computer-Assisted Qualitative Research: An Overview. In Qualitative Research Using R: A Systematic Approach, Yanto Chandra and Liang Shang (Eds.). Springer Nature, 21–31. https://doi.org/10.1007/978-981-13-3170-1\_2
- [11] Yanto Chandra, Liang Shang, Yanto Chandra, and Liang Shang. [n. d.]. An Overview of R and RQDA: An Open-Source CAQDAS Platform. ([n. d.]),
- [12] Bonnie Chinh, Himanshu Zade, Abbas Ganji, and Cecilia Aragon. 2019. Ways of Qualitative Coding: A Case Study of Four Strategies for Resolving Disagreements. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (Glasgow, Scotland Uk) (CHI EA '19). Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312879
- [13] Muhammad Faisol Chowdhury. 2015. Coding, sorting and sifting of qualitative data analysis: Debates and discussion. Quality & Quantity 49, 3 (2015), 1135–1143. https://doi.org/10.1007/s11135-014-0039-2
- [14] Amanda Coffey, Holbrook Beverley, and Atkinson Paul. 1996. Qualitative Data Analysis: Technologies and Representations. Sociological Research Online 1, 1 (1996), 80–91. https://doi.org/10.5153/sro.1 arXiv:https://doi.org/10.5153/sro.1
- [15] Jingfeng Cui, Zhaoxia Wang, Seng-Beng Ho, and Erik Cambria. 2023. Survey on sentiment analysis: evolution of research methods and topics. Artificial Intelligence Review 56 (2023), 8469—8510. https://doi.org/10.1007/s10462-022-10386-z
- [16] Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? Trends in Cognitive Sciences 27, 7 (2023), 597–600. https://doi.org/10.1016/j.tics.2023.04.008
- [17] Victoria Elliott. 2018. Thinking about the coding process in qualitative data analysis. Qualitative report 23, 11 (2018).
- [18] Helen Elliott-Mainwaring. 2021. Exploring using NVivo software to facilitate inductive coding for thematic narrative synthesis. British Journal of Midwifery 29, 11 (2021), 628–632. https://doi.org/10.12968/bjom.2021.29.11.628
- [19] Yunhe Feng, Sreecharan Vanam, Manasa Cherukupally, Weijian Zheng, Meikang Qiu, and Haihua Chen. 2023. Investigating Code Generation Performance of ChatGPT with Crowdsourcing Social Data. In 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC). 876–885. https://doi.org/10.1109/COMPSAC57700.2023.00117
- [20] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. International journal of qualitative methods 5, 1 (2006), 80-92.
- [21] Alexander J. Fiannaca, Chinmay Kulkarni, Carrie J Cai, and Michael Terry. [n. d.]. Programming without a Programming Language: Challenges and Opportunities for Designing Developer Tools for Prompt Programming. In Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg Germany, 2023-04-19). ACM, 1-7. https://doi.org/10.1145/3544549.3585737
- [22] Jane Forman and Laura Damschroder. 2007. Qualitative content analysis. In Empirical methods for bioethics: A primer. Emerald Group Publishing Limited, 39–62. https://doi.org/10.1016/S1479-3709(07)11003-7
- [23] Andrew Gao. [n. d.]. Prompt Engineering for Large Language Models. https://doi.org/10.2139/ssrn.4504303
- [24] Jie Gao, Kenny Tsu Wei Choo, Junming Cao, Roy Ka-Wei Lee, and Simon Perrault. 2023. CoAlcoder: Examining the Effectiveness of AI-assisted Human-to-Human Collaboration in Qualitative Analysis. ACM Trans. Comput.-Hum. Interact. 31, 1, Article 6 (nov 2023), 38 pages. https://doi.org/10.1145/3617362

- [25] Jie Gao, Yuchen Guo, Toby Jia-Jun Li, and Simon Tangi Perrault. 2023. CollabCoder: A GPT-Powered WorkFlow for Collaborative Qualitative Analysis. In Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing (Minneapolis, MN, USA) (CSCW '23 Companion). Association for Computing Machinery, New York, NY, USA, 354–357. https://doi.org/10.1145/3584931.3607500
- [26] Lucia Garcia and Francis Quek. 1997. Qualitative research in information systems: time to be subjective?. In Information Systems and Qualitative Research: Proceedings of the IFIP TC8 WG 8.2 International Conference on Information Systems and Qualitative Research, 31st May-3rd June 1997, Philadelphia, Pennsylvania, USA. Springer, 444-465. https://doi.org/10.1007/978-0-387-35309-8\_22
- [27] Robert P. Gauthier and James R. Wallace. 2022. The Computational Thematic Analysis Toolkit. Proc. ACM Hum.-Comput. Interact. 6, GROUP, Article 25 (jan 2022), 15 pages. https://doi.org/10.1145/3492844
- [28] Simret Araya Gebreegziabher, Zheng Zhang, Xiaohang Tang, Yihao Meng, Elena L. Glassman, and Toby Jia-Jun Li. 2023. PaTAT: Human-AI Collaborative Qualitative Coding with Explainable Interactive Rule Synthesis. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 362, 19 pages. https://doi.org/10.1145/3544548.3581352
- [29] Lisa Given. 2024. The SAGE Encyclopedia of Qualitative Research Methods. https://doi.org/10.4135/9781412963909
- [30] Nahid Golafshani. 2003. Understanding reliability and validity in qualitative research. The qualitative report 8, 4 (2003), 597-607.
