



Large Language Models in Qualitative Research: Uses, Tensions, and Intentions

Hope Schroeder*
MIT
Cambridge, Massachusetts, USA
hopes@mit.edu

Marianne Aubin Le Quéré*
Information Science
Cornell University
Ithaca, New York, USA
msa258@cornell.edu

Casey Randazzo
Rutgers University
New Brunswick, New Jersey, USA
cer124@rutgers.edu

David Mimno
Cornell University
Ithaca, New York, USA
mimno@cornell.edu

Sarita Schoenebeck
School of Information
University of Michigan
Ann Arbor, Michigan, USA
yardi@umich.edu

Abstract

Qualitative researchers use tools to collect, sort, and analyze their data. Should qualitative researchers use large language models (LLMs) as part of their practice? LLMs could augment qualitative research, but it is unclear if their use is appropriate, ethical, or aligned with qualitative researchers' goals and values. We interviewed twenty qualitative researchers to investigate these tensions. Many participants see LLMs as promising interlocutors with attractive use cases across the stages of research, but wrestle with their performance and appropriateness. Participants surface concerns regarding the use of LLMs while protecting participant interests, and call attention to an urgent lack of norms and tooling to guide the ethical use of LLMs in research. We document the rapid and broad adoption of LLMs across surfaces, which can interfere with intentional use vital to qualitative research. We use the tensions surfaced by our participants to outline recommendations for researchers considering using LLMs in qualitative research and design principles for LLM-assisted qualitative research tools.

CCS Concepts

• **Human-centered computing** → **HCI design and evaluation methods**; *Empirical studies in HCI*.

Keywords

Qualitative methods, large language models

ACM Reference Format:

Hope Schroeder, Marianne Aubin Le Quéré, Casey Randazzo, David Mimno, and Sarita Schoenebeck. 2025. Large Language Models in Qualitative Research: Uses, Tensions, and Intentions. In *CHI Conference on Human Factors in Computing Systems (CHI '25)*, April 26–May 01, 2025, Yokohama, Japan. ACM, New York, NY, USA, 17 pages. <https://doi.org/10.1145/3706598.3713120>

*Both authors contributed equally to this research.



This work is licensed under a Creative Commons Attribution 4.0 International License. *CHI '25, Yokohama, Japan*

© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1394-1/25/04
<https://doi.org/10.1145/3706598.3713120>

1 Introduction

The qualitative research process is grounded in deep engagement with participants and their data through an iterative sensemaking process. Today, Large Language Models (LLMs), particularly those accessed through chat-based interfaces like ChatGPT [59], are popular tools in research-related tasks, including within the social sciences [36, 50]. While a growing number of dedicated tools for qualitative work use LLMs to streamline research processes [19, 30, 32], direct interactions with ChatGPT and similar interfaces may also be increasingly used by qualitative researchers, a topic which has not received thorough investigation to this point.

LLMs distinguish themselves from prior tools for qualitative research. First, chat-based LLMs are marketed as general purpose tools, not dedicated qualitative analysis tools. Users interact with them in natural language, unlike most dedicated software platforms. LLMs go beyond the simple document-organization or analysis functions in classic tools, and facilitate highly flexible interpretations of documents, such as the ability to produce summaries or annotations, behaviors which mimic aspects of the human sensemaking process. Second, LLM use is currently dominated by one product, ChatGPT, that has rapidly entered general use by the public [63], including by qualitative researchers. Third, the dominant user interface metaphor [55] of ChatGPT and its competitors is that of a jack-of-all-trades assistant and/or collaborator. This pattern is intuitive thanks to the conversational interface, yet completely unguided. With wide-ranging and open-ended interactions with LLMs increasingly permeating everyday life, qualitative researchers are left to wonder whether or how these tools might apply to their research process or outputs. While some qualitative researchers support integrating LLMs into aspects of qualitative research, others, as Soden et al. [71] highlight, will be “concerned that increasingly all qualitative research, including that which draws on interpretive traditions of research design, is being evaluated from the perspective of positivism.”

There is a critical and immediate need to understand the degree to which LLMs are shaping qualitative research processes. On the one hand, many researchers may benefit from the inclusion of LLM-based features in qualitative data analysis (QDA) systems, as has been explored in HCI literature [30, 43, 60]. However, the recent explosion of work using LLMs to annotate text in more complex ways

than previously possible [6, 80] may risk the excitement of communities of scholars and tool builders without full consideration for potential tradeoffs to the sensemaking process. In particular, readily available tools that seek to automate the qualitative sensemaking process could decrease the amount of time researchers spend with their participant data, a potential risk to qualitative researchers' depth of understanding.

