Human-Computer Interaction & Emerging Technologies, Vol. 195, 2025, 142–152
https://doi.org/10.54941/ahfe1006232

Exploring Inductive and Deductive Qualitative Coding With AI: Investigating Inter-Rater Reliability Between Large Language Model and Human Coders

He Zhang[1], Chuhao Wu[1], Jingyi Xie[1], Fiona Rubino[2], Sydney Graver[2], Jie Cai[3], ChanMin Kim[4], and John M. Carroll[1]

[1]College of Information Sciences and Technology, Pennsylvania State University, University Park, PA 16802, USA
[2]College of Engineering, Pennsylvania State University, University Park, PA 16802, USA
[3]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[4]College of Education, Pennsylvania State University, University Park, PA 16802, USA

ABSTRACT

Qualitative research provides valuable insights into complex human phenomena, but its coding processes are often time-intensive and labor-intensive. The advent of Large Language Models (LLMs) has introduced new opportunities to streamline qualitative analysis. This study investigates the application of LLMs in both inductive and deductive coding tasks using real-world datasets, assessing their ability to complement traditional coding methods. To address challenges such as privacy concerns, prompt customization, and integration with qualitative workflows, we developed QualiGPT, an API-based tool that facilitates efficient and secure qualitative coding. Our evaluation shows that the consistency level between AI-generated codes and human coders is acceptable, particularly for inductive coding tasks where themes are identified without prior frameworks. In our case study using data from a Discord community, GPT-4 achieved a Cohen's Kappa of 0.57 in inductive coding, demonstrating moderate agreement with human coders. For deductive coding, the inter-rater reliability between human coders and GPT-4 reached a Fleiss' Kappa of 0.46, indicating a promising level of consistency when applying pre-established codebooks. These findings highlight the potential of LLMs to augment qualitative research by improving efficiency and consistency while maintaining the contextual depth that human researchers provide. We also observed that LLMs demonstrated higher internal consistency compared to human coders when using a codebook for deductive coding, suggesting their value in standardizing coding approaches. Additionally, we explored a novel paradigm where LLMs function not merely as coding tools but as collaborative co-researchers that independently analyze data alongside humans. This approach leverages LLMs' strengths in generating high-quality themes and providing genuine content references, thereby enriching researchers' insights while maintaining human oversight to ensure contextual understanding and ethical standards. Nevertheless, challenges remain regarding prompt engineering, domain-specific training, and the risk of fabricated information, underscoring the importance of human validation in the final analysis. This research advances human-AI collaboration in qualitative methods by exploring AI-assisted coding and highlighting future improvements in interaction design.

Keywords: Large language model, Prompt engineering, Qualitative analysis, Inductive coding, Deductive coding, Inter-rater reliability, Analytical evaluation

INTRODUCTION

Qualitative research offers a unique perspective into individuals' comprehension, attitudes, and insights regarding technology, phenomena, and specific topics (Preissle, 2006). Over time, an increasing number of researchers have acknowledged the significance of qualitative methodologies across diverse fields. In the HCI (Human-Computer Interaction) field, qualitative research methods are particularly important because they provide deep insights into how people collaborate with and through technology. 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, researchers face the challenge of efficiently processing and analyzing large volumes of qualitative data. Analyzing such data can be labor-intensive (Zhang et al., 2023), 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 (Saldana, 2015). As the production of qualitative data continues to surge, there is an escalating demand for innovative techniques to streamline and enhance the analysis process (Bazeley, 2013).

To address these challenges, researchers have employed qualitative analysis software with computer-assisted collaboration to boost data management and efficiency (Elliott-Mainwaring, 2021). Since the release of GPT-3 in 2022 and later versions, LLMs have sparked a technological revolution. We explored leveraging these models for qualitative analysis to enhance efficiency and accuracy, choosing OpenAI's ChatGPT and its API for broad applicability.

