

LATA: A Pilot Study on LLM-Assisted Thematic Analysis of Online Social Network Data Generation Experiences

QILE WANG, University of Delaware, USA

MOATH ERQSOUS, University of Delaware, USA

KENNETH E. BARNER, University of Delaware, USA

MATTHEW LOUIS MAURIELLO, University of Delaware, USA

Large Language Models (LLMs) have gained attention in research and industry, aiming to streamline processes and enhance text analysis performance. Thematic Analysis (TA), a prevalent qualitative method for analyzing interview content, often requires at least two human experts to review and analyze data. This study demonstrates the feasibility of LLM-Assisted Thematic Analysis (LATA) using GPT-4 and Gemini. Specifically, we conducted semi-structured interviews with 14 researchers to gather insights on their experiences generating and analyzing Online Social Network (OSN) communications datasets. Following Braun and Clarke's six-phase TA framework with an inductive approach, we initially analyzed our interview transcripts with human experts. Subsequently, we iteratively designed prompts to guide LLMs through a similar process. We compare and discuss the manually analyzed outcomes with responses generated by LLMs and achieve a cosine similarity score up to 0.76, demonstrating a promising prospect for LATA. Additionally, the study delves into researchers' experiences navigating the complexities of collecting and analyzing OSN data, offering recommendations for future research and application designers.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI); Human computer interaction (HCI);** • **Computing methodologies** → **Artificial intelligence;** • **Software and its engineering;**

Additional Key Words and Phrases: Online Social Networks, Data Collection, Thematic Analysis, Inductive Coding, Large Language Models (LLMs), ChatGPT, Google Gemini

ACM Reference Format:

Qile Wang, Moath Erqsous, Kenneth E. Barner, and Matthew Louis Mauriello. 2025. LATA: A Pilot Study on LLM-Assisted Thematic Analysis of Online Social Network Data Generation Experiences. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW124 (April 2025), 28 pages. <https://doi.org/10.1145/3711022>

1 Introduction

Leveraging new technology has the potential to provide valuable insights and enable more in-depth analysis beyond traditional human capabilities. Large Language Models (LLMs) are extensive neural network-based Machine Learning (ML) models trained on massive amounts of internet-based text. These models are designed to comprehend and generate text in a human-like way. For example, Generative Pre-trained Transformer (GPT) refers to a series of LLMs, including ChatGPT¹ which

¹<https://chat.openai.com/>

Authors' Contact Information: Qile Wang, University of Delaware, USA, kylewang@udel.edu; Moath Erqsous, University of Delaware, USA, merqsous@udel.edu; Kenneth E. Barner, University of Delaware, USA, barner@udel.edu; Matthew Louis Mauriello, University of Delaware, USA, mlm@udel.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM 2573-0142/2025/4-ARTCSCW124

<https://doi.org/10.1145/3711022>

was primarily developed for conversational response and was first released in 2022 by OpenAI². With an estimated over 100 million monthly active users, it became the fastest-growing app in history [52]. Following its official release, ChatGPT has attracted considerable attention and can be applied to a broad range of applications, such as translating text, debugging code, summarizing documents, and scheduling appointments. Zhao et al. undertook an extensive literature review, summarizing the available resources for developing LLMs and exploring their applications in healthcare, education, finance, law, and research [73]. With the industry attention captured, Google introduced an LLM named Bard in March 2023, which was later rebranded to Gemini³. Some LLMs such as BERT [19], Alpaca [61], LLAMA [63], etc, have been made open source. While GPT demonstrates strong capabilities, LLMs like ChatGPT are proprietary technologies. The transparency of model parameters and pre-trained data may be lacking. In addition, several publications have highlighted its inherent limitations, such as lack of understandability, incorrect or nonsensical response (i.e., hallucinations), and ethical concerns [37, 40, 72]. These concerns may include potential biases in responses, privacy breaches, sustainability concerns, and the risk of misuse.

Given that LLMs are language models capable of analyzing textual data, we aim to investigate their potential to assist with qualitative analysis, particularly in the context of thematic analysis (TA) for semi-structured interviews. We call this approach LLM-Assisted Thematic Analysis (LATA). Specifically, TA was first introduced by Braun & Clark and is a method used in qualitative research to identify, analyze, and report patterns from a dataset such as interviews or written materials [8]. Their six-phase framework has become a widely used and valuable method to understand the perspectives and experiences of individuals and communities [9]. Nevertheless, performing TA may require considerable manual efforts and resources in terms of time, expertise, and specific software for managing and analyzing qualitative data. Kiger & Varpio highlight that findings from TA may be unconvincing or unsupported due to overlapping themes, inconsistency, and lack of outstanding data extracts [34]. Moreover, different researchers may have various findings and interpretations [45]. Therefore, we are motivated to explore the possibilities, techniques, and limitations through a detailed pilot study using LLMs. Unfortunately, interview data annotated using the 6-phase framework is not widely available. Thus, we gathered our own data and conducted TA. As a pilot study, we selected the topic of data generation experiences from Online Social Networks (OSNs) as it has some known themes or challenges that should surface in our analysis alongside novel insights driven by our participants' experiences. Existing challenges include managing large and diverse data, handling noise and heterogeneous content, adapting to dynamic information in OSNs, and creating efficient data mining algorithms for social interaction analysis [54]. As early work, our questions are exploratory and include:

- RQ1: How do researchers collect data from OSNs, and how can we enhance it?
- RQ2: How do the outcomes of LATA compare to those obtained via human experts?
- RQ3: How effectively can LLMs assist in the inductive TA process, and what are some limitations?

To answer these questions, we collected questionnaires, including demographic information, and then conducted a need-finding interview case study to gather firsthand opinions from researchers about OSN data collection experiences. Using the 6 phases TA framework proposed by Braun and Clarke [8], we first analyzed the interview content manually. We then deliberately instructed GPT-4 to conduct a similar process using our interview transcriptions. We draw comparisons between GPT-driven and manual TA, discussing opportunities and challenges. We further experimented with TA using another leading LLM, Google Gemini 1.5 Pro and discussed the overall design

²<https://openai.com/>

³<https://gemini.google.com/>

implementation of LATA. The general method employed in this study is illustrated in Figure 1. Each component is discussed in subsequent sections. We observed moderate agreement between excerpts selected by GPT-4 and human raters (Cohen’s Kappa = 0.72). The final theme summaries showed LATA comparable to human analysis in our dataset, with cosine similarity scores ranging from 0.45 to 0.76 from GPT-4 and 0.37 to 0.54 from Gemini. The answers to these research questions offer insights for refining prompts and establishing benchmarks for automating TA. In addition, from the interview content, we identify existing obstacles in OSN data collection and discuss strategies to enhance the user experience of next-generation data collection systems.

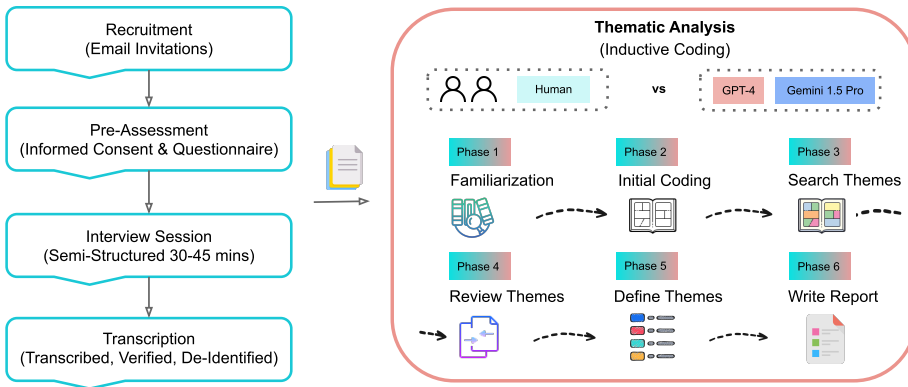


Fig. 1. Overall methods used in this study

2 Related Work

We discuss qualitative analysis using LLMs and survey current works on data collection from OSNs. We acknowledge that Twitter rebranded to X.com in July 2023; however, as our research commenced prior to this change, and our discussions with participants often referenced the platform as Twitter and its content as Tweets, we will continue to use these terms for consistency.

2.1 LLMs Prompts

The input query for LLMs, known as the prompt, is a textual input in the form of instructions, questions, or other data. A prompt is typically crafted by a person and is expressed in human language. Andrej Karpathy, former Tesla head of AI, tweeted that “The hottest new programming language is English” [33]. Responses from LLMs are greatly influenced by the specific prompt given. Strobelt et al. note that identifying a suitable prompt for new NLP tasks requires multiple trials [60]. They also emphasized that variation in prompt templates and wording could result in significant differences in terms of accuracy [60]. Therefore, prompt engineering is becoming essential for effective communication with LLMs [65].

Prompt engineering often involves designing and refining prompts to ensure LLMs can comprehend and produce desired outcomes by in-context learning. Several strategies exist for crafting prompts. For example, the Chain-of-thought (CoT) prompting assists LLMs in reasoning by breaking down steps. This approach is particularly effective when providing a few examples (i.e., few-shot learning) [64]. Further, it has been observed that in the Zero-shot (i.e., no example given) CoT approach, accuracy is increased substantially by adding the phrase “Let’s think step by step” [35]. The author focuses on only logical reasoning, exemplified by addressing mathematical problems [35].

In contrast, Arvidsson & Axell conducted interview studies and reviewed 28 related publications to offer guidelines for prompt engineering for tasks or categories in reasoning, summarizing, persona, disambiguation, etc. [4]. A prompt example in the persona setting would be, “I want you to act as the system.” A prompt example could start with “Within this scope” to avoid disambiguation. The authors recommend providing context to all prompts to prevent hallucination in the output using open-ended prompts and adding more context tokens to prompts, which results in a more detailed output. Regarding using specific keywords, the author also recommends emphasizing subject and style keywords rather than focusing on connecting words [4]. To our knowledge, no prompt recommendations are aimed explicitly at TA. Therefore, we designed and improved prompts to produce a detailed TA and offer suggestions for this and similar use cases.

2.2 Qualitative Analysis using LLMs

Large Language Models (LLMs) like ChatGPT are gaining popularity for their ability to accelerate many information tasks, including evidence that leveraging LLMs could assist with different kinds of qualitative analysis. For example, ChatGPT appears proficient at various NLP tasks, including data augmentation, sentiment analysis, and text classification [29]. Alizadeh et al. compared text annotation techniques on news articles and Twitter data using open-source LLMs and Amazon Mechanical Turk (MTurk) workers in various learning settings [2]. The authors discovered that GPT-3 generally outperforms MTurk workers in binary and multi-classification tasks, providing compelling evidence for utilizing GPT as a practical alternative for text annotation. Similarly, Tornberg examined GPT-4 could achieve higher accuracy and lower bias than human classifiers in binary classification on a publicly available Twitter dataset in Zero-shot learning [62]. In other qualitative analyses, LLMs have been explored in evaluating or providing scoring for text data [11, 39]. Shen et al. highlighted that the rapid development of LLMs has significantly impacted multiple aspects of social computing research, including data collection, generation, and analysis; however, there is currently an absence of well-defined guidelines for conducting such research [56].

Thematic analysis (TA) is commonly used in qualitative research to extract meaningful insights from textual data. The main frameworks can be broken down into 6 phases: familiarization, initial coding, searching themes, reviewing themes, defining themes, and writing reports [8]. Very few studies have explored the use of LLMs to assist in TA. One commonly employed method for coding in TA is the deductive approach, in which researchers prepare and provide predefined codes with definitions. A preliminary study used GPT-3.5 to perform deductive coding on four public datasets and found that the level of agreement is comparable to human-only coders performing the same tasks [14]. They also analyzed that human coders took much longer than GPT [14]. Similarly, Gao et al. developed a web-based tool using GPT-3 to make code suggestions from excerpts and discovered that employing this tool could reduce the workload of individual coding and improve mutual comprehension [25]. In another study, Xiao et al. found that GPT-3 using provided code definition and examples (i.e., few-shot) has better agreement with human experts [68]. Their Cohen’s Kappa score reached 0.61 for two categories but decreased to 0.38 when considering four categories. These studies have shown that LLMs can be valuable tools for supporting deductive coding in qualitative analysis.