Furthermore, the speed of LLM development has outpaced guidance on their ethical use [9]. Most policies currently guiding use focus on the danger of uploading sensitive information to cloud services [58] or text generation in publications [39]. The HCI community is contending with the need for developed policies that tackle how to use AI ethically in research [5, 70]. While guidelines are beginning to emerge, there are no guidelines specific to qualitative work today.

This work critically reflects on the role of LLMs in qualitative research and offers considerations and recommendations for researchers considering using LLMs. Through our work, we answer the following research questions:

- RQ1: In what ways are LLMs being adopted into qualitative research processes today, and how might they be incorporated in the future?
- RQ2: What do qualitative researchers say they stand to lose or gain from adopting LLMs into their research processes, and what tensions might arise?
- RQ3: What are potential models for the ethical incorporation of LLMs into the qualitative research process?

Qualitative research is a broad and rich tradition with scholars from varied disciplines and ways of thinking. This paper focuses on the subset of researchers, including mixed-method researchers, who are considering using LLMs in their qualitative research processes. Those researchers should be clear-eyed about what is gained and lost when LLMs are used in data collection, analysis, or writing.

This work surfaces both the curiosities and concerns of qualitative and mixed-method researchers across disciplines who are wrestling with what the introduction of LLMs means for their qualitative research process. Through twenty semi-structured interviews with researchers across academic communities, we sought to understand any current use, motivations for use, perceptions of usefulness, and concerns about the use of LLMs in qualitative research. Our findings show that qualitative researchers can envision leveraging LLMs for their work to help with a range of uses including generating interview materials and analyzing research artifacts. Specifically, the interactivity of ChatGPT has spurred researchers to think creatively about what chat-based interfaces can bring to their work. Simultaneously, researchers harbor fundamental concerns about ethics, unequal adoption of new technologies, model bias, and performance, leading some to abstain from the use of LLMs to conduct qualitative research, and others to experiment despite concerns regarding participant privacy. New tools and practices are rapidly being adopted, but scholarly communities are not yet equipped to conduct ethical and effective research with them. As such, we are at a junction that urgently requires thorough evaluation and recalibration of both our methodological approaches and our thinking on research practices. As a step forward, we outline considerations and recommendations for future qualitative

researchers interested in using LLMs, and we present design opportunities and principles for creating LLM-based qualitative data analysis tools given our findings.

2 Related Work

Qualitative research is a broad term with different definitions and applications across disciplines [37]. In the late 1960s, the method emerged in retaliation to quantitative and positivist approaches to science characterized by hypothesis testing and objectivity [38]. HCI scholars often turn to qualitative methods when studying populations or contexts that require deep exploration and understanding of experiences, behaviors, and interactions [7, 8, 27, 64–66].

In contrast to quantitative methods that prioritize numerical data and generalizability, qualitative research aims to uncover how participants perceive, understand, and make meaning of their experiences [41]. Qualitative researchers often utilize diverse techniques, such as interviews, observations, and narratives [3, 57] to capture the complexity of lived experiences [15]. The epistemological foundation of qualitative research is *social constructionism* [11], in which meaning and knowledge is subjective and constructed in and through communication [78], enabling researchers to capture the social contexts that shape participants' lives [74]. Through deep engagement with participants' worlds, qualitative scholars seek to describe and interpret phenomena in a way that reflects participants' understanding of their experiences [41], while also contributing to theory development through inductive analysis [69].

As a field, HCI is both home to researchers with qualitative approaches as well as mixed-method researchers, a fact which has sometimes led to tension in the community regarding how contributions should be evaluated. In particular, a stigma surrounding the perceived rigor of qualitative research [18] has sparked debates over how best to evaluate such methods [46, 51, 67]. As a result, conducting qualitative research can involve a complex and sometimes creative negotiation with the norms and evaluative criteria of one's research communities [21]. Within the HCI community, there have been concerns that increasingly quantitative evaluation techniques, such as providing participant demographics, are being unduly required of qualitative scholars [71]. LLMs may lend the impression that qualitative inquiry can be automated, and their integration into Qualitative Data Analysis (QDA) software may increasingly impose positivist approaches that conflict with interpretivist traditions.

Approaches to analyzing qualitative data (e.g., grounded theory, thematic analysis) were originally designed for manual, pen-and-paper applications [22]. In the long history of qualitative research practice, many new technologies have been introduced, including the recording and digitization of interview data, popularization of automated transcription, and introduction of phone interviews, eventually followed by online interviews [16, 17]. The widespread adoption of software tools for qualitative data analysis marked a significant and, at times, contentious shift in the field [34]. Early critiques of qualitative software centered around concerns that these tools created too much distance from the data [33], compromised the depth of analysis [13], or could be misused by researchers who lacked an understanding of them [34].