We developed QualiGPT[1], an integrated tool using API access and prompt engineering, to overcome challenges in data privacy, prompting, and workflow. We applied an LLM to a real dataset and compared its performance with manual coding in inductive and deductive tasks using inter-rater reliability (IRR) (McHugh, 2012). The LLM achieved acceptable agreement with human coders, especially in inductive coding where natural theme emergence highlights AI's potential to boost coding efficiency while maintaining contextual depth and ethical oversight. This work contributes to the discourse on human-AI collaboration in qualitative research and outlines future directions for LLM-assisted coding.

RELATED WORK

Qualitative research is a vital approach to understanding human experiences, rooted in subjectivity and practical interactivity (Preissle, 2006) and widely used across disciplines. Rigorous data analysis is essential to this method (Sarker et al., 2013). In this section, we briefly review common qualitative

[1] https://github.com/KindOPSTAR/QualiGPT

coding processes and challenges, examining both manual and technology-assisted approaches.

Manual Qualitative Coding

Coding is the most crucial part of qualitative research. 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 (Zhang et al., 2023). 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.

Inductive Coding

Inductive coding is a widely used qualitative method that derives themes directly from data without relying on a predefined codebook (Thomas, 2003; Forman and Damschroder, 2007). Human coders play a crucial role by identifying emerging patterns and key concepts, like open coding, which often requires expert guidance (Fereday and Muir-Cochrane, 2006; Naeem et al., 2023; Wiltshire and Ronkainen, 2021). This process is challenging for novices due to the complexity of qualitative data (Kalman, 2019), the need for reflective skills like memo-writing (Stuckey, 2015), effective management of iterative processes (Noble and Smith, 2015), and a collaborative mindset to avoid over-interpretation.

Deductive Coding

Deductive coding is a key qualitative method where annotators apply a pre-established framework (MacQueen et al., 1998; Fereday and Muir-Cochrane, 2006) by assigning codes from a codebook to data segments (Goodell et al., 2016). While this approach requires less expertise in uncovering implicit patterns than inductive coding, it demands a deeper understanding of each category's connotations (Azungah, 2018).

Consistency is equally crucial in deductive coding, and maintaining coding consistency becomes a significant challenge (O'Connor and Joffe, 2020). 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 (Zade et al., 2018; Chowdhury, 2015). 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 (Pearse, 2019). 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 (Chinh et al., 2019; Lyu et al., 2024; Cai et al., 2024).

AI-Assisted Qualitative Coding

The combination of AI and qualitative research has begun to redefine how researchers approach qualitative data and analysis (Haque et al., 2022; Xiao et al., 2023). Technologies, especially AI algorithms, provide potential for improved efficiency in analyzing large datasets. 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 (Feng et al., 2023). 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 (Marathe and Toyama, 2018). 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 (Cui et al., 2023; Khurana et al., 2023; Shaik et al., 2022). By comparing AI findings with human analysis, researchers can increase the validity and reliability of their findings (Gebreegziabher et al., 2023).

PROCEDURE

In this study, we adapted prompts from previous research (Gao et al., 2023b; Xiao et al., 2023) that cover Task Background, Task Description, Processing Method, and Expected Output. The customizable prompts allow expert role-playing and data type selection (e.g., social media), and we controlled analysis depth by specifying the number of key themes and adding instructions for inductive coding to ensure comparability with human-generated themes. Using GPT's default temperature (0.7), we submitted these prompts for analysis. To evaluate LLM performance, we applied them to a dataset of 1,000 public Discord posts (excluding sensitive user data) and compared the resulting codes with those from manual coding to assess IRR in both inductive and deductive processes.

Case Study One - Inductive Coding by Using LLM

In Case Study 1, we randomly selected 200 data entries for inductive coding using both manual and LLM-assisted methods. For the manual approach, topic modeling on the raw dataset yielded 8 topics; we then chose two topics and randomly extracted 100 entries each. Two research assistants independently labeled each entry with short tags. Since the dataset is anonymous, the first round of coding took several hours, with the full process, including discussions, lasting nearly a week.