Another common coding approach in TA is the inductive method, where the codebook is not provided prior. Paoli employed a similar 6-phase methodology to generate personas from 14 public interview data with GPT-3 but focused on only two themes, achieving an acceptable level of quality [18]. Mirroring this study, Paoli tested GPT-3 (3.5-Turbo) on two open-source interview datasets and found GPT-3 could accurately identify or infer the majority of the themes [17]. Despite some of the GPT-3 responses being inconsistent, the author suggests that using LLMs could serve as an additional coder for verifying human analysis.

Motivated by Paoli’s concept, Dai et al. designed a human–LLM collaboration framework for TA and tested it on two other public datasets using mixed methods (i.e., inductive and deductive approaches) [15]. The authors suggest that employing GPT-3 with only one human coder can achieve performance comparable to that of two human coders in TA. Furthermore, Zhang et al. collected perspectives on LLMs qualitative analysis and then used the web-based ChatGPT for streamlined TA [71]. The authors offered several recommendations for crafting prompts in qualitative analysis. However, we know no case study has been conducted to investigate all six phases of TA using only LLMs. Considering the present literature, we utilized GPT-4 for conducting TA and applied it to our interview transcripts using inductive coding in a Zero-shot learning environment. We then compared each phase of the six-phase process proposed by Braun and Clarke [8] in detail.

To further test the feasibility of LATA, we experimented with Google Gemini, another cutting-edge LLM with the potential to revolutionize areas such as education [32] and health [38]. A series of recent studies have compared the performance of GPT and Google Gemini in various applications. Qi et al. found that GPT excels in precision, while Gemini provides detailed and expansive answers [48]. Another study demonstrates Gemini is exceptional in recall compared to GPT-4 for spam email detection [41]. Our work adds to this body of work by exploring how these features of LLMs will influence results in TA.

2.3 OSN Data Generation

We opted to conduct our own interview and focus on researchers’ experiences with collecting data from Online Social Networks (OSNs). To give some background, OSNs are applications that create online communities where people can share various data types such as text, images, and videos with other users [55]. The most used platforms are Twitter, Facebook, Instagram, and Reddit [30]. In recent years, data from OSNs has been widely used toward research efforts in various fields such as social sciences [54], business [20, 24], health [3, 6, 7], education [28], and transportation [50]. In a meta-review of 229 studies, Snelson describes an increasing trend in OSN research, many using quantitative and qualitative methods [57]. Utilizing natural language processing (NLP) techniques for systematically analyzing qualitative data could reduce both cost and time [1]. However, collecting and analyzing data from large social platforms can be challenging, particularly for researchers without a software engineering background. Existing qualitative approaches to dataset generation are often manual, time-consuming, expensive, and challenging to reproduce.

While several tools and methods are available for collecting data from OSNs, it remains a difficult task. Researchers may be able to search existing datasets from Google Dataset Search ⁴ and download them directly from open data sources such as Kaggle ⁵, WHO ⁶, UCI ⁷, CKAN ⁸, and Linked Open Data Cloud ⁹. The most direct way to collect OSN data is using Application Programming Interfaces (APIs) and their associate packages, such as the Python library Tweepy ¹⁰ for collecting Twitter data. Similarly, the Python package “quoras” can be used to collect multi-language data from Quora [16]. Many OSNs offer APIs that allow researchers to access data in a structured format. This can be a quick and efficient way to collect data, but it may be subject to rate limits and often requires specific technical skills to access. When APIs are unavailable or limited, an alternative method involves web scraping, an automated process of extracting content from web pages. It can be applied to almost

⁴<https://datasetsearch.research.google.com/?ref=jrashford.com>

⁵<https://www.kaggle.com/>

⁶<https://www.who.int/data>

⁷<https://archive.ics.uci.edu/>

⁸<https://ckan.org/>

⁹<https://www.lod-cloud.net/>

¹⁰<https://www.tweepy.org/>

all web pages and is often more flexible than APIs [22]. Following OSN data collection, researchers commonly employ analytic techniques such as topic modeling, time series analysis, sentiment analysis, and network analysis [26, 31]. Stieglitz et al. highlighted challenges, such as acquiring high-quality data, identifying relevant topics, and presenting data meaningfully [59]. Kumar and Goswami discussed difficulties in finding certain data representations for content and network analysis [54]. During our interview, we aim to have a broader discussion with researchers across disciplines to understand the challenges they face when collecting data from OSNs, the methods they use to collect the data, and how they view future systems improving current practice.

3 Research Design

We conducted semi-structured interviews to generate data that would allow us to explore the possibility of using LLMs to assist with TA. This section outlines the study undertaken to gain insight into OSN data collection experiences. The study received approval from the Institutional Review Board of the affiliated university (Protocol # 2037012-1). Figure 1 illustrates an overview of methods used in this study.

3.1 Procedures

By sending email invitations to active faculty throughout the university, we recruited 14 researchers from diverse backgrounds during the Summer of 2023. Before the interview, each participant was asked to sign a consent form and complete an intake questionnaire. This intake questionnaire included basic background information and survey questions related to data collection from OSNs. Every study session had a scheduled duration of 45 minutes. The interview segment followed a semi-structured format, featuring a consistent set of initial questions and follow-up inquiries based on participant responses. The semi-structured approach allows us to explore each participant's experience, thereby gaining deeper insights into their process. The interviews were conducted using Zoom and video/audio recording with participants' permission. On average, these interviews lasted approximately 22.42 minutes ($SD = 5.94$). Two interview sessions were not digitally recorded and excluded from the duration calculation. Participants in this study did not receive compensation. After completing our initial set of interviews, we reviewed our data briefly as a research team and found that conversations were repeating information without new perspectives emerging. We decided to discontinue recruitment because we had reached saturation and additional recruitment was unlikely to provide new information or enhance the depth of our findings.

3.2 Participants

The majority of participants are affiliated with the same university. Only those with prior research experience using OSNs were eligible to participate in the study. The age of our participants ranged from 28 to 64 years old ($M = 41.43$, $SD = 11.33$). In terms of racial background, 10 participants identified as White, five as Asian, and one preferred not to identify. Regarding educational background, 10 (71.43%) held a doctoral or professional degree, while three (21.43%) had a master's degree. Table 1 summarizes all participants' demographic information and relevant project(s) discussed during the interview. Each research topic was selected by the participant through discussion with the interviewer to help ground conversation on their research experience involving OSNs. Gender distribution shows eight (57.14%) of the participants identified as male and six (42.86%) as female. Most participants, six (42.86%), have their primary research area in Social Sciences, including psychology, sociology, and economics. This is followed by four (28.57%) specializing in business and management. The majority of participants have substantial expertise in their respective fields. Notably, 11 (78.57%) participants have over five years of research experience.

Table 1. Summary of Participant Demographics and Research Topics

| ID | Gender | Research Domain | Yrs Expr. | Relevant Project(s) |
|-----|--------|----------------------------|-----------|--|
| P1 | Male | Engineering & Technology | 6-10 | Russian Twitterbots during 2016 election |
| P2 | Male | Business & Management | 6-10 | Analysis of Grubhub food delivery reviews |
| P3 | Male | Business & Management | 11-15 | Labor negotiations, ratification, and strike predictions |
| P4 | Female | Engineering & Technology | 20+ | Using Graphic caption with low vision people, and building AAC communication devices to help speech impairment |
| P5 | Male | Social Sciences | 1-5 | Trump's aggressive Twitter behavior |
| P6 | Male | Business & Management | 16-20 | Investigate the relationship between stock price and Twitter sentiments |
| P7 | Female | Social Sciences | 6-10 | Stereotype salience on Twitter |
| P8 | Female | Education | 20+ | Public perceptions of the scientific method in COVID-19 Tweets |
| P9 | Female | Social Sciences | 16-20 | Exploring public opinions on social media marketing |
| P10 | Male | Social Sciences | 11-15 | Geo-tagged Twitter data on food and water insecurity |
| P11 | Female | Social Sciences | 6-10 | Anti-Japanese and Korea boycott movement |
| P12 | Male | Business & Management | 16-20 | Analysis of firm performance in relation to cybersecurity breaches |
| P13 | Male | Social Sciences | 1-5 | Study of personal satisfaction & time spent online |
| P14 | Female | Health Sciences & Medicine | 1-5 | Fertility knowledge and emerging young adults |

3.3 Thematic Analysis

3.3.1 Human Expert Analysis. Digitally recorded interviews were transcribed using audio files via the online software Otter.ai.¹¹ The transcriptions were reviewed and corrected manually to ensure accuracy and quality. Personally identifiable information, if any, was then redacted from the transcriptions. Using an inductive approach [47], two researchers iteratively double-coded all 14 files using Dedoose¹² and resolved disagreements through discussions with the research team. Based on the initial interview questions and participant responses, we identified 51 codes and merged them into 13 themes detailed in Table 3. This table presents information on theme description, example excerpts extracted from the transcriptions, and basic statistics. For instance, the theme *Data Quality and Validation* contains 18 excerpts from 13 participants, and an example excerpt for this theme is from participant #2.

3.3.2 LLM-based Analysis. We aim to investigate whether traditional TA can be integrated into LLMs as a complementary method. We followed the same methodologies by Braun and Clarke [8] with an inductive approach, in which we asked GPT to open code for each transcription. We chose to utilize GPT-4, initially released on March 14, 2023 and is currently the most advanced model from OpenAI¹³. According to tech report [44], GPT-4 performs much better than GPT-3. We used GPT-4 API (model: gpt-4) over the web-based chat system because of its flexibility and customized functions. As a comparison, we also employed Google Gemini API (model: gemini-1.5-pro-latest) [51], which was the most advanced model offered by Google at the time of the experiment. In both models, one of the main parameters called *Temperature* controls the element of randomness of the output. This parameter can be set from 0 to 2, with a default value of 1. A lower temperature tends to focus on the most likely outcomes, producing more predictable and deterministic results. On the other hand, a higher temperature value leads to more varied and creative outputs, but it could also increase the risk of generating incoherent responses. Similar to many study approaches [13, 15, 69, 74], we set the *Temperature* = 0 to ensure reproducibility. It gives us a more predictable response, although some

¹¹<https://otter.ai/>

¹²<https://www.dedoose.com/>

¹³<https://openai.com/>

variability may still persist. For example, one study shows that using a temperature setting of 0 yields the best performance when using ChatGPT for code refinement [27]. Our interview scripts have not been published online. Therefore, our data is unseen by the pre-trained LLM models. This will help avoid overfitting issues and prevent LLMs responses from being memorized. Responses from LLMs are highly sensitive to each prompt, so we have employed an iterative design approach, refining the prompt for each phrase until we achieve an acceptable response. Regarding ethical considerations, OpenAI asserts that data sent through the API will no longer be utilized to enhance its model as of March 2023. Google also asserts that Gemini does not utilize prompts or its generated responses as training data for its models. Finally, we compared the response from GPT-4 with manual TA regarding accuracy, consistency, and depth of understanding. Moreover, we assessed the level of agreement between excerpts selected and labeled by humans and GPT. We then computed cosine similarities for the final reports. For simplicity, we refer to GPT-4 as GPT in subsequent sections.

4 Results

4.1 Survey Results

In addition to demographic information from participants, we also included a few survey questions related to their experience in data collection from OSNs. As a supplement to the interview content, these survey results help us better understand how to improve OSN data collection.