- [31] L Suzanne Goodell, Virginia C Stage, and Natalie K Cooke. 2016. Practical qualitative research strategies: Training interviewers and coders. Journal of nutrition education and behavior 48, 8 (2016), 578–585. https://doi.org/10.1016/j.jneb.2016.06.001
- [32] Ulla H Graneheim, Britt-Marie Lindgren, and Berit Lundman. 2017. Methodological challenges in qualitative content analysis: A discussion paper. Nurse education today 56 (2017), 29–34. https://doi.org/10.1016/j.nedt.2017.06.002
- [33] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: A Case Study. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 433, 19 pages. https://doi.org/10.1145/3544548.3580688
- [34] Thomas Hansson, Greg Carey, and Rafn Kjartansson. 2010. A multiple software approach to understanding values. Journal of Beliefs & Values 31, 3 (2010), 283–298. https://doi.org/10.1080/13617672.2010.521005
- [35] Mubin Ul Haque, Isuru Dharmadasa, Zarrin Tasnim Sworna, Roshan Namal Rajapakse, and Hussain Ahmad. 2022. "I think this is the most disruptive technology": Exploring Sentiments of ChatGPT Early Adopters using Twitter Data. arXiv:2212.05856 [cs.CL]
- [36] Hossein Hassani and Emmanuel Sirmal Silva. 2023. The role of ChatGPT in data science: how ai-assisted conversational interfaces are revolutionizing the field. Big data and cognitive computing 7, 2 (2023), 62. https://doi.org/10.3390/bdcc7020062
- [37] Vonna L Hemmler, Allison W Kenney, Susan Dulong Langley, Carolyn M Callahan, E Jean Gubbins, and Shannon Holder. 2022. Beyond a coefficient: An interactive process for achieving inter-rater consistency in qualitative coding. Qualitative Research 22, 2 (2022), 194–219. https://doi.org/10.1177/1468794120976072
- [38] Thomas F. Heston and Charya Khun. [n. d.]. Prompt Engineering in Medical Education. 2, 3 ([n. d.]), 198–205. Issue 3. https://doi.org/10.3390/ime2030019
- [39] Judith A Holton. 2007. The coding process and its challenges. The Sage handbook of grounded theory 3 (2007), 265–289. http://digital.casalini.it/9781446275726
- [40] Jialun Aaron Jiang, Kandrea Wade, Casey Fiesler, and Jed R. Brubaker. 2021. Supporting Serendipity: Opportunities and Challenges for Human-AI Collaboration in Qualitative Analysis. Proc. ACM Hum.-Comput. Interact. 5, CSCW1, Article 94 (apr 2021), 23 pages. https://doi.org/10.1145/3449168
- [41] Mahmut Kalman. 2019. "It requires interest, time, patience and struggle": Novice researchers' perspectives on and experiences of the qualitative research journey. Qualitative Research in Education 8, 3 (2019), 341–377. https://doi.org/10.17583/qre.2019.4483
- [42] Brianna L Kennedy and Robert Thornberg. 2018. Deduction, induction, and abduction. The SAGE handbook of qualitative data collection (2018), 49–64. http://digital.casalini.it/9781526416063
- [43] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2023. Natural language processing: State of the art, current trends and challenges. Multimedia tools and applications 82, 3 (2023), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4
- [44] Michael Liebrenz, Roman Schleifer, Anna Buadze, Dinesh Bhugra, and Alexander Smith. 2023. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. The Lancet Digital Health 5, 3 (2023), e105–e106. https://doi.org/10.1016/S2589-7500(23)00019-5