Advancements in natural language processing (NLP) and machine learning have widened methods for data analysis over the past several decades, and the sudden rise of LLMs has, once again, opened up new possibilities. LLMs are being used for text annotation in interpretive contexts [6, 80], including qualitative coding [19, 23]. Past work has considered both the potential tensions and compatibilities between computational tools and interpretive work, grounding the conversation regarding the potential usefulness of computational tools in ways that can support qualitative work [10, 56]. In HCI, tools that leverage advancements in machine learning have been developed to assist qualitative researchers with data and text analysis since before the recent LLM era. The new generation of LLMs has ushered in new kinds of tools, with developers creating tools for qualitative researchers that bear researcher goals in mind, like agency over the coding process [60], convergent discussion on coding among collaborators [30], or incorporating trauma-informed design principles [75]. Studies introducing and evaluating these tools contain insights regarding the promises and pitfalls of LLMs in qualitative research, but largely focus on users' perceptions of the particular tools being developed for a task-specific use case, such as data annotation. Given the sudden, widespread ubiquity of LLMs through task-agnostic interfaces like ChatGPT, experiences with AI tools have become more common while being less scaffolded by purpose-built software, warranting investigation of current LLM use by qualitative researchers in a task-agnostic context.

Beyond the evaluation of specific software interfaces, prior work has also explored the concept of using AI tools to augment qualitative work. Jiang et al. [43] and Feuston and Brubaker [26] tackle the question of how qualitative researchers may respond to incorporating AI in their workflows. Jiang et al. [43] identify that AI tools for qualitative work should “honor serendipity, human agency, and ambiguity.” Feuston and Brubaker [26] highlight that intentional use and ordering of AI tools in qualitative tasks is paramount, and that the use of any computational tools may entail shifts in qualitative processes towards scale, abstraction, and delegation. These works lay key groundwork for this study, but were both published prior to the popular adoption of LLMs. Both papers presuppose the use of AI in the qualitative research workflow in bounded, task-specific ways, an idea which largely contrasts with the fluid ways that researchers often use LLMs throughout workflows today. The sudden adoption of LLMs warrants an urgent re-examination of AI tools in qualitative work, with early research suggesting that 81% of researchers use LLMs in their research practices [50]. Many papers have examined the application of an LLM, or of ChatGPT in particular, to the problem of a thematic analysis for an applied case, with examples across fields as varied as medicine and education [62]. Researchers often report promise regarding LLMs' ability to capture some or most themes humans do, but also perceive a lack of depth, nuance, or variety in identified themes [42, 48, 53, 62, 72, 76]. Simultaneously, some work raises or notes researchers' concerns regarding the reliability and validity of using LLMs in qualitative research, and privacy concerns that may arise when using them [28, 79]. As LLMs permeate society, sociotechnical gaps [1] may arise from the use of popular tools like ChatGPT for qualitative research purposes, where the social expectations of research do not match the underlying technical capabilities [24, 49]. In light of

these societal shifts, we endeavor a fresh examination of the current uses, attitudes, and tensions that arise when using LLMs in qualitative research. Through highlighting a range of considerations for qualitative researchers to scaffold their intentional use of these tools, we chart a path forward for using LLMs conscientiously in qualitative work.

3 Methods

To understand how qualitative researchers are interacting with generative AI tools or choosing not to, we conducted a series of twenty semi-structured interviews in June-August 2024. We analyzed the interviews using inductive thematic analysis, inspired by grounded theory [14], in which concepts emerge from the data [41]. These interview procedures were approved by the Institutional Review Board at Cornell University.

Participants were recruited through personal networks, email outreach about past work, and through in-person events and conferences. The inclusion criterion for this interview was simple: we sought to speak with researchers who consider themselves qualitative scholars, including mixed-method researchers. Since we wanted to include a variety of perspectives, we did not recruit participants based on their use or knowledge of LLMs, and we recruited participants across academic domains. Since qualitative research perspectives can vary by field, we considered it important to allow potential interviewees to self-define as qualitative researchers.

These recruitment strategies resulted in twenty participant interviews, summarized in Table 1. Our participants spanned seven academic fields, include sixteen PhD students, and are balanced between nine primarily qualitative and eleven mixed-method scholars. Beyond this high-level information, we did not explicitly collect demographic information from participants. Nonetheless, we note that a majority of participants presented as women and as researchers of color, and five participants referenced being non-native English speakers in interviews. Outreach in our personal networks resulted in a population that was largely located in North American Universities, with one exception.