Table 2. Ranking Orders using Dowdall Scores for Challenge and Feature Items

| Rank | Challenge Items | Score | Feature Items | Score |
|------|---|-------|---|-------|
| 1st | Lack of Data Access | 78.21 | Historical data retrieval | 65.77 |
| 2nd | Technical Challenges | 35.45 | User-friendly interface and navigation | 44.54 |
| 3rd | Managing time constraints | 32.09 | Advanced search filters and queries | 43.93 |
| 4th | Poor Data Quality | 31.28 | Real-time data monitoring and collection | 32.16 |
| 5th | Dealing with changes in social media platforms' algorithms and interfaces | 26.02 | Integration with multiple social media platforms | 24.94 |
| 6th | Dealing with potential biased content | 24.81 | Customizable data analytic, visualization and reporting | 22.29 |
| 7th | Dealing with large volumes of data and information overload | 22.40 | Data Anonymization | 19.48 |
| 8th | Ensuring data privacy and compliance with ethical guidelines | 17.58 | Collaborative Data Collection | 18.67 |
| 9th | Dealing with multilingual content | 15.06 | | |

4.1.1 CSAT & Challenges Rankings. One of the survey questions seeks to evaluate end-user or customer satisfaction (CSAT) with their existing data collection and analysis workflows, using a 1 to 10 scale where 10 represents the highest level of satisfaction. The responses are ranging from three to 10 ($M = 5.79$, $SD = 1.93$). To identify the current limitations in data collection from OSNs, we presented participants with nine predefined challenges with OSN data collection and analysis. We requested that participants rank these items according to their perceived difficulty. One represents the highest level of challenge, while nine indicates the least. The selection of most of these challenge items is designed based on relevant prior studies that have been identified. Among the top three most challenging issues, 12 (85.71%) chose "Lack of Data Access," followed by eight (57.14%) for "Technical Challenges," and six (42.86%) for "Poor Data Quality." Additionally, we apply the Dowdall method [21], a Borda-ranked voting system modification, to assess the overall significance of challenges in Table 2. The Dowdall scores range from 0 to 100. The higher the

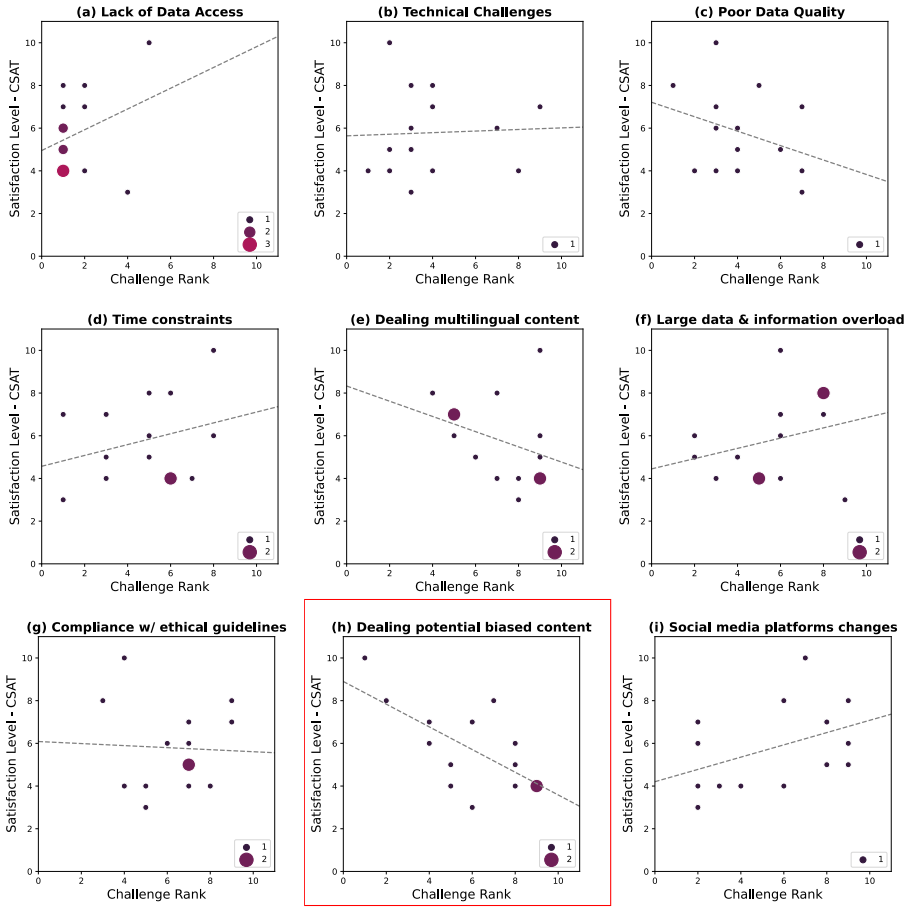


Fig. 2. Relationship between CSAT and Challenge Items Ranking. (On the X-axis, 1 signifies the most critical challenge)

number, the higher the priority. It appears that managing time constraints poses a slightly more significant challenge than dealing with data quality, albeit with a minor difference.

Figure 2 illustrates the correlation between CSAT and the rankings for each challenge item. The size and colors of the points indicate their occurrence. The red-highlighted sub-figure(h) has a statistically significant negative correlation ($r(12) = -.66, p < .01$). Figure 2 (e) “Dealing with multilingual content,” and (h) “Dealing potential biased content” tend to have low priority and less significance to participants who have lower CSAT scores. This may be understandable, as researchers might need to address other challenges before delving into how to handle potentially biased content. Conversely, subfigure (f) exhibits a slightly positive trend. Participants who express less satisfaction with their current workflow tend to perceive dealing with extensive data as a more significant challenge. Similarly, (d) “Managing time constraints” shows a positive trend, and (i) “Dealing with changes in social media platforms such as algorithm and interfaces” were ranked as being more challenging, corresponding with lower CSAT scores.

4.1.2 CSAT & Features Rankings. In addition to challenge ranking, we asked participants to rank eight predefined features they found valuable in a data collection tool from OSNs. The ranking 1

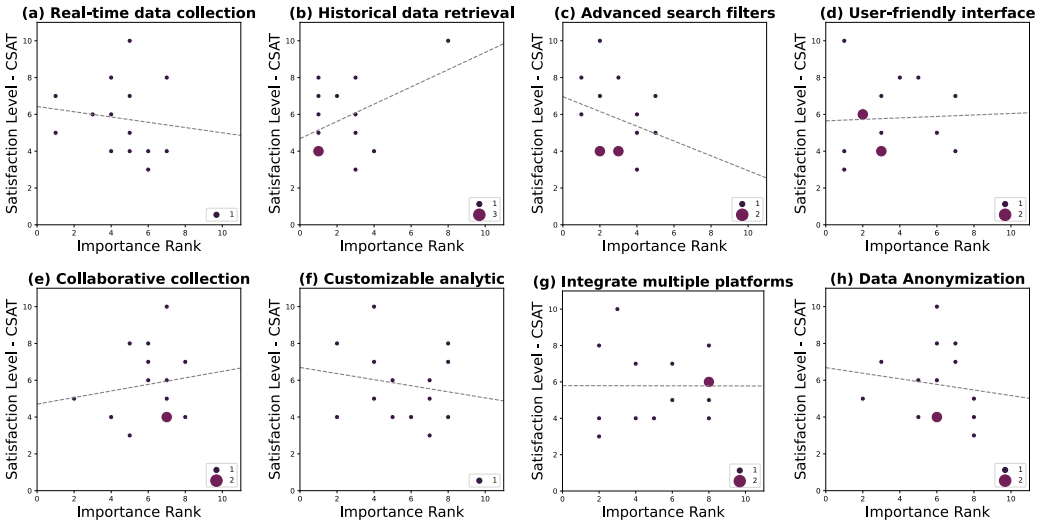


Fig. 3. Relationship between CSAT and Features Items Ranking. (On the X-axis, 1 signifies the most desired feature)

represents the most important to them, while 8 is the least important. We selected these features because they are crucial for establishing a comprehensive, efficient, and ethically responsible OSN data collection platform. Table 2 displays the average feature ranking as determined by the Dowdall method. Notably, having collaborative data collection ranks as the least desired feature.

Similarly, Figure 3 shows relationships between these feature items ranking and CSAT scores. Overall, having the “Historical data retrieval” feature is the most desired functionality ($M = 2.36$, $SD = 1.87$). Participants with lower CSAT scores consider it more important, as seen in sub-figure (b). The next desired feature is (c) “Advanced search filters and queries” ($M = 2.93$, $SD = 1.28$) with moderate negative relationships, where people with higher satisfaction find it more valuable. Third is (d) “User-friendly interface and navigation” ($M = 3.43$, $SD = 2.03$). On the contrary, participants assign low priority to features in (h) “Data Anonymization,” (e) “Collaborative Data Collection,” and (g) “Integration with multiple social media platforms.”

4.2 Thematic Analysis by Human

We provide a summary of 13 emerging themes. Table 1 has a comprehensive list of discussion topics covered in the participant interviews. This analysis section was completed prior to GPT and was not influenced by any responses generated by GPT.

4.2.1 Research Field and Background. We begin the interview by inviting participants to give a brief overview of their research backgrounds. Our diverse participants include finance, education, social science, business, and hospitality faculty members.

4.2.2 Research Questions and Motivations. Our study involved a diverse group of experts focusing on various OSN-related projects. P1 investigated Russian Twitter bots’ influence during the 2016 U.S. election, while P2 analyzed Grubhub restaurant reviews. P7 and P8 delved into public perceptions and understanding related to COVID-19 using Twitter data. P13, a sociologist, examined the impact of Instagram on teenage girls’ body image and noted TikTok’s rising generational significance. All participants were enthusiastic about discussing their research projects and explaining their choice of topics, including cybersecurity, cyberbullying, and financial services.

Table 3. Summary of Manual Analysis Theme and Basic Statistics. The second column #E represents the number of excerpts in the theme, #M denotes the count of media(interview transcriptions) fitting that theme.

| Theme | (E, M) | Description | Example Excerpt |
|-----------------------------------|----------|--|--|
| Research Field and Background | (11, 11) | Discussing job role, general research background, and domains | "My research is hybrid. So for example, sometimes during summer or winter when I have free time, I am interested in economics related topics" (P9) |
| Research Questions and Motivation | (36, 14) | Explaining the research questions and motivations behind a project related to OSNs | "I'm very interested in learning how people, in general outside the classroom environment, understand science and think about science." (P8) |
| General Data Description | (24, 13) | Indicating description of the data size, and time taken to collect, etc | "I collected 12 million tweets because we collected 15,000 tweets per day for 800 days" (P7) |
| Data Collection Method | (38, 14) | Pointing out data collection and the source of the data | "I think it was called the Trump Twitter archive. So we collect 10,000 tweets within a certain parameter." (P5) |
| Data Quality and Validation | (18, 13) | Highlighting the uses of specific methods to ensure and improve data quality | "that was just a lot of checking of different reviews, ... , and then also just kind of making sure that the right reviews were with the right restaurant." (P2) |
| Data Collection Challenge | (18, 12) | Discussing challenges or issues faced by the participant before analyzing the data | "The main challenge was the version two API, at that time was like a very recent rollout. And so the, like parameter information on Twitter like own guide for how to use its API was not fully up to date" (P7) |
| Data Preprocessing | (8, 5) | Mentioning the criteria of cleaning the data before analyzing | "If most of the tweet has words that are not in English, remove the tweet." (P7) |
| Data Analytic Method | (28, 14) | Describing methods, tools, or models used to analyze the data | "...count how many of these words in the tweets positive versus negative words, and then create some sort of overall sentiment for the tweet." (P6) |
| Data Analytic Challenge | (11, 7) | Demonstrating challenges or issues encountered during the analysis process | "our models were not that complicated. But it would be frustrating if something happened, and you lost it." (P4) |
| Data Storage | (5, 5) | Referencing method or software used to store the collected data | "I keep that on a sort of a thumb drive backed up, you know, by a password." (P5) |
| Ethical Consideration | (17, 12) | Considering an opinion about the privacy and ethical practices during data collection and publishing results | "I think it's very dependent on the user of the internet and the tools or resources that they have access to that they should have no expectation of privacy" (P3) |
| Project Collaboration | (11, 10) | Noting the number of people involved in the project and background of their collaborator(s) | "So this team had four people on it." (P4) |
| Future Improvement | (17, 11) | Discussing future improvement and any additional suggestions for future projects | "Oh, God, I would like to not spend so much time trying to wrangle the API; that was such a pain in the neck." (P1) |

4.2.3 General Data Description. Beyond the technical challenge of data collection and analysis, it is crucial to discuss and understand the inherent characteristics of the data. Strategically planning, such as selecting the right collaborator or data source, is invaluable. Four participants mentioned that their data volume ranged from hundreds to millions of records, collected over weeks to several months. While most participants worked with English text, two worked with data in other languages. For example, P10 worked with a multi-language dataset, and P7 examined the activity of Portuguese speakers on Twitter in Brazil.