For inductive coding, by using LLM, we removed the manually coded labels from the data and submitted it to LLM for analysis, and we designed prompts that allow LLM to explore the data that was processed in the manual inductive coding section and generate corresponding codes without prior knowledge. Specifically, we interact with LLM through prompts. First, we employ role-playing to activate LLM’s capabilities, such as “*You are now an excellent qualitative data analyst and qualitative research expert.*” Then, we inform it about the required task and provide the task background, for instance, “*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 input data format, 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.*” If GPT provides an incorrect format or is not processing the task correctly, we regenerate the output.

Case Study Two - Deductive Coding by Using LLM

Codebook Development

Before deductive coding, we developed a codebook based on inductive results from Case 1. Two research assistants compared and merged labels for each entry, forming an initial codebook of 171 labels. With input from the senior researcher, the team reviewed and refined this codebook, removing irrelevant labels, to produce a final set of 54 labels (with label 0 for irrelevant topics and label 53 for relevant but unspecified topics) for use in the deductive coding process.

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 the task to deductive coding and asked GPT-4 to code the same 200 data entries using the code 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 have modified the task description section in the prompts as follows: “*For each line in the dataset, assign the most appropriate code(s) from the codebook. If multiple codes apply, list all relevant codes. Return the results in a table format with the following columns: ‘|Original Text | Assigned Code(s)|’ Start the table with* ‘**********’ *and end with* ‘**********’*. Do not include a header row in the output table.*” We also tested the deductive coding on the latest models (GPT-4o and Claude 3.5). The coding results and corresponding data indices from each round were also stored in a table for subsequent comparative analysis.

Results and Evaluation

Table 1: IRR for inductive and deductive coding across various coders.

Index	Type of Coding	Coders	Num of Coders	Kappa Value[a]
1	Inductive Coding	[Human coders], GPT-4	2	0.57
2	Deductive Coding	RA1, RA2	2	0.73
3		RA1, RA2, GPT-4	3	0.44–0.50
4		GPT-4(s)	3	0.87
5		RA1, RA2, GPT-4o	3	0.38
6		RA1, RA2, Claude 3.5	3	0.42

[a] Round to two decimal places

After coding, we calculated IRR for the deductive coding task. Human coders achieved a Kappa of about 0.73 (substantial agreement). With GPT-4, Fleiss' Kappa ranged from 0.44–0.50 (averaging 0.46, moderate agreement), while three independent GPT-4 coding results exhibited nearly perfect agreement at around 0.87. The newer models, GPT-4o and Claude 3.5, each recorded Fleiss' Kappa values between 0.38–0.42 (moderate agreement). Detailed results are shown in Table 1.

DISCUSSION

In recent times, the advent and evolution of LLMs such as GPT-3.5 Turbo and GPT-4 have opened new avenues for automating tasks that were traditionally labor-intensive. In this section, we further discuss the AI's role and collaboration in qualitative analysis.

LLMs as Tools for Inductive and Deductive Coding

Our study explores the practical applications of LLMs in both inductive and deductive qualitative coding tasks, demonstrating their potential to enhance traditional methodologies. In inductive coding without a predefined codebook, GPT-4 demonstrated an acceptable, moderate level of agreement with human coders. This balance allows the model to generate meaningful themes without simply mimicking human judgment. This suggests that LLMs can effectively identify emergent themes and patterns within qualitative data, a process that typically demands significant human expertise and intuition (Fereday and Muir-Cochrane, 2006). This level of reliability indicates that LLMs can serve as valuable complementary tools, enhancing the efficiency of qualitative analysis while still allowing for the depth and contextual insights that human researchers provide. On the other hand, in deductive coding with a predefined codebook, GPT-4 exhibited moderate performance, indicating that LLMs can effectively apply predetermined codes—a key asset for qualitative research (MacQueen et al., 1998). Qualitative analysis inherently involves a degree of subjectivity, allowing researchers to derive unique insights from the data (Garcia and Quek, 1997). However, this subjectivity can result in varied interpretations of the same dataset by different researchers, necessitating discussions and consensus-building among co-researchers (Saldana, 2015). LLMs, when employed as coding tools, offer a means to process large volumes of data swiftly and consistently, potentially reducing the labor-intensive nature of manual coding (Zhang et al., 2023). While LLMs have not yet matched human-level consistency, their ability to handle structured data efficiently suggests they can identify content that might be overlooked by human coders, thereby augmenting the overall coding process (Poldrack et al., 2023).