The interview protocol was designed to elicit perspectives on LLMs in qualitative research, with a focus on the data collection and analysis processes. For this study, data collection was defined as *any part of the research process that contributes towards how researchers obtain data from participants*. Data analysis was defined as *any part of the research process that relates to the analysis of gathered data*. We opened the interview with questions about the researcher's qualitative work, practices, and their general familiarity with large language models. We then asked questions about data collection and analysis, asking the researcher if they have or could imagine using LLMs for any portion of these research processes. If they said they had used LLMs for that purpose, we asked them about their motivations for doing so, experiences of doing so, including perceptions of model performance, and ethical considerations during the process. If they had not used LLMs for the steps we outline, we asked them why they had not, and whether they had concerns with others using LLMs for these purposes. We concluded by discussing the participants' overall perceptions of LLMs in qualitative research, and soliciting both their perceived potential use cases as well as

Participant ID	Academic Domain	Position	Approach to Research
P1	HCI	PhD Student	Mixed-Method
P2	Communication	PhD Student	Mixed-Method
P3	Political Science	Research project manager	Primarily Qualitative
P4	HCI	PhD Student	Mixed-Method
P5	Social Work	PhD Student	Mixed-Method
P6	Communication	PhD Student	Primarily Qualitative
P7	Political Science	PhD Student	Mixed-Method
P8	HCI	PhD Student	Primarily Qualitative
P9	HCI	Professor	Mixed-Method
P10	HCI	PhD Student	Primarily Qualitative
P11	HCI	PhD Student	Primarily Qualitative
P12	HCI	PhD Student	Primarily Qualitative
P13	Sociology	Research Scholar	Mixed-Method
P14	Political Science	PhD Student	Mixed-Method
P15	Sociology	Former Professor	Primarily Qualitative
P16	Communication	PhD Student	Primarily Qualitative
P17	Sociology	Post Doc	Mixed-Method
P18	Political Science	PhD Student	Mixed-Method
P19	Anthropology	PhD Student	Primarily Qualitative
P20	Psychology	PhD Student	Mixed-Method

Table 1: Participant Summary Table. Participants self-reported their academic domains, position, and approach to research.

their concerns. Interviews lasted 42 minutes on average (min: 24, max: 59). The full interview protocol is included in the Appendix.

We iteratively analyzed the data using inductive thematic analysis techniques inspired by grounded theory [14]. While data collection was ongoing, three researchers open coded two interviews each, and each compiled a proposed codebook centered around identifying main themes, which were consolidated into one hierarchical codebook. These initial themes motivated us to balance recruitment between computing and non-computing scholars, and participants with varying approaches to qualitative research. Data collection continued until we obtained a balance of academic domains and research approaches, and the main identified themes remained stable indicating theoretical saturation [35]. Using our first-round codebook as a starting point, two further interviews were fully coded by three researchers each, and the codes were refactored to yield the final codebook. We then divided the twenty interviews among coders, and met regularly to discuss the process and update codes where needed. The two primary authors re-reviewed each transcript and the applied codes for thoroughness, verifying with each other when uncertain. Throughout the analysis process, we also constructed various artifacts, including interview notes, memos, and tables that summarized key context about each participant. We used Dedoose for collaborative code application to interview transcripts. We primarily avoided using LLMs for this research, with the exception of finding specific works of related literature, which we then verified externally. To protect the identity of participants, any specific research studies or topics have been changed.

We reflect on how our own positionality influenced this study. The authors are mixed-method researchers in HCI, NLP, and communication. In our own work, we have felt interest in incorporating LLMs into our research processes, while being unsure how to ethically navigate adopting this technology. This work is an effort to consider these opportunities and challenges head-on and to shine a light on the value of qualitative work in HCI.

4 Findings

In these findings, we outline how the researchers we interviewed adopted LLMs into qualitative research, identifying their main risks and concerns, then concluding by reflecting on the fundamental philosophical tensions raised.

4.1 Current and Potential Uses of Qualitative Researchers Adopting LLMs

We sought to understand how our participants had or had not experimented with using LLMs in their qualitative research process. Of the researchers we interviewed, fourteen reported actively exploring the use of LLMs in data collection or analysis. The most common use cases mentioned, highlighted in Figure 1, were using LLMs to help generate recruitment materials, attempts to use LLMs to speed up qualitative coding, and using LLMs for ideation and feedback.

ChatGPT was the most frequently mentioned LLM tool by participants. When we opened the conversation by asking participants about their experiences with “generative AI tools or large language models,” most would immediately talk about ChatGPT. Some researchers specified the model they used; P11 opened with “*I use LLMs, specifically ChatGPT.*” Others spoke more hesitantly, such as P13: “*I’ve started using ChatGPT, I’m trying to be more familiar with it because it is the future, and you don’t want to be a dinosaur.*” Overall, perceptions of LLMs writ large were strongly influenced by experiences with ChatGPT as the most popular interface through which to explore LLMs. Participants felt that LLMs were more advanced in areas where ChatGPT performed better, thanks to their interactions directly through the web interface. For example, more participants felt that LLMs were helpful “*brainstorming buddies*” (P20), than felt they were helpful annotators, a use case for which ChatGPT’s web interface was not designed.