4.2.4 Data Collection Method. Most participants employed multiple methods to collect data from OSNs, ranging from web crawling and API usage to manual data collection. Specifically, more

than one-third of the participants (P1, P2, P7, P8, P11) utilized APIs. These individuals either had a technical background or collaborated with technically proficient co-authors. P2 specializes in scraping a large volume of reviews, especially in the hospitality industry, where online reviews are accumulated. P8 stressed the importance of using short and unique search terms for effective data collection. P11 found around 30% of the collected data from API was irrelevant and had to change to manual curation. Almost half of the participants (6/14) engaged extensively in manual data collection work. P3 manually reviewed Reddit posts and multiple news sources to gather insights, highlighting the time-sensitive content of some events. P6 mentioned the possibility of purchasing pre-collected datasets, but he did not find suitable data for his specific research needs. Crowdsourcing platforms such as MTurk and Cloud Research were also discussed as data collection options.

4.2.5 Data Quality and Validation. Participants indicated that the success of their analytical efforts often hinges on the quality of the data. Similar to prior work, the idea of “Data Cascades” refers to accumulating events stemming from data influencing downstream analysis results [53]. This discussion emphasizes the negative consequences of traditional AI practices and generative AI when data quality is overlooked. Most participants engaged in the manual data review to improve data quality. P1 conducted spot checks on a small subset of data before relying on third-party software for broader screening. P2 conducted extensive manual checks randomly, reviewing as many entries as possible, though not all. P14 raised concerns about the possibility of social media posts originating from bots and often verified this through available IP addresses. P5 highlighted the potential pitfalls of using data-mining techniques. Despite the longer timeframe, their research involved manual coding with collaborators to ensure code reliability, similar to P6, P8, and P12. Based on human judgment, P3 utilized multiple data points and rated similar content for trustworthiness. Overall, these participants’ collection experiences emphasized the crucial connection between data quality and the effectiveness of analytical pursuits.

4.2.6 Data Collection Challenges. Most of the participants faced challenges with collecting and preprocessing datasets from OSNs. Participants raised concerns about misinformation and data accessibility. For example, P5 said “*I sought to use Twitter’s API, it is incredibly, I think, difficult to use. It’s not easy to get access.*” Two participants expressed the need for collaboration with other researchers to overcome challenges.

4.2.7 Data Preprocessing. Data cleaning is a standard step before undertaking data analysis. While this approach is common for data scientists, it might be less familiar to individuals from other backgrounds. Some of the participants anonymized and classified the dataset manually. Notably, participants frequently gathered more data than they ultimately used, often beginning their research with a small, exploratory sample. Participants frequently used Natural Language Processing (NLP) libraries for text data. These tools remove stop words and generate fundamental statistics, such as word count and sentiment scores. P7 set the filter to collect only English tweets but observed language other than English requiring manual removal.

4.2.8 Data Analytic Method. Participants adopted various standard analytical techniques to process data from OSNs. A significant number of the participants used sentiment analysis for textual data. Other participants have used topic modeling, regression models, statistical methods, and word embeddings. Most of participants (9/14) have employed quantitative analysis on their projects, utilizing tools like Stata, SPSS, and R packages.

4.2.9 Data Analytic Challenge. We asked about their challenges while the participants used data analytic techniques and tools. Half of participants (7/14) encountered difficulties. Participants faced technical issues, missing packages, and the need to convert existing scripts to other programming

languages. As P7 described, *“I believe it’s like a java file, and I’m not well versed in that, so I was trying to convert it into a python file.”* Participants emphasized the need to know the limitations of machine learning to apply them without challenges. Participants who used sentiment analysis experienced problems due to removing the stopping words, reading textual data, and picking a scoring system.

4.2.10 Data Storage. Proper data storage is especially critical for OSN data collection due to the sensitive nature of user-generated content. Ensuring robust and secure storage not only upholds the integrity of the research but also safeguards users’ privacy. Many participants use password-protected cloud storage, complemented by thumb drives. Some participants emphasized the importance of securely managing the data, as P5 stated, *“Only the people who are working on that research project have access to that information”*. Any oversights could jeopardize the study’s credibility, the researchers’ reputation, and the risk of data loss.

4.2.11 Ethical Considerations. The ethical considerations show the common opinion of the participants regarding the OSN data. P1, P2, P3, and P4 delved into the right of ownership of the data published on social media platforms. While some participants (3) mentioned the importance of the IRB approval, P8 and P5 highlighted the issue of obtaining consent for collecting publicly available data *“I really can’t get consent from anybody because you don’t know who they are, you don’t know if they have like serious accounts or not.”* (P8). Removing identifiable information was addressed as crucial for ensuring the anonymization of users.

4.2.12 Project Collaboration. None of the participants worked on the project individually. They collaborated with graduate students or professionals from different fields, reflecting the interdisciplinary nature of most projects. A typical research team consists of 2 to 3 members, and their collective expertise helps address complex and technical challenges.

4.2.13 Future Improvement. Participants provided suggestions for future enhancements, emphasizing the need for better research platforms and resources. P5 and P7 indicated the significance of access to historical and open data. P9 advocated for a dedicated social media data collector to ensure objective results and broader data capture. Meanwhile, P14 called for platforms to provide educational resources and guidance for budding researchers.

4.3 Thematic Analysis by GPT

We began the experiment with GPT-4 and later incorporated Gemini. Presently, directly requesting GPT to conduct a TA without supervision is not practicable. We carefully crafted effective prompts and input data for each step to facilitate a comprehensive TA using the OpenAI API. These prompts at each stage were initially inspired by Paoli’s work [17] and were further refined based on GPT responses. The following sections highlight the prompts optimized for effectiveness at each stage of the TA process. Specifically, adjustments were made to ensure they generated relevant and detailed outputs that facilitated the identification of themes and extraction of insights from the data. All resulting tables were organized by authors using content generated from GPT.

4.3.1 Phase 1: Familiarization. We ask GPT to get familiar with the data using each interview text.

Prompt: *“Please get familiar with the provided interview transcription or meeting minutes. Each interview is one-on-one. There should be only two speakers recorded. Please provide a maximum of 2 sentences summary for the interviewee based on the transcriptions or meeting minutes. Here is the interview from Participant {ID} {interview content}”*

Table 4. Summary of Interview for each Participant by GPT

| ID | Interview Summary |
|----|--|
| 1 | The interviewee is a data engineer who supports researchers in data collection and management for numerous projects at the university. They have a particular interest in the ongoing information war between major powers and the potential manipulation of public opinion through large language models and have worked on a project analyzing the impact of Twitter bots during the 2016 elections. |
| 2 | The interviewee is an assistant professor in the Hospitality and Sport Business Management Department, with a background in revenue management for hospitality applications. Their research involves using big data methodologies to analyze online reviews of hospitality products, with a focus on how these reviews impact a property's profitability and the importance of responding to reviews. |
| 3 | The interviewee is a teacher of a human relations administration course, focusing on understanding workplace dynamics, particularly employee-employer relationships. They incorporate analytics into their coursework, using real-time events for project-based learning, and have previously conducted a project analyzing sentiments during labor negotiations between UPS and the Teamsters Union. |
| 4 | Participant #4 is a researcher who primarily focuses on accessibility for people with disabilities and natural language processing. They have used social media data, particularly from Twitter, to develop technologies that could be useful for people with disabilities, and have also worked on projects involving sentiment analysis and predictive algorithms for communication devices for people who cannot speak. |
| 5 | The interviewee is a researcher who focuses on media effects from a quantitative communication perspective, particularly in sports and science. They have conducted a project analyzing cyberbullying tactics used by former President Donald Trump on Twitter, and they emphasize the importance of ethical considerations, data anonymization, and manual coding in their research process. |
| 6 | The interviewee is a researcher in Finance, specifically in asset pricing. They discussed a project they worked on that aimed to determine if tweets influence stock prices, detailing the process of data collection, analysis, and the challenges they faced, such as data organization, classification, and ensuring data quality. |
| 7 | The interviewee is a researcher who studies discriminatory behavior, specifically decision making and prejudicial behavior on social media. They have conducted a project where they collected tweets to analyze changes in stereotypes associated with Asians due to the COVID-19 outbreak, using tools like Twitter API, BeautifulSoup, Glove for word embeddings, and WEAT for comparing stereotype distance. |
| 8 | The interviewee is a science educator who conducts research in the field of science education, specifically focusing on how people of all ages understand science. They have recently been interested in the public understanding of science, using social media as a source of data, and have conducted a study analyzing tweets related to COVID-19 to understand how people conceptualize the nature of science. |
| 9 | Participant #9 is a researcher with a PhD in Social Science in Economics and a master's in statistics. They have worked in the pharma industry for many years and their research is a hybrid of economics and teaching, often involving the collection of fresh datasets for students to analyze. They have previously guided students on social media development and marketing, specifically researching the impact of social media marketing over time. |
| 10 | Participant #10 is conducting a research project that involves analyzing Google reviews of parks and geotagged tweets related to food and water insecurity, with a focus on English and Swahili languages. The participant uses Twitter Archiver and NLP models like BERT and GPT for data collection and analysis, and is interested in misinformation tweets and social issues like misinformation, conspiracy theories, populism, polarization, and extremism. |
| 11 | Participant #11 has experience in manual data curation from online social networks, specifically from the UN conference data where they downloaded 800-900 speech scripts from different countries. They have also worked on a project involving Twitter, where they scraped tweets using specific keywords, and expressed interest in the NLP expansion and models of SMIDGen, and would like to see features that can find different languages, hashtags, and multiple keywords. |
| 12 | The interviewee is a researcher in the field of cybersecurity, focusing on the behavioral aspects of cybersecurity, particularly employee-related behaviors that could pose a threat to an organization's cybersecurity. The interviewee's research involves analyzing data from various sources, including publicly available data on companies and data from social media, using software for sentiment analysis and manual review processes. |
| 13 | The interviewee is an academic advisor in the Department of English at a university, with a background in sociology. They have experience in conducting research, particularly in the field of social sciences, and have used datasets such as the General Social Survey for their work. They have also utilized social media to gather survey responses and expressed interest in the potential of social media data for sociological research, particularly in studying attitudes and behaviors related to race, gender, and political polarization. |
| 14 | The interviewee is a nurse practitioner who conducted research on fertility knowledge among young adults in America, aged 18 to 25. She used crowdsourcing platforms like MTurk and Cloud Research to collect data, and used Excel and Jump for data analysis, focusing on knowledge rather than sentiment. |

The responses from GPT are detailed in Table 4. Each summary demonstrates a comprehensive understanding of the interview content and aligns with human analysis in Table 1.

4.3.2 *Phase 2: Generating Initial Codes.* We ask GPT to code interesting excerpts systematically.

Prompt: “I would like you to perform qualitative coding for an interview or meeting minute. I would like to see the results in a table. The first column is the code name. Each code should have less than 4 words. The second column is the code definition. Each definition should have less than 15 words. The third column is the total number of excerpts identified for that code. The fourth column displays all excerpts for that code. Do not shorten. Keep the original excerpts. You should not code the questions asked by the interviewer. You should apply codes to excerpts such as sentences or parts of sentences. For each interview, you should identify 10-20 codes. Please return the table as a JSON object. Here is the interview from Participant {ID} {interview content}”

Per our inductive approach, We did not supply a codebook but requested GPT to perform open coding. GPT produced 54 unique codes, comparable to human analysis (i.e., 51). While not all are expressed in the exact wording, some codes share similar semantic meanings. The total number of excerpts selected by GPT was 176, while human analysis identified 243. Table 5 shows the full list of code names. Multiple codes share similar meanings because GPT processed each interview independently, posing a challenge for GPT to compile a complete code book autonomously. Overall, 46 (85.19%) initial code names generated from GPT are similar to the manual analysis. They either match the name exactly or can be inferred semantically. In contrast, 8 (14.81%) initial codes such as “Research Transparency,” “Teaching Role,” and “Social Media Research” were overlooked previously from human coding.