- [45] Yao Lyu, He Zhang, Shuo Niu, and Jie Cai. 2024. A Preliminary Exploration of YouTubers' Use of Generative-AI in Content Creation. In Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24). Association for Computing Machinery, New York, NY, USA, Article 20, 7 pages. https://doi.org/10.1145/3613905.3651057
- [46] Liye Ma and Baohong Sun. 2020. Machine learning and AI in marketing Connecting computing power to human insights. International Journal of Research in Marketing 37, 3 (2020), 481–504. https://doi.org/10.1016/j.ijresmar.2020.04.005
- [47] Calum Macdonald, Davies Adeloye, Aziz Sheikh, and Igor Rudan. 2023. Can ChatGPT draft a research article? An example of population-level vaccine effectiveness analysis. Journal of global health 13 (2023). https://doi.org/10.7189/jogh.13.01003
- [48] Kathleen M MacQueen, Eleanor McLellan, Kelly Kay, and Bobby Milstein. 1998. Codebook development for team-based qualitative analysis. Cam Journal 10, 2 (1998), 31–36. https://doi.org/10.1177/1525822X98010002030
- [49] Megh Marathe and Kentaro Toyama. 2018. Semi-Automated Coding for Qualitative Research: A User-Centered Inquiry and Initial Prototypes. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173922

[50] T. Mather, S. Kumaraswamy, and S. Latif. 2009. Cloud Security and Privacy: An Enterprise Perspective on Risks and Compliance. O'Reilly Media. https://books.google.com/books?id=BHazecOuDLYC

- [51] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. Biochemia medica 22, 3 (2012), 276-282.
- [52] Haradhan Kumar Mohajan et al. 2018. Qualitative research methodology in social sciences and related subjects. Journal of economic development, environment and people 7, 1 (2018), 23–48. https://mpra.ub.uni-muenchen.de/id/eprint/85654
- [53] Muhammad Naeem, Wilson Ozuem, Kerry Howell, and Silvia Ranfagni. 2023. A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. *International Journal of Qualitative Methods* 22 (2023), 16094069231205789. https://doi.org/10.1177/16094069231205789
- [54] Helen Noble and Joanna Smith. 2015. Issues of validity and reliability in qualitative research. Evidence-based nursing 18, 2 (2015), 34–35. https://doi.org/10.1136/eb-2015-102054
- [55] OpenAI. 2024. Hello GPT-4O. https://openai.com/index/hello-gpt-4o/ Accessed: 2024-07-02.
- [56] Cliodhna O'Connor and Helene Joffe. 2020. Intercoder reliability in qualitative research: debates and practical guidelines. International journal of qualitative methods 19 (2020), 1609406919899220. https://doi.org/10.1177/1609406919899220
- [57] Geetanjali Panda, Ashwani Kumar Upadhyay, and Komal Khandelwal. 2019. Artificial intelligence: A strategic disruption in public relations. Journal of Creative Communications 14, 3 (2019), 196–213. https://doi.org/10.1177/0973258619866585
- [58] Trena Paulus, Jessica Lester, and Paul Dempster. [n. d.]. Digital Tools for Qualitative Research. SAGE. googlebooks:ZgZPAgAAQBAJ
- [59] Noel Pearse. 2019. An illustration of deductive analysis in qualitative research. In 18th European conference on research methodology for business and management studies. 264.
- [60] Margarita S Peredaryenko and Steven Eric Krauss. 2013. Calibrating the human instrument: Understanding the interviewing experience of novice qualitative researchers. The qualitative report 18, 43 (2013), 1.