Outside of ChatGPT, participants shared that they used a suite of available tools for qualitative research, such as NVivo, Atlas.ti, MAXQDA or Dedoose, and AI sometimes underpins these tools’ features. Participants were sometimes unsure whether these tools

leveraged LLM-based functionality. For example, participants (P4, P8, P9, P13) wondered: “*does Grammarly count?*” Others mentioned explicitly that they engaged in AI features offered by qualitative coding platforms. For example, P7 reported using the in-context code suggestions that Atlas.ti provides. Participants did not necessarily feel agency over the AI features embedded in these tools, as P1 explained: “*Atlas sends me a thematic analysis that essentially aggregates stuff, I don’t really like it,*” and P4 summarized AI features in these tools as something they see “*not by choice.*” Overall, participants often relied on software platforms to help with qualitative research, though not all appreciated the apparent increase in “smart” AI-based features.

In contrast, some researchers shared they purposefully avoid advanced technology in their qualitative research process. While all participants used their computers for analysis, they also described paring back their approach to remain close to the data, particularly to data from participant interviews. P6 explained that while they were “*not quite as old school as the people who would cut out their coded texts and move it around, I basically do that in Word documents.*” P2 echoed this sentiment: “*I didn’t like the restrictions and things the platform was putting on me, and I got really frustrated; I’m just familiar with Excel.*” These attitudes highlight how, to some researchers, complex software interferes with their process or can be hard to learn, and simpler solutions may meet their needs.

4.1.1 LLMs during Data Collection. Seven participants reported using LLMs to help with creating or editing recruitment or interview materials. Often, this assistance took the form of researchers using a service like ChatGPT to edit or tweak their writing. Even those who were reluctant to use LLMs for research conceded that they could be helpful for writing, such as P10 who said “*the only real reason that I use ChatGPT is to have it reword or rephrase,*” and P4 explained that “*it will give me a really good sentence or switch a word.*” This capability was particularly helpful for researchers whose native language is not English; for example, P8 said they “*feel more confident once [they] have passed [their] writing to the AI tools.*” LLMs also assisted participants in tailoring recruitment materials to their subject population. P9 (a non-American, non-native English speaker) told us they used LLMs to make messaging more appropriate for American research participants. They shared, “*one thing I noticed that LLMs can do beautifully is make things American,*” and therefore “*an LLM’s general answer or general email to an American future participant will probably be better than my take on it.*” P5 reported that LLMs helped them deploy recruitment materials in multiple languages, explaining that they “*used it for the translation of interviews and consent forms and other required documents,*” before having a native speaker collaborator double check the output. LLMs were also used by participants to help them create interview and recruitment materials from scratch. For example, P8 recalled “*I was working with a deadline, and I was feeling stuck. I was not sure what kinds of questions to ask, but I did have a set of broad topics [...], so I put those into ChatGPT and I was like, ‘Can you help me come up with this type of interview question?’*” While quite a few of our participants had experimented with using these tools to help generate research materials, not all found these tools helpful: “*the performance was really great,*” P5 said, while P8 thought “*I feel like I could have done better,*” and P10 elaborated that

its outputs were “*really broad, versus the empirical work I’ve done is very specific and granular.*”

In terms of promising potential uses, researchers seemed to understand why one might use LLMs to help create recruitment materials. Some experienced researchers explained that they had past recruitment templates already, but for those who do not, P10 said, “*I can imagine that maybe using something like this would be helpful for beginner researchers.*” P20 expressed a similar attitude: “*I would consider it in the future, if I’m drafting a new consent letter or things like that, but we had all of that from previous ones.*” Nonetheless, some researchers pointed to specific reasons why they may not consider LLMs appropriate for these use cases, such as when materials are being produced for at-risk or non-Western populations, issues which will be further discussed in the following sections.

4.1.2 LLMs during Data Analysis. Another avenue that participants had explored, though often less successfully, was using an LLM to process and make sense of their qualitative data. Many researchers we interviewed did not attempt to use LLMs in data analysis due to concerns about privacy and performance, issues we explore in the next section. Among those interested in LLM-facilitated analysis of qualitative data, we saw various functions being performed with LLMs: transforming or retrieving data, data exploration, generating initial codes or themes, and applying qualitative codes.