Table 5. Initial Codes Generated by GPT. #E represents the number of excerpts identified. Codes in bold indicate matched semantically with human analysis

| Code Name | #E | Code Name | #E | Code Name | #E |
|----------------------------|----|-------------------------------|----|-------------------------|----|
| Data Collection | 32 | Project Challenges | 2 | Data Incorporation | 1 |
| Data Analysis | 22 | Project Duration | 2 | Project Example | 1 |
| Data Quality | 11 | Team Size | 2 | Data Annotation | 1 |
| Ethical Considerations | 11 | Project Description | 2 | Computation Limitations | 1 |
| Research Area | 8 | Data Management | 2 | Data Privacy | 1 |
| Challenges | 5 | Data Processing | 2 | Collaboration | 1 |
| Trending Topics | 5 | Data Volume | 2 | Research Team | 1 |
| Data Challenges | 5 | Research Topic | 2 | Research Transparency | 1 |
| Improvements | 4 | Research Question | 2 | Research Tools | 1 |
| Social Media Research | 4 | Data Source | 2 | Data Limitations | 1 |
| Data Collection Challenges | 3 | Data Sources | 2 | Data Usage | 1 |
| Data Storage | 3 | Data Collection Method | 1 | Qualitative Research | 1 |
| Future Improvements | 3 | Data Quality Verification | 1 | Research Focus | 1 |
| Data Quality Assurance | 3 | Privacy Concerns | 1 | Data Collection Need | 1 |
| Data Cleaning | 3 | Project Improvement | 1 | Data Analysis Tools | 1 |
| Data Collection Tools | 3 | Data Documentation Importance | 1 | Qualitative Analysis | 1 |
| Role Description | 2 | Research Background | 1 | Social Media Data | 1 |
| Research Interest | 2 | Teaching Role | 1 | Data Collection Service | 1 |

4.3.3 *Phase 3: Searching for Themes.* We ask GPT to organize codes and identify potential themes.

Prompt: “I would like you now to search for themes as part of thematic analysis. I will give you the code name and code definition. You should determine which code can be merged together based on commonalities or overarching concepts. I would like to see the result in a table. The first column is the name of the theme. The second column is all codes that fit into the theme. The third column is a

brief summary of the theme. Please return the table as a JSON object. Here are the codes and their definitions [response form Phase 2].”

The response from this Phase is Trial #1 in Table 6. We discuss the result in the next Phase.

4.3.4 Phase 4: Reviewing Themes. We ask GPT to review the meaning of each code and related themes.

Prompt: This is the same prompt as in Phase 3.

We used the same prompt as in the previous phase. However, we adjusted the temperature setting in this phase to capture additional information that might have been missed. In general, setting *Temperature* = 0.5 often offers a blend of creativity and coherence, and *Temperature* = 1 could exhibit greater creativity and capture unexpected outputs. This approach allows us to request GPT re-evaluate the data to generate less predictable themes that may provide additional insights. Upon manual inspection of the themes and descriptions, we present results from four experiments and display them in a separate column (Table 6). Each trial yielded approximately 7 to 10 emergent themes, with some codes remaining ungrouped. For instance, Trial 1’s themes encompassed 49 unique codes, while Trial 4 covered the fewest, with 41 unique codes. The initial 5 themes (i.e., rows) in each trial exhibit strong similarities, while subsequent themes demonstrate variations in naming. The themes will be finalized in the next phase.

Table 6. Themes Identified by GPT. Each column is the response from GPT, and the parameter *T* (temperature) regulates uncertainty. *n* represents the number of unique initial codes covered by all themes.

| Trial 1 (<i>T</i> = 0, <i>n</i> = 49) | Trial 2 (<i>T</i> = 0.5, <i>n</i> = 45) | Trial 3 (<i>T</i> = 1, <i>n</i> = 49) | Trial 4 (<i>T</i> = 1, <i>n</i> = 41) |
|--|--|--|--|
| Research Context | Research Profile | Participant Information | Role and Research Information |
| Data Collection | Data Collection | Data Collection | Data Collection |
| Data Analysis | Data Analysis | Data Management | Data Analysis |
| Data Quality | Data Quality and Management | Data Analysis | Data Quality |
| Ethical and Privacy Considerations | Ethical and Privacy Considerations | Ethical and Privacy Considerations | Ethics and Privacy |
| Project Management | Project Management and Improvement | Project Logistics | Challenges and Improvements |
| Data Management | Data Sources and Use | Future Aspects | Project Details |
| Data Usage | | Research Tools | Data Management |
| Research Tools and Sources | | | Other Research Aspects |
| Special Research Areas | | | |

4.3.5 Phase 5: Defining and Naming Themes. We ask GPT to refine the specifics of each theme and determine what aspect of the data each theme captures. We set the temperature parameter back to 0 to ensure greater reproducibility.

Prompt: “Please review all the themes and revise them if needed. Please provide a summary of each theme. The results should be presented in a table. The first column is the name of the theme. The second column is codes associated with the theme. Every unique code from the original data should be included and only once. The third column is a brief summary of the theme. Please return the table as a JSON object. Below are the themes, codes, and original summary of the themes: {themes}.”

Using the previously selected themes from multiple trials, descriptions, and codes, GPT refined its analysis to produce a final theme. Table 7 shows the 11 refined themes from the initial codes. This final theme table covers 45 unique codes, comprising 153 excerpts. GPT omitted 9 initial codes. Except for the themes of “Research Tools and Sources” and “Special Research Areas”, the remaining 8 (80%) of the generated themes by GPT align with the manual analysis themes listed in Table 3. There are 3 codes that overlap with the themes. For example, the initial code “Data Collection

Tools” was categorized under both the “Data Collection” theme and “Research Tools and Sources.” Although we directed GPT to minimize code overlap, this aspect is beyond our control. Overall, human experts labeled 18.7 excerpts per theme on average, while GPT coded 15.3 per theme on average.

Table 7. Final Theme by GPT. #E indicates the number of excerpts. #M represents the number of transcriptions that align with the theme. Themes names in bold are matched with manual analysis

| Theme | Description | Codes Merged | (E, M) |
|---|---|---|----------|
| Data Analysis | Encompasses the methods and tools used for data analysis, inclusive of qualitative analysis. | Data Analysis, Data Analysis Tools, Qualitative Analysis | (24, 12) |
| Data Collection | Details on the methods, tools, challenges, and needs associated with the process of data collection. | Data Collection Method, Data Collection Challenges, Data Collection, Data Collection Tools, Data Collection Need, Data Collection Service | (41, 14) |
| Data Management | Involves the importance of documenting data collection methods and the methods used to store and manage data. | Data Documentation Importance, Data Storage, Data Management | (6, 6) |
| Data Quality | Relates to the methods employed to verify, maintain, and assure the quality and accuracy of data. | Data Quality Verification, Data Quality, Data Quality Assurance | (15, 10) |
| Data Usage | Relates to how data is used, processed, cleaned, and annotated in the research. | Data Incorporation, Data Usage, Data Volume, Data Cleaning, Data Processing, Data Annotation | (10, 6) |
| Ethical and Privacy Considerations | Covers concerns and considerations related to privacy and ethics in the process of data collection and analysis. | Privacy Concerns, Ethical Considerations, Data Privacy | (13, 11) |
| Project Management | Focuses on the management aspects of the research project, discussing challenges, improvements, duration, team size, and collaboration. | Project Improvement, Project Challenges, Project Duration, Team Size, Future Improvements, Improvements, Collaboration, Research Team | (16, 10) |
| Research Context | The context of the research, including the participant’s role, their research interests and background, the main area or topic of their research. | Role Description, Research Interest, Research Background, Teaching Role, Research Area, Research Topic, Research Question, Research Tools, Research Focus | (20, 14) |
| Research Tools and Sources | Refers to the tools used for data collection and analysis, as well as data sources. | Data Sources, Data Collection Tools, Data Analysis Tools, Research Tools | (2, 1) |
| Special Research Areas | Focuses on specialized areas of the research like social media research and qualitative research. | Social Media Research, Social Media Data, Qualitative Research | (6, 3) |

We computed Cohen’s Kappa and Percent Agreement using ReCal[23] to measure inter-rater reliability (IRR) between raters (i.e. GPT and humans). Due to the inductive approach employed, excerpts selected from humans and GPT exhibit many variations. Consequently, we inspected each excerpt and aligned them when there were overlapping sentences or words. We prepared two datasets to calculate agreements. One dataset has 242 excerpts that are selected from all the human and GPT. The other dataset has 111 excerpts that were selected only from GPT. To highlight, humans missed 4 excerpts that GPT selected. To give an example, the excerpt “I use social media, I do a lot of things on like opinions and attitudes... P(13),” was missed in the human analysis. Including these four excerpts scarcely appears to modify the interpretation of the overall themes and data, so we excluded it. Furthermore, the IRR was computed based on 8 matching theme levels (bolded in Table 7) because labels were originally assigned under different code names between raters. We present results in Table 8. The overall Cohen’s Kappa value reached 0.72 (Percent Agreement 93.02%) when considering only GPT selected excerpts, indicating moderate agreement as suggested

Table 8. Human and GPT Agreement

| Theme | All Excerpts (n = 242) | | GPT Selected Excerpts (n = 111) | |
|-----------------------|------------------------|----------------------|---------------------------------|----------------------|
| | Cohen's κ | % Agreement | Cohen's κ | % Agreement |
| Data Analysis | 0.14 | 42.56 | 0.71 | 92.79 |
| Data Collection | 0.1 | 38.02 | 0.44 | 82.88 |
| Data Management | 0.04 | 45.45 | 0.88 | 99.1 |
| Data Quality | 0.11 | 45.04 | 0.9 | 98.2 |
| Data Usage | 0.03 | 43.39 | 0.47 | 94.59 |
| Ethical Consideration | 0.12 | 45.87 | 1 | 100 |
| Project Management | 0.13 | 36.36 | 0.44 | 79.28 |
| Research Context | 0.12 | 44.63 | 0.88 | 97.3 |
| Mean (SD) | 0.10 (± 0.04) | 43.67 (± 3.57) | 0.72 (± 0.23) | 93.02 (± 7.79) |

in [43]. Since humans had selected 131 more excerpts than GPT, the average Cohen's Kappa score was 0.10 (Percent Agreement 43.67%) when considering all experts from both raters.

4.3.6 Phase 6 Writing the Report. We ask GPT to provide a detailed analysis of each theme with associated data extracts.

Prompt: "As the final stage of thematic analysis, please provide a summary and insights for the themes based on the excerpts. Please keep it concise and the summary should not exceed 100 words. You may mention a few quotes with the participant ID. Below are the themes, excerpts, and participant ID: {data}."

Table 10 summarizes each theme derived from GPT. For more precise visualization and space efficiency, we substituted most of the term "participant" with the abbreviation "P." The remainder of the content in the second column originates directly from GPT. The final column aligns with themes from the human TA based on manual inspection and inference, except for two themes that couldn't find a match. In general, the summaries provided are comprehensive and insightful. To assess the performance of GPT and for more quantitative comparisons, we applied sentence embeddings built from BERT [19] on each final theme from manual and GPT summaries. We computed semantic textual similarity using the cosine similarity score and presented the results in Table 9. These similarity scores, ranging from 0.45 to 0.76, highlight a promising opportunity to employ GPT for TA. In addition, Figure 4 shows similarity scores are highest for matching themes compared to non-matching themes summaries.