- $[61] \ \ Margaret \ Phillips \ and \ Jing \ Lu. \ [n. d.]. \ A \ Quick \ Look \ at \ NVivo. \ 30, 2 \ ([n. d.]), 104-106. \ \ https://doi.org/10.1080/1941126X.2018.1465535$
- [62] F. Beryl Pilkington. [n. d.]. The Use of Computers in Qualitative Research. 9, 1 ([n. d.]), 5-7. https://doi.org/10.1177/089431849600900103
- [63] Russell A Poldrack, Thomas Lu, and Gašper Beguš. 2023. AI-assisted coding: Experiments with GPT-4. arXiv:2304.13187 [cs.AI]
- [64] Judith Preissle. 2006. Envisioning qualitative inquiry: a view across four decades. International Journal of Qualitative Studies in Education 19, 6 (2006), 685–695. https://doi.org/10.1080/09518390600975701 arXiv:https://doi.org/10.1080/09518390600975701
- [65] Rémi Rampin and Vicky Rampin. [n. d.]. Taguette: Open-Source Qualitative Data Analysis. 6, 68 ([n. d.]), 3522.
- [66] Carl Ratner et al. 2002. Subjectivity and objectivity in qualitative methodology. In Forum Qualitative Social forschung/Forum: Qualitative Social Research, Vol. 3. https://doi.org/10.17169/fqs-3.3.829
- [67] renaissancerachel. [n. d.]. 15 Best Qualitative Data Analysis Software of 2023. Renaissance Rachel. https://renaissancerachel.com/best-qualitative-data-analysis-software/
- [68] Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. 1–7.
- [69] Tim Rietz and Alexander Maedche. 2021. Cody: An Al-Based System to Semi-Automate Coding for Qualitative Research. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 394, 14 pages. https://doi.org/10.1145/3411764.3445591
- [70] Rashmi R Salavi, Mallikarjun M Math, and UP Kulkarni. 2019. A survey of various cryptographic techniques: From traditional cryptography to fully homomorphic encryption. In *Innovations in Computer Science and Engineering: Proceedings of the Sixth ICICSE 2018*. Springer, 295–305. https://doi.org/10.1007/978-981-13-7082-3 34
- [71] J. Saldana. 2015. The Coding Manual for Qualitative Researchers. SAGE Publications. https://books.google.com/books?id=jh1iCgAAQBAJ
- [72] Suprateek Sarker, Xiao Xiao, and Tanya Beaulieu. 2013. Guest Editorial: Qualitative Studies in Information Systems: A Critical Review and Some Guiding Principles. MIS Quarterly 37, 4 (2013), iii—xviii. http://www.jstor.org/stable/43825778
- [73] Thanveer Shaik, Xiaohui Tao, Yan Li, Christopher Dann, Jacquie McDonald, Petrea Redmond, and Linda Galligan. 2022. A Review of the Trends and Challenges in Adopting Natural Language Processing Methods for Education Feedback Analysis. IEEE Access 10 (2022), 56720–56739. https://doi.org/10.1109/ACCESS.2022.3177752
- [74] Yiqiu Shen, Laura Heacock, Jonathan Elias, Keith D Hentel, Beatriu Reig, George Shih, and Linda Moy. 2023. ChatGPT and other large language models are double-edged swords. , e230163 pages. https://doi.org/10.1148/radiol.230163
- [75] Sruti Srinivasa Ragavan, Zhitao Hou, Yun Wang, Andrew D Gordon, Haidong Zhang, and Dongmei Zhang. 2022. GridBook: Natural Language Formulas for the Spreadsheet Grid. In 27th International Conference on Intelligent User Interfaces (Helsinki, Finland) (IUI '22). Association for Computing Machinery, New York, NY, USA, 345–368. https://doi.org/10.1145/3490099.3511161
- [76] Heather L Stuckey. 2015. The second step in data analysis: Coding qualitative research data. Journal of Social Health and Diabetes 3, 01 (2015), 007–010. https://doi.org/10.4103/2321-0656.140875
- [77] Yan Tai, Weichen Fan, Zhao Zhang, and Ziwei Liu. 2024. Link-Context Learning for Multimodal LLMs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 27176–27185.
- [78] David R Thomas. 2003. A general inductive approach for qualitative data analysis. (2003).