Researchers wanted to use LLMs as assistants for retrieving, processing, transforming, or reformatting their data using natural language. P9 reported that they leverage LLMs to perform fuzzy searches on qualitative data: “*With retrieval via an LLM, I can say something like, ‘Where in this document is an example of such and such?’ and just talk more conceptually or more abstractly, and the tool will find it.*” Participants also saw LLMs as being potentially useful to reformat data. For example, P4 imagined that “*it would be nice if there was like a large language model agent to grab all the qualitative codes [and put them] onto a Google Jamboard because I have to rely on copy pasting.*”

On the analysis front, some participants explained that they valued being able to use LLM-aided software to flexibly explore participant data. In this data exploration use case, participants reported using chat-based models like ChatGPT to conduct initial sensemaking with and about their data. In P15’s words, the ability to easily analyze qualitative data with an LLM was a “*game changer,*” since “*there’s no friction between the person trying to make sense of a lot of data and the information.*” One participant said they once used ChatGPT to help them make sense of the theory behind their qualitative data, explaining that they had interviewed people about their experiences with grief. They were “*getting all this lovely, well-gelled data where people were repeatedly saying the same things, everything was perfectly categorized*” but the researcher was “*having a hard time naming those categories in a way that was as inclusive as it could be.*” In this example, the participant asked ChatGPT to come up with a list of potential themes, based not on their participant data but on prior theories, and it gave them “*this list of things, and I was like, hmm, this is really neatly packaged [...] it’s giving me labels for the things I’m seeing that are also very intuitive.*” These themes then served as a jumping off point for deeper engagement with the data, and guided the researchers’ thinking for constructing their final qualitative codebook.