Table 9. Semantic Cosine Similarity Score between GPT Final Report and Manual Analysis. Model 1 (all-MiniLM-L6-v2¹⁴) and Model 2 (paraphrase-MiniLM-L6-v2¹⁵) are utilized for sentence embeddings.

| | GPT Final Themes | Model 1 | Model 2 |
|----|------------------------------------|--------------------|--------------------|
| 1 | Data Analysis | 0.549 | 0.702 |
| 2 | Data Collection | 0.756 (max) | 0.656 |
| 3 | Data Management | 0.723 | 0.716 |
| 4 | Data Quality | 0.585 | 0.727 (max) |
| 5 | Data Usage | 0.679 | 0.678 |
| 6 | Ethical and Privacy Considerations | 0.707 | 0.684 |
| 7 | Project Management | 0.592 | 0.621 |
| 8 | Research Context | 0.447 (min) | 0.578 (min) |
| 9 | Research Tools and Sources | - | - |
| 10 | Special Research Areas | - | - |

¹⁴<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

¹⁵<https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2>

Table 10. Interpretation of each Theme by GPT. The final column denotes themes analogous to those identified in human thematic analysis.

| Theme | GPT Analysis Overview | Link |
|------------------------------------|--|---------------------------|
| Data Analysis | The themes mostly revolve around data analysis from different perspectives. Participants use various tools, metrics, and methods to understand data. Methods include sentiment analysis (P1), text mining and topic modeling (P2), historical analysis (P3), inferential and statistical analysis (P5), merging different data sets to interpret results (P6), and using word embeddings and advanced models (P7). Participants also mentioned using qualitative analysis and Excel (P14), trend analysis (P9), and modeling (P11). Some reported removing unneeded information before analysis (P11). | 4.2.8 4.2.9 |
| Data Collection | Participants use various methods for data collection, from Twitter API and scraping tools to search engines and online sources. Participants commonly use APIs; however, they highlighted issues such as response limits (P1) and turnaround periods (P8). Others mentioned using Python (P2) and third-party software to bypass API restrictions (P5). Data sets were sourced from public information (P5 and P13), and P2 mentioned students compressing and scraping reviews. Data collection periods ranged from a day (P9) to multiple days (P1). | 4.2.4 |
| Data Management | Data Management emerges as a core theme across participant responses. Participants acknowledge the importance of data documentation for later analyses (P1). The excerpt illustrates a range of storage methods, including cloud-based systems (P2), anonymized physical storage (P5), and commonly used software like Excel (P9, P14). Ensuring anonymity and access restrictions were also key concerns in managing data (P5). | 4.2.10 |
| Data Quality | The theme revolves around 'Data Quality' and the various measures taken by the Participants to ensure the accuracy, relevance, and validity of their data. Notably, Participants involved in research trusted government agencies or reliable websites (P1, P11), conducted a lot of checks, and reviewed the data (P2). They utilized human judgment (P3), disposed of unfit data (P4), and employed validation methods like subsampling and correlation checks (P6). To ensure quality, various annotation tools, coding system, and rubrics were used (P10). | 4.2.5 |
| Data Usage | Participants emphasized the importance of data usage, through incorporating analytics into coursework, collecting extensive data such as millions of tweets, and focusing on organizing data sets for analysis. Challenges like huge data sets and language cleanup were encountered during processing as stated by P6 and P7. Some involved in survey research ensured crucial questions were answered sequentially to ensure data integrity and completeness (P14). Overall, Participants utilized systematic strategies while dealing with different aspects of data. | 4.2.7 |
| Ethical and Privacy Considerations | The key theme 'Ethical and Privacy Considerations' navigates around researchers' ethical hesitations dealing with the privacy issues of public Twitter users. Many respondents struggled with the idea of utilizing public information without explicit consent or ensuring anonymity; P5 remarked "just because we can scrape it doesn't mean we should." Others, like P6, noted the data was already anonymized, and P9 ensured they only used general demographic data to respect privacy. | 4.2.11 |
| Project Management | The 'Project Management' theme suggests challenges and improvements Participants would like in their projects. Participants experienced troubles managing APIs (P1) while also underestimating tasks like text mining (P2). They expressed a need for more defined project goals (P2), more access to historical data and bigger data sets (P5, 6), and more direct handling of raw data (P6). They also expected changes due to Twitter's policy changes (P7) and expressed interest in new tools like social media data collectors (P9). Team sizes varied (P2, P4). Project duration was mostly long-term, around a semester or six months (P2, P6, P10). | 4.2.6 4.2.12 4.2.13 |
| Research Context | The main theme here is the diverse research contexts of different Participants. They come from various fields like data collection, finance, communication, science education, and cybersecurity etc. P1 supports data collection and management for research projects while P6 studies how social media influences stock prices. P9 hybridizes their research with different domains like economics and statistics during their available time. All the Participants provide unique perspectives on their main area of research. | 4.2.1 4.2.2 |
| Research Tools and Sources | The theme explores 'Research Tools and Sources' used by Participants. P13 highlighted that population level surveys are easily accessible and do not require special governmental clearance. They feel secure using data from government sources. Even though this source seems non-problematic, the Participant suggests potential implications of using government-derived information. | |
| Special Research Areas | The conversations reveal a common theme of 'Special Research Areas', with a focus on the role of social media in society. Participants, such as number 9, discussed guiding students on the impact of social media marketing. P13 highlighted how sociologists study social media from both qualitative and quantitative data perspectives. P14 conveyed the significance of studying social media, especially in regards to their work with specific age groups. | |

4.4 Thematic Analysis by Gemini

We applied the same prompts and settings to Gemini and conducted a detailed comparison and analysis, as outlined in the supplementary materials. Similar to GPT-4, Gemini effectively generated feedback for each phase of the thematic analysis. Specifically, Gemini generated 123 unique initial codes, twice the number compared to those generated by human experts and GPT-4. Gemini produced the final seven themes, including Data Analysis & Text Mining, Data Collection & Processing, Information Warfare & LLM Threats, Media Effects & Public Perception, Research Design & Methodology, Researcher Background, and Twitter Data & Social Media. Although the maximum summary length in the prompt is set to less than 100 words, the final summaries generated by Gemini tend to be shorter, averaging 53 words with a variation of 14. In contrast, summaries produced by GPT-4 are longer (75 ± 15 words). We observed greater hallucinations in the generated content when we attempted to increase the maximum number of words. Applying the same method to compare final insights between Gemini and human experts, we computed the cosine similarity, resulting in a range of 0.37 to 0.54 with a mean of 0.46. Please see the supplementary materials for full details.

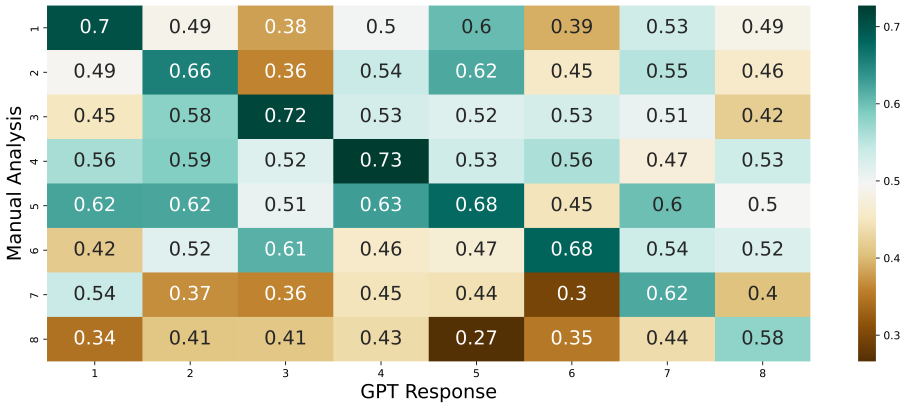


Fig. 4. Similarity Score Heatmap Per Theme. Each score represents cosine similarity when comparing manual and GPT analyses from a specific theme. Diagonal entries have the maximum value. Model 2 (paraphrase-MiniLM-L6-v2¹⁵) embedding was used.

5 Discussion

We address our research questions, reflect on the methods used, and describe future opportunities for research in this area.

5.1 Data Collection Experience from OSN

Our RQ1 aims to explore how researchers collect data from OSNs and how we can enhance this process. As OSN data attracts significant attention across diverse research domains, their use in research will continue to grow in popularity [58]. Our combined analysis with LLMs showed that participants employed various data collection techniques from OSNs, such as the Twitter API, scraping tools, and manual online searches. Almost all participants struggle to identify a reliable tool to collect OSN data. This obstacle affects research advancement and result quality, highlighting the demand for designing an efficient and effective data collection resource.

Challenges. Regardless of the popularity of APIs, issues like response limits and extended processing times posed challenges. Although prior studies (e.g., [10, 42, 70]) highlighted potential difficulties over decades, significant gaps persist in the data collection and analysis stages. Predominant challenges include ethical dilemmas, data accessibility, and ensuring data integrity. Through one-on-one interviews with experienced researchers, it became evident that lack of data access and technical challenges remain prime concerns. As Participant 5 stated *“the biggest hurdle is just getting started...I didn’t know where to begin...I don’t have a coding background.”* This kind of frustration could obstruct many from conducting their research. Many researchers often need to collaborate with technically proficient colleagues for a straightforward solution. In addition, data quality remains another concern, which often leads to potential misinterpretations and requires additional manual effort for data review and filtering. Some researchers pointed out the frustrations of creating accounts on many different social media platforms to access public content. Other concerns raised include potential selection bias due to deleting historical posts. Moreover, a common discussion was the lack of API documentation. Participants emphasized the importance of ensuring replicability and facilitating future data code sharing.

Process Improvements. Two process improvements surfaced from our findings: the importance of historical data and the desire for better search and query filters. Researchers often prioritize accessing historical data from online network platforms, while some indicate that real-time data holds little value. As OSN platforms evolve, there is less access to historical data, and accessing such data has become more costly. For instance, Twitter discontinued the academic research API in June 2023 and imposed stricter restrictions on API access with a more expensive subscription plan [66]. Incorporating advanced search filters or query capabilities and creating a user-friendly interface could significantly improve researchers’ satisfaction. To facilitate the collection process with any tool, it is essential to keep the documentation and tutorials up to date for clear guidance. Lastly, it is surprising that participants have given low priority to features such as collaborative collection, compatibility with multiple OSN platforms, and data anonymization, even for those with higher satisfaction with their existing processes.

5.2 Comparison of Human and LLMs Results

We have experimented with both GPT and Gemini in TA to address RQ2. While both LLMs are capable of conducting TA on our interview data, we found that GPT performs better.

Responses from GPT using our interview data are promising. During phase 1 familiarization, GPT provided a comprehensive high-level summary, showcasing an effective understanding of each interview. Given that the response from this phase does not progress to the next step, it serves as a trial test for GPT to comprehend input data. If the summarization is less comprehensible, it could pose challenges in processing the rest of TA phases. In the second phase (i.e., Generating Initial Codes), GPT annotated comparable excerpts with a level of coding similar to human analysis. GPT tends to overlook excerpts, especially those from meeting notes (which haven’t been digitally recorded), resulting in a much lower number compared to human excerpts. In phases 3 and 4 (i.e., Searching and Reviewing Themes), GPT effectively identified themes similar to human analysis. The variations in naming led to different themes in total. The difference could provide potential adjustments for researchers in their theme selection. Based on results from Section 4.3.5, the majority of the themes selected from GPT are matched with human analysis following phase 5 (i.e., Defining and Naming Themes). All themes in Table 7, with the exceptions of *“Research Tools and Sources”* and *“Special Research Areas”* align with our manual TA. When examining IRR with GPT-selected excerpts in Table 8, the overall Cohen’s Kappa is even higher than the deductive approach in one study [68]. In detail, our analysis shows nearly perfect agreement among raters for the *“Ethical Consideration”* theme. Strong agreement is observed for the themes of *“Data Management,”* *“Data*

Quality,” and “Research Context.” “Data Analysis” exhibits moderate agreement. In contrast, the themes of “Data Collection,” “Data Usage,” and “Project Management” show weak agreement. It is important to note that IRR for all excerpts in the dataset presents baselines with significant considerations, highlighting areas for future work. These metrics can serve as benchmarks for enhancing TA processes, such as designing prompts. Lastly, in phase 6 (Writing the Report), most interpretations generated by GPT are insightful and valuable.

GPT may offer additional information that human experts missed. As an example, for the theme “*Ethical and Privacy Considerations*”, GPT managed to extract additional information, such as P6’s observation that the data was already anonymized, and highlighted a compelling quote from P5: “*just because we can scrape it doesn’t mean we should.*” For the theme “*Project Management*”, GPT deduced that participants were aware of changes within social media platforms and expressed an interest in newer data collection tools. Furthermore, within the “*Data Collection*” theme, GPT highlighted discussions about utilizing third-party software to prevent API restrictions, as noted by P5. Similar behavior was observed that GPT tends to pay more attention to certain edge cases that human experts may have overlooked.