LLMs as Collaborative Co-Researchers in Qualitative Analysis

Building upon their role as coding tools, our research investigates the novel paradigm of integrating LLMs as independent co-researchers within qualitative studies. In this framework, human researchers and LLMs independently analyze the qualitative data. Subsequently, the results from both parties are collated for collective discussion to achieve consensus. This collaborative model leverages the strengths of LLMs in generating high-quality themes and providing genuine content references from the original text, thereby enriching the researcher's insights (Haque et al., 2022).

The subjectivity inherent in qualitative analysis, while advantageous for deriving unique insights, can lead to inconsistent interpretations among different coders (Garcia and Quek, 1997). By incorporating LLMs as co-researchers, researchers can mitigate these inconsistencies through the amalgamation of human and machine-generated analyses. Additionally, advanced LLMs with capabilities such as voice interaction open new avenues for more interactive and dynamic collaboration, positioning LLMs not just as tools but as active participants in the research process.

Despite these advancements, challenges remain. While acceptable, the current stage of IRR for LLMs highlights the need for ongoing refinement in prompt engineering and domain-specific training (Reynolds and McDonell, 2021; Gao, 2023; Wei et al., 2022). Furthermore, concerns regarding the accuracy of LLM-generated content, such as the potential for fabricating information, underscore the necessity of maintaining human oversight in the coding process (Poldrack et al., 2023). When used as independent co-researchers, LLMs can facilitate rapid coding and offer diverse perspectives, but the final validation and decision-making remain firmly in the hands of human researchers to ensure the integrity and reliability of the qualitative analysis (Elliott, 2018).

CONCLUSION

This study demonstrates GPT-4's potential to enhance qualitative research by effectively supporting both inductive and deductive coding processes, aligning well with human judgment. The development of QualiGPT improved workflow integration and efficiency through an API-based approach. However, since the study was conducted before newer models were released and local configuration options were explored, some performance aspects remain unexamined.

ACKNOWLEDGMENT

This work was supported by the Center for Socially Responsible Artificial Intelligence (CSRAI) at Pennsylvania State University. This work is part of the "Optimizing Large-Scale Language Model-Based AI Integration and Human-Computer Interaction in Educational Scenarios" project, funded by the Big Ideas Grant (BIG) Summer 2023, CSRAI.

REFERENCES

Alexander J. Fiannaca, Chinmay Kulkarni, Carrie J Cai, and Michael Terry. 2023. 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, pages 1–7. ACM.

Andrew Gao. 2023. Prompt Engineering for Large Language Models.

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, CHI EA '19, page 1–6, New York, NY, USA. Association for Computing Machinery.

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.

Cliodhna O'Connor and Helene Joffe. 2020. Inter-coder reliability in qualitative research: Debates and practical guidelines. International journal of qualitative methods, 19:1609406919899220.

David R Thomas. 2003. A general inductive approach for qualitative data analysis.

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):3713–3744.

Elaine Welsh et al. 2002. Dealing with data: Using nvivo in the qualitative data analysis process. In Forum qualitative sozialforschung/Forum: qualitative social research, volume 3.

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):159–180.

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.

Heather L Stuckey. 2015. The second step in data analysis: Coding qualitative research data. Journal of Social Health and Diabetes, 3(01):007–010.

Helen Elliott-Mainwaring. 2021. Exploring using nvivo software to facilitate inductive coding for thematic narrative synthesis. British Journal of Mid-wifery, 29(11):628–632.

Helen Noble and Joanna Smith. 2015. Issues of validity and reliability in qualitative research. Evidence-based nursing, 18(2):34–35.

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, CHI '18, page 1–11, New York, NY, USA. Association for Computing Machinery.