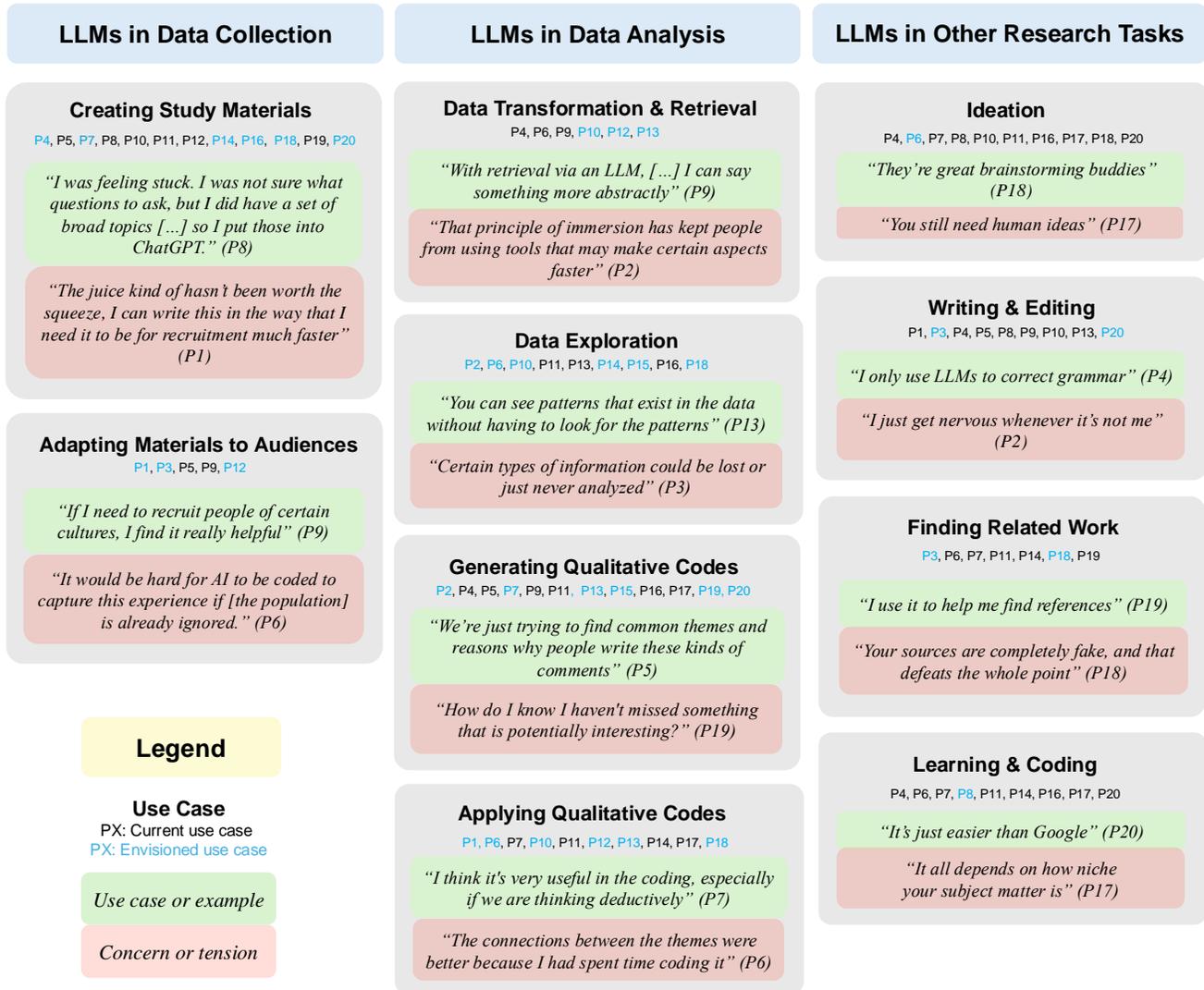


Figure 1: Current and Envisioned Use Cases for LLMs within the Qualitative Research Process. The figure highlights how participants had previously used and might imagine using LLMs moving forward in their research processes. We also highlight different perspectives and tensions for each use case, with participant quotes. Participants are only tagged under a use case if they view LLMs as being potentially beneficial to that end, and use cases are only included if mentioned by three or more participants.

Experiences were more mixed for participants who directly tried to use LLMs to generate or assign qualitative codes to data. In these use cases, participants were seeking to replace or augment the process of generating and/or applying a codebook to label qualitative data. The most common experience was perhaps that of P5, who “tried to use ChatGPT to get the very general theme and the results were not that great, so I decided to code it manually, but I tried it.” P9 engaged in a process they refer to as “semi-coding,” where they provided ChatGPT a chunk of text they already had intuitions about, and wanted to see if the LLM agreed with their perspective. Ultimately, however, they felt that the codes were “uninteresting

in a way, or at least they were first level codes.” This experience was also shared by P11, who explained that when they “asked it to assign some initial codes [on their dataset], oftentimes the response will be very high level and abstract, so it’s not trustworthy at all.” P17 disagreed, and had set up specific coding infrastructure that allowed them to identify whether an individual piece of social media data contains a rare occurrence of interest to their research. P17 was seeking to identify and code specific types of gender-based online harassment in Reddit comments and derived the most advanced annotation pipeline we saw:

“I come up with 16 or 18 thematic categories, and then use them to describe the system that I’m looking at. That’s qualitative work... I have... millions of comments, and most of those are garbage... So it’s a needle in a haystack, and computation allows you to find and organize all those needles so that you can do the qualitative work.”

In their workflow, P17 derived codes based on initial data, and then constructed an LLM pipeline to annotate each post, using the LLM to take a first pass at organizing the data. P17 expressed that this process does not preclude deep qualitative work, since they still do a manual layer of qualitative analysis.

Despite mixed current experiences using ChatGPT for data analysis, many participants felt curious or excited by the prospect of using LLMs to automate qualitative coding. The biggest factor that motivated the use of LLMs for annotation was the tedium and time investment required by manual coding. P10 said *“I’m currently in the process of coding, and sometimes I’m kind of like, it’s taking a long time.”* Others mentioned that automating coding could limit human error and help them optimize for inter-rater reliability. P13 coordinated a large project with multiple coders, and they described that *“having a more thoughtful upfront conversation about that, and then using a tool to execute on that [code scheme], I think, would be way more accurate and better science than what we ended up doing.”* Of those interested in using LLMs for coding, researchers had conflicting perspectives about whether LLMs would best be used for an initial, exploratory round of coding, or whether they should be used to apply manually developed codes. Some specifically identified that LLM-aided coding could help with the first round of coding, or when initially getting a grip on the data. P13 thought LLMs could *“save you from that first tedious round of coding and then you could maybe be looking at the more nuanced codes faster over time.”* Others described the desire to use LLMs as a way to “gut check” their coding. For example, P18 explained that they had considered if *“me and another human are coding the data, and then ChatGPT is also coding the data, and we do an intercoder reliability check with all of us.”* Nonetheless, some participants (P2, P4, P8, P16) struggled to see the automation of coding as aligned with their perception of qualitative work, stating *“there’s a part of qualitative research where it kind of feels like I’m in this for the manual work and that’s why I’m doing it”* (P2). Some perceived LLM involvement in some research tasks as fundamentally in tension with qualitative research values, which we will further revisit in a later section.

4.1.3 LLMs in Other Research Tasks. Beyond data collection and analysis, participants also reported leveraging the interactivity of readily available LLMs, particularly ChatGPT. Participants felt that ChatGPT could help them ideate better and be more creative researchers: *“I think LLMs are really great for bouncing ideas off of and thinking creatively,”* P18 said, while P16 said they use LLM outputs *“a basis or a starting point.”* Participants also explained that they used ChatGPT to generate explanations for complicated concepts or coding implementations, like P7, who explained that they ask ChatGPT to help explain specific theories, stating *“if I didn’t understand what kind of methods they used in a particular paper or [...] I’m not able to make some connections or linkages, I just pose a question to ChatGPT.”* In these instances, the social semblance of LLMs takes

center stage, and our participants would sometimes refer to ChatGPT as more of a collaborator or peer than a tool. For example, P18 referred to ChatGPT as a *“great brainstorming buddy,”* P20 called it a *“thought partner,”* P9 defined it as *“a rational person, a smart layman, if you will,”* and P12 said that working with ChatGPT was *“almost like working with another person.”* For P11, this connection went further, as they found that interacting with ChatGPT while doing research helped them stay engaged with their research:

“As a researcher, we are always heads down and maybe we do not have much interaction with peer researchers who are human, I feel like I really value this kind of interaction, somebody I can talk to. I feel like this research process is not boring, it’s engaging.”