- [79] Sara Thunberg and Linda Arnell. 2022. Pioneering the use of technologies in qualitative research—A research review of the use of digital interviews. International Journal of Social Research Methodology 25, 6 (2022), 757–768. https://doi.org/10.1080/13645579.2021.1935565

- [80] Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C. Comeau, Rezarta Islamaj, Aadit Kapoor, Xin Gao, and Zhiyong Lu. [n. d.]. Opportunities and Challenges for ChatGPT and Large Language Models in Biomedicine and Health. https://doi.org/10.48550/arXiv.2306.10070 arXiv:2306.10070 [cs, q-bio]
- [81] Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed., Cham: Springer International Publishing 10, 3152676 (2017), 10–5555. https://doi.org/10.1007/978-3-319-57959-7
- [82] Shuyue Wang and Pan Jin. [n. d.]. A Brief Summary of Prompting in Using GPT Models. ([n. d.]). https://doi.org/10.32388/IMZI2Q
- [83] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. [n. d.]. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. 35 ([n. d.]), 24824–24837. https://proceedings.neurips.cc/paper\_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html
- [84] E. Weitzman and M.B. Miles. 1995. Computer Programs for Qualitative Data Analysis. SAGE Publications. https://books.google.com/books?id= E4Y5DQAAQBAJ
- [85] Elaine Welsh et al. 2002. Dealing with data: Using NVivo in the qualitative data analysis process. In Forum qualitative sozialforschung/Forum: qualitative sozial research, Vol. 3. https://doi.org/10.17169/fqs-3.2.865
- [86] Gareth Wiltshire and Noora Ronkainen. 2021. A realist approach to thematic analysis: making sense of qualitative data through experiential, inferential and dispositional themes. Journal of Critical Realism 20, 2 (2021), 159–180. https://doi.org/10.1080/14767430.2021.1894909
- [87] Chunqiu Steven Xia, Yinlin Deng, and Lingming Zhang. 2024. Top Leaderboard Ranking = Top Coding Proficiency, Always? EvoEval: Evolving Coding Benchmarks via LLM. arXiv:2403.19114 [cs.SE] https://arxiv.org/abs/2403.19114
- [88] Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In Companion Proceedings of the 28th International Conference on Intelligent User Interfaces (Sydney, NSW, Australia) (IUI '23 Companion). Association for Computing Machinery, New York, NY, USA, 75–78. https://doi.org/10. 1145/3581754.3584136
- [89] Jingyi Xie, Rui Yu, He Zhang, Sooyeon Lee, Syed Masum Billah, and John M. Carroll. 2024. Emerging Practices for Large Multimodal Model (LMM) Assistance for People with Visual Impairments: Implications for Design. arXiv:2407.08882 [cs.HC] https://arxiv.org/abs/2407.08882
- [90] Himanshu Zade, Margaret Drouhard, Bonnie Chinh, Lu Gan, and Cecilia Aragon. 2018. Conceptualizing Disagreement in Qualitative Coding. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3173574.3173733
- [91] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 1–21.
- [92] He Zhang, Chuhao Wu, Jingyi Xie, ChanMin Kim, and John M. Carroll. 2023. QualiGPT: GPT as an easy-to-use tool for qualitative coding. arXiv:2310.07061 [cs.HC] https://arxiv.org/abs/2310.07061
- [93] He Zhang, Chuhao Wu, Jingyi Xie, Yao Lyu, Jie Cai, and John M. Carroll. 2023. Redefining Qualitative Analysis in the AI Era: Utilizing ChatGPT for Efficient Thematic Analysis. arXiv:2309.10771 [cs.HC]
- [94] He Zhang, Jingyi Xie, Chuhao Wu, Jie Cai, ChanMin Kim, and John M. Carroll. 2024. The Future of Learning: Large Language Models through the Lens of Students. arXiv:2407.12723 [cs.HC] https://arxiv.org/abs/2407.12723
- [95] Yan Zhang and Barbara M Wildemuth. 2009. Qualitative analysis of content. Applications of social research methods to questions in information and library science 308, 319 (2009), 1–12.
- [96] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. [n. d.]. Calibrate Before Use: Improving Few-shot Performance of Language Models. In Proceedings of the 38th International Conference on Machine Learning (2021-07-01). PMLR, 12697-12706. https://proceedings.mlr.press/ v139/zhao21c.html