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):62.

Hussam Alkaissi and Samy I McFarlane. (2023). Artificial hallucinations in chatgpt: implications in scientific writing. Cureus, 15(2):1–4.

J. Saldana. 2015. The Coding Manual for Qualitative Researchers. SAGE Publications.

Jane Forman and Laura Damschroder. 2007. Qualitative content analysis. In Empirical methods for bioethics: A primer, pages 39–62. Emerald Group Publishing Limited.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. 35:24824–24837.

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, pages 1–21.

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):80–92.

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, CHI '24, New York, NY, USA. Association for Computing Machinery.

Jie Gao, Kenny Tsu Wei Choo, Junming Cao, Roy Ka-Wei Lee, and Simon Perrault. 2023a. Coaicoder: Examining the effectiveness of AI-assisted human-to-human collaboration in qualitative analysis. ACM Trans. Comput.-Hum. Interact., 31(1), nov.

Jie Gao, Yuchen Guo, Toby Jia-Jun Li, and Simon Tangi Perrault. 2023b. Collabcoder: A gpt-powered workflow for collaborative qualitative analysis. In Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing, CSCW '23 Companion, page 354–357, New York, NY, USA. Association for Computing Machinery.

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: 8469–8510.

Judith Preissle. 2006. Envisioning qualitative inquiry: A view across four decades. International Journal of Qualitative Studies in Education, 19(6):685–695.

Kathleen M MacQueen, Eleanor McLellan, Kelly Kay, and Bobby Milstein. 1998. Codebook development for team-based qualitative analysis. Cam Journal, 10(2):31–36.

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):578–585.

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, pages 1–7.

Lea Bishop. (2023). A computer wrote this paper: What chatgpt means for education, research, and writing. Research, and Writing (January 26, 2023).

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, pages 444–465. Springer.

Mahmut Kalman. 2019. “It requires interest, time, patience and struggle”: Novice researchers’ per-spectives on and experiences of the qualitative research journey. Qualitative Research in Education, 8(3):341–377.

Mary L McHugh. 2012. Interrater reliability: The kappa statistic. Biochemia medica, 22(3):276–282.

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, CHI ‘18, page 1–12, New York, NY, USA. Association for Computing Machinery.

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): e105–e106.

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.

Muhammad Faisol Chowdhury. 2015. Coding, sorting and sifting of qualitative data analysis: Debates and discussion. Quality & Quantity, 49(3):1135–1143.

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, 22:16094069231205789.

Noel Pearse. 2019. An illustration of deductive analysis in qualitative research. In 18th European conference on research methodology for business and management studies, page 264.

Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. Proc. ACM Hum.-Comput. Interact., 3(CSCW), November.

P. Bazeley. (2013). Qualitative Data Analysis: Practical Strategies. SAGE Publications.

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, CHI ‘23, New York, NY, USA. Association for Computing Machinery.

Russell A Poldrack, Thomas Lu, and Gasper Begus. 2023. AI-assisted coding: Experiments with gpt-4.

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. 2023. Opportunities and Challenges for ChatGPT and Large Language Models in Biomedicine and Health.

Shuyue Wang and Pan Jin. 2023. A Brief Summary of Prompting in Using GPT Models.

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, CHI ’23, New York, NY, USA. Association for Computing Machinery.

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): iii–xviii.

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, 10:56720–56739.

Theophilus Azungah. 2018. Qualitative research: Deductive and inductive approaches to data analysis. Qualitative research journal, 18(4):383–400.

Thomas F. Heston and Charya Khun. 2023. Prompt Engineering in Medical Education. 2(3):198–205.

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, New York, NY, USA. Association for Computing Machinery.

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.

Yunhe Feng, Sreecharan Vanam, Manasa Cheruku-pally, 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), pages 876–885.

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, IUI '23 Companion, page 75–78, New York, NY, USA. Association for Computing Machinery.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-shot Performance of Language Models. In Proceedings of the 38th International Conference on Machine Learning, pages 12697–12706. PMLR.