There are moderate agreements between human experts and GPT. We compared and assessed IRR metrics between GPT and human raters. GPT typically chose shorter excerpts and occasionally selected the same word or sentence multiple times. In contrast, human researchers coded longer experts. Overall, our manual and GPT responses indicate a similarity score ranging from 0.45 to 0.76 when applying sentence embeddings. These results are promising as they demonstrate significant possibilities for utilizing LLMs for TA.

Using Gemini in TA is feasible but shows limited performance. Gemini tends to use quotes directly more often than GPT-4 or not at all. Although Gemini offers the option to return the response in JSON format, we experienced inconsistent JSON formats. It often has extra unnecessary characters that make it challenging to handle or merge data for subsequent processing. Additional data handling mechanisms may be required for verification if an automated process is being built. GPT-4 and Gemini 1.5 Pro are believed to have comparable parameters in their models, although the exact numbers have not been disclosed. We have added additional comparisons to the supplementary materials.

5.3 Reflection of LLMs Methods

Exploring LLMs with TA was conducted after the manual TA. We first employed the standard GPT-4 model, which accommodates a maximum context window size of 8,192 tokens (around 6,000 words). Given that none of our interview scripts exceeded this token limit, there was no need to divide the files. While the gpt-4k-32k model¹⁶ can handle up to 32,768 tokens (roughly 25,000 words), we did not have access at the time of the study. Consequently, it was not feasible to aggregate all interview files into a single prompt. On the other hand, Google Gemini 1.5 Pro features a 1 million context window, potentially capable of accommodating all interview data at once. To maintain consistency and comparability with GPT, we experimented with LATA process and Gemini one file at a time. Further research could explore how this distinction influences performance.

As mentioned earlier in the paper, varying prompts and study designs could have significantly different responses. We dedicated significant effort to prompt engineering, experimenting with multiple prompts for each process phase. Obtaining intermediate results from GPT poses a considerable challenge. Therefore, We segmented the process into distinct phases to navigate GPT through the TA. Researchers may need to meticulously handle and prepare the data between each phase for GPT’s subsequent processing, which sometimes could be very tedious but likely less so than

¹⁶<https://platform.openai.com/docs/models/gpt-4>

conducting a fully manual TA. Additionally, we encountered three different JSON formats from Gemini and had to manually convert each one. To address our RQ3, we discuss the effectiveness and limitations.

Conducting TA with LLMs presents numerous advantages. Undoubtedly, LLM-powered TA runs faster than human coders' analysis. While LLMs might need supervision from a single researcher, traditional analysis often demands input from at least two human researchers. Moreover, LLMs can repeatedly execute the process for enhanced accuracy and offer easy reproducibility. Achieving such consistency can be challenging for human experts. Chen et al. envision the technical perspective for future human-centered generative AI (HGAI) and discussed that qualitative research often inherently carries personal biases from researchers, which can influence their interpretations; however, this subjectivity could be diminished or reduced when utilizing AI such as GPT [12]. This paper also advises the need for the HCI community to establish clear guidelines and evaluation criteria for qualitative studies using LLMs. Despite the usefulness of GPT, the purpose of using LLMs is not to replace them with traditional methods but rather to provide supplementary assistance [25].

An emerging concern associated with LLMs is their potential to produce inaccurate or flawed information, often referred to as 'hallucinations.' Ashwin et al. investigated serious bias (i.e., annotation accuracy) that LLMs may introduce, especially on specific topics [5]. In our case, GPT disregarded the initial codes like "Challenges," "Data Challenges," and "Data Limitations," which were not particularly helpful, and we believe they should be merged into a single final theme. In addition, Table 4 displays summaries generated by GPT for each interview. Out of these, 10 (71.43%) start with the term "The interviewee," while 4 (28.57%) begin with the word "Participant." A human researcher might opt for consistent phrasing, but GPT lacks this nuanced understanding. This inconsistency might arise because our interview files are provided to GPT individually. The data format and resulting response are challenging to maintain across phases due to the lack of complete control over LLMs. For example, the code "Data Analysis Tools" appeared under two themes in Table 7, despite our request to avoid duplication. However, we can filter out such duplicate information afterward and prepare the data for the next stage. Using the API, GPT and Gemini do not automatically make inferences from past prompts, and it's nearly impossible to feed in all historical prompts due to limited input size. While challenges persist in utilizing LLMs for TA, employing LLMs could be an additional tool to aid or validate human analysis.

5.4 Ethical Considerations

We have carefully reviewed and removed sensitive information from the interview transcripts in this pilot study. As mentioned in section 3.3.2, we trust that OpenAI and Google company do not store our data on any of their servers. Our LATA involves uploading individual transcripts one at a time. The intermediate data are processed offline without involving LLMs.

Since LLMs are pre-trained algorithms using extensive datasets, numerous researchers have raised ethical concerns [46, 67]. These concerns often focus on issues such as social bias and harmful content, prompting questions about their capabilities and objectives. Recent studies have proposed practical solutions to address these concerns. Qu et al. [49] suggest employing machine unlearning to discard sensitive or biased data selectively. Another study underscores the importance of robust defensive strategies to mitigate security threats [36]. Our study found few instances where the LLMs generated inaccurate or irrelevant information. Importantly, we did not encounter any harmful content produced by the LLMs. The LATA process also aims to maintain ethical standards by ensuring that humans are consistently involved in the process and reviewing feedback at each stage. However, future research could explore the implementation of anonymity measures to improve this

process during the data preparation phase and to identify bias in output throughout the analysis and reporting process.

6 Limitations and Future Opportunities

Concerning the work presented here, it is important to acknowledge several limitations. First, our participant numbers are limited, but we are confident that the conditions of our study reflect common interviewing and data analysis practices suitable for a pilot study. Saturation had been achieved, indicating that further interviews are unlikely to offer new insights. Despite a diverse participant pool, certain research domains such as natural sciences, humanities, and environmental science are underrepresented. While broadening our participant base could enrich the range of perspectives, it is unlikely to alter the overall scope of our study. Secondly, most recent LLM papers are exclusively available on archive servers such as arXiv¹⁷. The reason is primarily due to the rapid emergence of this technology. Although we include references to these papers, it is crucial to acknowledge that they may not have been peer-reviewed. Thirdly, we approached this pilot study of LATA with an inductive method. Therefore, coding at the response level occurred without explicit instruction in the prompt for LLMs. To enhance examination and calculate direct IRR, one alternative approach could involve instructing humans and GPT to code at the sentence level and requiring all raters to evaluate each text segment. During the interview study, two of our sessions were not digitally recorded, posing a significant challenge for LLMs in interpreting the content as an interview. In addition to constraints on input characters for prompts, there are additional rate limits. At the time of our analysis, GPT-4 is still in development, and the rate limit is 10,000 tokens per minute (TPM) and 200 requests per minute (RPM), which is considerably stricter than those of GPT-3.5. Errors such as “server overload” with GPT were a recurrent issue, and it’s uncertain if this has affected the responses. Gemini often shows exhausted resources and requires a manual time break between each request. With the rapid development of LLMs, one could also experiment with various LLMs methods to fully understand and evaluate their effectiveness in TA.

Our findings have several important implications for the design and analysis of LATA-like systems. First, the importance of delivering precise and unambiguous instructions cannot be overstated. While we have showcased our prompt design using our interview content, it might not yield the same effectiveness when directly borrowed for use with other datasets or contexts. Including the research question in the prompt may help the LLMs interpret interview data or focus on a specific theme analysis. One study highlights future research directions such as automatically generating prompts [73]. More studies are needed to investigate whether a universal framework can be developed to design TA-supporting prompts. Secondly, if the transcription file exceeds the maximum prompt tokens, a mechanism will be needed to break down the file into multiple pieces. The initial code must then pass through these pieces to ensure more consistent coding. Thirdly, intermediate data should be handled carefully to ensure accuracy. Responses from LLMs may vary in format, even when provided in JSON. Additionally, privacy concerns arise as LLMs can memorize and replicate information from their training data, which may introduce bias and fairness issues.

Our study employed the inductive approach with LLMs for TA. Future work could provide LLMs with a codebook and associated definitions curated by domain experts to assess if this refines the outcomes. For such a deductive approach, various learning settings, like Zero-shot, One-shot, and Few-shots can be considered. While we used a separate tool, the capabilities of the LLMs could even be practiced for transcribing audio files and further automating TA.

¹⁷<https://arxiv.org/>

7 Conclusions

This work demonstrated the possibilities of conducting a 6-phase deductive TA using GPT-4 and Gemini-1.5 Pro with rigorous prompt engineering. The potential of using LLMs to aid in qualitative analysis suggests an exciting avenue for future research. We observed promising responses from GPT-4, demonstrating a capability of performing TA comparable to human expert analysis within our interview content. As a pilot study, this work could establish a benchmark process for future studies to follow and compare. Furthermore, from the interview content itself, we found that the complexities of collecting data from OSNs remain significant challenges for researchers and often require substantial manual effort. The primary obstacles include accessibility concerns, implementation hurdles, and noisy data quality. The participants emphasized the need for improved tools, especially for accessing historical data and optimizing search capabilities. Addressing these challenges as the social and computational science gap evolves is crucial to ensure effective and efficient data collection from OSNs.

8 Acknowledgments

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s). ChatGPT was utilized to generate section 4.3 of this work, including text and tables. Responses from ChatGPT were reviewed and revised for clarity.

References

- [1] Marissa D Abram, Karen T Mancini, and R David Parker. 2020. Methods to integrate natural language processing into qualitative research. *International Journal of Qualitative Methods* 19 (2020), 1609406920984608. doi:10.1177/1609406920984608
- [2] Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2023. Open-source large language models outperform crowd workers and approach ChatGPT in text-annotation tasks. *arXiv preprint arXiv:2307.02179* (2023).
- [3] Fahdah Alshaikh, Farzan Ramzan, Salman Rawaf, and Azeem Majeed. 2014. Social network sites as a mode to collect health data: a systematic review. *Journal of medical Internet research* 16, 7 (2014), e171. doi:10.2196/jmir.3050
- [4] Simon Arvidsson and Johan Axell. 2023. Prompt engineering guidelines for LLMs in Requirements Engineering. (2023).
- [5] Julian Ashwin, Aditya Chhabra, and Vijayendra Rao. 2023. Using Large Language Models for Qualitative Analysis can Introduce Serious Bias. *arXiv:2309.17147* [cs.CL]
- [6] Yahya Benkaouz and Mohammed Erradi. 2015. Towards a Decentralized OSN for a Privacy-preserving e-health System. *Procedia Computer Science* 63 (2015), 284–291. doi:10.1016/j.procs.2015.08.345
- [7] Teresa Blanco, Álvaro Marco, and Roberto Casas. 2016. Online social networks as a tool to support people with special needs. *Computer Communications* 73 (2016), 315–331. doi:10.1016/j.comcom.2015.09.020
- [8] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa
- [9] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis*. American Psychological Association. doi:10.1007/978-1-4614-5583-7_311
- [10] Phillip Brooker, Julie Barnett, Timothy Cribbin, and Sanjay Sharma. 2016. Have we even solved the first ‘big data challenge?’ Practical issues concerning data collection and visual representation for social media analytics. In *Digital methods for social science: An interdisciplinary guide to research innovation*. Springer, 34–50. doi:10.1057/9781137453662_3
- [11] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201* (2023).
- [12] Xiang’Anthony’ Chen, Jeff Burke, Ruofei Du, Matthew K Hong, Jennifer Jacobs, Philippe Laban, Dingzeyu Li, Nanyun Peng, Karl DD Willis, Chien-Sheng Wu, et al. 2023. Next Steps for Human-Centered Generative AI: A Technical Perspective. *arXiv preprint arXiv:2306.15774* (2023). arXiv:2306.15774
- [13] Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. 2023. The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences* 120, 51 (2023), e2316205120. doi:10.1073/pnas.2316205120
- [14] Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. 2023. LLM-Assisted Content Analysis: Using Large Language Models to Support Deductive Coding. *arXiv preprint arXiv:2306.14924* (2023).

- [15] Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging Large Language Model for Thematic Analysis. *arXiv preprint arXiv:2310.15100* (2023).
- [16] Dipto Das and Bryan Semaan. 2020. quoras: A Python API for Quora Data Collection to Increase Multi-Language Social Science Research. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*. 251–256. doi:10.1145/3406865.3418333
- [17] Stefano De Paoli. 2023. Performing an Inductive Thematic Analysis of Semi-Structured Interviews With a Large Language Model: An Exploration and Provocation on the Limits of the Approach. *Social Science Computer Review* (2023), 08944393231220483.
- [18] Stefano De Paoli. 2023. Writing user personas with Large Language Models: Testing phase 6 of a Thematic Analysis of semi-structured interviews. *arXiv preprint arXiv:2305.18099* (2023). arXiv:2305.18099
- [19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [20] A Ferreira and T Du Plessis. 2009. Effect of online social networking on employee productivity. *South African Journal of Information Management* 11, 1 (2009), 1–11. doi:10.4102/sajim.v11i1.397
- [21] Jon Fraenkel and Bernard Grofman. 2014. The Borda Count and its real-world alternatives: Comparing scoring rules in Nauru and Slovenia. *Australian Journal of Political Science* 49, 2 (2014), 186–205. doi:10.1080/10361146.2014.900530 arXiv:https://doi.org/10.1080/10361146.2014.900530
- [22] Deen Freelon. 2018. Computational research in the post-API age. *Political Communication* 35, 4 (2018), 665–668. doi:10.31235/osf.io/56f4q
- [23] Deen G Freelon. 2010. ReCal: Intercoder reliability calculation as a web service. *International Journal of Internet Science* 5, 1 (2010), 20–33.
- [24] Mircea Fuciu and Hortensia Gorski. 2013. Marketing research regarding the usage of online social networking sites by high school students. *Procedia Economics and Finance* 6 (2013), 482–490. doi:10.1016/s2212-5671(13)00166-4
- [25] Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhan, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2023. CollabCoder: A GPT-Powered Workflow for Collaborative Qualitative Analysis. *arXiv preprint arXiv:2304.07366* (2023). doi:10.1145/3584931.3607500
- [26] Norjihhan Abdul Ghani, Suraya Hamid, Ibrahim Abaker Targio Hashem, and Ejaz Ahmed. 2019. Social media big data analytics: A survey. *Computers in Human behavior* 101 (2019), 417–428. doi:10.1016/j.chb.2018.08.039
- [27] Qi Guo, Junming Cao, Xiaofei Xie, Shangqing Liu, Xiaohong Li, Bihuan Chen, and Xin Peng. 2024. Exploring the potential of chatgpt in automated code refinement: An empirical study. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*. 1–13. doi:10.1145/3597503.3623306
- [28] Suraya Hamid, Jenny Waycott, Sherah Kurnia, and Shanton Chang. 2015. Understanding students' perceptions of the benefits of online social networking use for teaching and learning. *The Internet and higher education* 26 (2015), 1–9. doi:10.1016/j.iheduc.2015.02.004
- [29] 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). doi:10.3390/bdcc7020062
- [30] Libby Hemphill, Margaret L Hedstrom, and Susan Hautaniemi Leonard. 2021. Saving social media data: understanding data management practices among social media researchers and their implications for archives. *Journal of the Association for Information Science and Technology* 72, 1 (2021), 97–109.
- [31] Qixuan Hou, Meng Han, and Zhipeng Cai. 2020. Survey on data analysis in social media: A practical application aspect. *Big Data Mining and Analytics* 3, 4 (2020), 259–279. doi:10.26599/bdma.2020.9020006
- [32] Muhammad Imran and Norah Mansour Almusharraf. 2024. Google Gemini as a next generation AI educational tool: a review of emerging educational technology. *Smart Learn. Environ.* 11 (2024), 22. https://api.semanticscholar.org/CorpusID:270035023
- [33] Andrej Karpathy. 2023. "The hottest new programming language is English". https://twitter.com/karpathy/status/1617979122625712128?lang=en [Twitter] Accessed: 2023-11-12.
- [34] Michelle E Kiger and Lara Varpio. 2020. Thematic analysis of qualitative data: AMEE Guide No. 131. *Medical teacher* 42, 8 (2020), 846–854.
- [35] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.
- [36] Ashutosh Kumar, Sagarika Singh, Shiv Vignesh Murty, and Swathy Ragupathy. 2024. The Ethics of Interaction: Mitigating Security Threats in LLMs. *ArXiv abs/2401.12273* (2024). arXiv:2401.12273 https://api.semanticscholar.org/CorpusID:267095035
- [37] Minhyeok Lee. 2023. A mathematical investigation of hallucination and creativity in gpt models. *Mathematics* 11, 10 (2023), 2320. doi:10.3390/math11102320
- [38] Thomas J Lee, Daniel J Campbell, Shriya Patel, Afif Hossain, Navid Radfar, Emaad Siddiqui, and Julius M Gardin. 2024. Unlocking Health Literacy: The Ultimate Guide to Hypertension Education From ChatGPT Versus Google Gemini.

- Cureus* 16 (2024). <https://api.semanticscholar.org/CorpusID:269632563>
- [39] Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2023. Calibrating LLM-Based Evaluator. *arXiv preprint arXiv:2309.13308* (2023).
- [40] Brady D Lund and Ting Wang. 2023. Chatting about ChatGPT: how may AI and GPT impact academia and libraries? *Library Hi Tech News* 40, 3 (2023), 26–29. doi:10.2139/ssrn.4333415
- [41] Ketut Mardiansyah and Wayan Surya. 2024. Comparative Analysis of ChatGPT-4 and Google Gemini for Spam Detection on the SpamAssassin Public Mail Corpus. (2024). doi:10.21203/rs.3.rs-4005702/v1
- [42] Pablo Martí, Leticia Serrano-Estrada, and Almudena Nolasco-Cirugeda. 2019. Social media data: Challenges, opportunities and limitations in urban studies. *Computers, Environment and Urban Systems* 74 (2019), 161–174. doi:10.1016/j.compenvurbysys.2018.11.001
- [43] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282. doi:10.11613/bm.2012.031
- [44] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]
- [45] Anna-Marie Ortloff, Matthias Fassl, Alexander Ponticello, Florin Martius, Anne Mertens, Katharina Krombholz, and Matthew Smith. 2023. Different Researchers, Different Results? Analyzing the Influence of Researcher Experience and Data Type During Qualitative Analysis of an Interview and Survey Study on Security Advice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.
- [46] Giada Pistilli. 2022. What lies behind AGI: ethical concerns related to LLMs. <https://api.semanticscholar.org/CorpusID:248913224>
- [47] Kevin Proudfoot. 2023. Inductive/Deductive hybrid thematic analysis in mixed methods research. *Journal of Mixed Methods Research* 17, 3 (2023), 308–326.
- [48] Zhangyang Qi, Ye Fang, Mengchen Zhang, Zeyi Sun, Tong Wu, Ziwei Liu, Dahua Lin, Jiaqi Wang, and Hengshuang Zhao. 2023. Gemini vs GPT-4V: A Preliminary Comparison and Combination of Vision-Language Models Through Qualitative Cases. *ArXiv abs/2312.15011* (2023). arXiv:2312.15011 <https://api.semanticscholar.org/CorpusID:266550760>
- [49] Youyang Qu, Ming Ding, Nan Sun, Kanchana Thilakarathna, Tianqing Zhu, and Dusit Tao Niyato. 2024. The Frontier of Data Erasure: Machine Unlearning for Large Language Models. *ArXiv abs/2403.15779* (2024). arXiv:2403.15779 <https://api.semanticscholar.org/CorpusID:268681648>
- [50] Taha H Rashidi, Alireza Abbasi, Mojtaba Maghrebi, Samiul Hasan, and Travis S Waller. 2017. Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. *Transportation Research Part C: Emerging Technologies* 75 (2017), 197–211. doi:10.1016/j.trc.2016.12.008
- [51] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).
- [52] Jürgen Rudolph, Shannon Tan, and Samson Tan. 2023. War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *Journal of Applied Learning and Teaching* 6, 1 (2023).
- [53] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15. doi:10.1145/3411764.3445518
- [54] Androniki Sapountzi and Kostas E Psannis. 2018. Social networking data analysis tools & challenges. *Future Generation Computer Systems* 86 (2018), 893–913. doi:10.1016/j.future.2016.10.019
- [55] Fabian Schneider, Anja Feldmann, Balachander Krishnamurthy, and Walter Willinger. 2009. Understanding online social network usage from a network perspective. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*. 35–48. doi:10.1145/1644893.1644899
- [56] Hong Shen, Tianshi Li, Toby Jia-Jun Li, Joon Sung Park, and Diyi Yang. 2023. Shaping the Emerging Norms of Using Large Language Models in Social Computing Research. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 569–571.
- [57] Chareen L Snelson. 2016. Qualitative and mixed methods social media research: A review of the literature. *International Journal of Qualitative Methods* 15, 1 (2016), 1609406915624574. doi:10.1177/1609406915624574
- [58] Nedson D. Soares, Regina Braga, José Maria N. David, Kenya B. Siqueira, and Victor Stroele. 2023. Data Analysis in Social Networks for Agribusiness: A Systematic Review. *IEEE Access* 11 (2023), 8422–8432. doi:10.1109/ACCESS.2023.3237984
- [59] Stefan Stieglitz, Milad Mirbabaie, Björn Ross, and Christoph Neuberger. 2018. Social media analytics—Challenges in topic discovery, data collection, and data preparation. *International journal of information management* 39 (2018), 156–168. doi:10.1016/j.ijinfomgt.2017.12.002
- [60] Hendrik Strobel, Albert Webson, Victor Sanh, Benjamin Hoover, Johanna Beyer, Hanspeter Pfister, and Alexander M Rush. 2022. Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. *IEEE*

- transactions on visualization and computer graphics* 29, 1 (2022), 1146–1156. doi:10.1109/tvcg.2022.3209479
- [61] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models*. <https://crfm.stanford.edu/2023/03/13/alpaca.html> 3, 6 (2023), 7.
- [62] Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588* (2023).
- [63] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [64] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.
- [65] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382* (2023).
- [66] Gregor Wiedemann, Felix Victor Münch, Jan Philipp Rau, Phillip Kessling, and Jan-Hinrik Schmidt. 2023. Concept and challenges of a social media observatory as a DIY research infrastructure. *Publizistik* (2023), 1–23. doi:10.1007/s11616-023-00807-6
- [67] Zhiheng Xi, Zheng Rui, and Gui Tao. 2023. Safety and Ethical Concerns of Large Language Models. In *China National Conference on Chinese Computational Linguistics*. <https://api.semanticscholar.org/CorpusID:261341825>
- [68] 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*. 75–78. doi:10.1145/3581754.3584136 arXiv:2304.10548
- [69] Gyeong Hoon Yi, Jiwoo Choi, Hyeongyun Song, Olivia Miano, Jaewoong Choi, Kihoon Bang, Byungju Lee, Seok Su Sohn, David Buttler, Anna Hiszpanski, et al. 2024. MaTableGPT: GPT-based Table Data Extractor from Materials Science Literature. *arXiv preprint arXiv:2406.05431* (2024).
- [70] Zohreh Zahedi and Rodrigo Costas. 2018. General discussion of data quality challenges in social media metrics: Extensive comparison of four major altmetric data aggregators. *PLoS one* 13, 5 (2018), e0197326. doi:10.1371/journal.pone.0197326
- [71] He Zhang, Chuhaio 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 preprint arXiv:2309.10771* (2023).
- [72] Min Zhang and Juntao Li. 2021. A commentary of GPT-3 in MIT Technology Review 2021. *Fundamental Research* 1, 6 (2021), 831–833. doi:10.1016/j.fmre.2021.11.011
- [73] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).
- [74] Mingkai Zheng, Xiu Su, Shan You, Fei Wang, Chen Qian, Chang Xu, and Samuel Albanie. 2023. Can gpt-4 perform neural architecture search? *arXiv preprint arXiv:2304.10970* (2023).

Received January 2024; revised July 2024; accepted October 2024