However, some researchers were more critical of ChatGPT occupying these kinds of interpersonal roles. P5 expressed frustration that ChatGPT does not push back enough on their ideas to be a collaborator, since ChatGPT *“never criticizes me, even though I asked: pretty, pretty please criticize me.”* P4, meanwhile, was dismayed at the prospect of replacing any of their processes with ChatGPT, saying *“I have a very supportive advisor and group of collaborators; they will give me great human feedback and I wouldn’t I even think about using language models for that purpose.”*

4.2 Perceptions and Risks of LLM Use in Qualitative Research

While researchers identified potential use cases for LLMs in research, and expressed curiosity about new methods, most had not used LLMs extensively in research. During our interviews, participants shed light on barriers to adoption, including a lack of established norms and concerns about output quality and appropriateness.

4.2.1 Uneven adoption. Researchers reported logistical hurdles to adopting AI tools, which related to their comfort with technology and inequality around funding. For example, participants (P1, P3, P10, P13, P18) expressed that adopting new technologies takes time and effort: *“We’re going to have to learn this, it’s gonna take forever, and we’re never going to get around to doing the actual project”* (P13). These efforts could be even higher for researchers not in computing fields. Researchers also may not be certain how to find tools that meet their needs; P20 described how *“you don’t know what’s out there, sometimes you don’t even know what something’s called.”* P2 and P5 recounted that cost influenced their choice of models, P2 stating *“since the Claude model is quite expensive, I didn’t use it, and OpenAI models are much cheaper, so I could test it out.”* Overall, logistical issues led researchers to be more reliant on readily available interfaces like ChatGPT, even when these are not optimized for their use case, an issue exacerbated for non-technical researchers.

4.2.2 Uncertainty about norms, guidelines, and expectations. Participants shared that university or publication venue policies are shaping their perception of what is acceptable (P4, P11, P14), but many guidelines fall short. For example, participants referenced that university policies surrounding LLMs to produce writing informed their decisions about using LLMs for this use case, but also fuel uncertainty about using LLMs for other purposes. P14 reflected, *“I guess, as students, ... it’s kind of concerning that we might*

be penalized for using ChatGPT, so we're probably very, very careful" and "I don't want this to backfire and cause any problems." Sixteen participants in our study were graduate students, who occupy the unique role of collecting and analyzing study data directly, while feeling little power over the direct creation of guidelines. Some researchers mentioned efforts to seek additional clarification, such as P13 who reached out to legal experts in the university to clarify whether they are allowed to "use AI" when analyzing their data.

Some participants also raised concerns that LLMs introduce IP and originality issues that make them cautious about their use in research (P11, P14, P18, P20), particularly for tasks in research that create original ideas or findings. For example, P11 shared, "I feel like the IP is kind of mixed up... because ChatGPT does give me creative answers, but then sometimes I feel I just take it for granted, and I probably did not think about whether this creative answer has its own intellectual property." Opinions on acceptable use in light of this originality challenge were often inconclusive, with P19 asking, "Is it bad to ask it to do stuff, and then look it over, does that still count as your material? I think we're getting into those waters now." Others note that as LLMs are increasingly integrated into workflows, policies that require disclosure, like the ACM policy, may be hard to enforce in practice. P18 wrestled with this concept, too: "At some point it may be somewhat impossible to truly distinguish what a human contributed or thought about, from what an AI that they were working with contributed."

4.2.3 Navigating Participant Privacy Implications. Our participants were most concerned with how LLM use impacts data security, consent, and obligations to participants in qualitative research. Participants expressed a general lack of understanding and distrust for data protections awarded to participant data sent to proprietary models and AI companies. For example, P7 ruminated: "how will the companies use that data [...] how is their data management protocol? Because the data that we share with LLMs, we don't know how that is managed at the back end."

Researchers also wondered whether classic models of participant consent that cover sharing data with third party proprietary models like ChatGPT are doing enough to protect participants, like P8: "There's a bigger question of... primacy of what is going to be giving consent." Some participants held strong positions on protecting participant data by not sharing it with proprietary LLMs. P10 shared, "Obviously, I wouldn't use my participant data with GPT-4 and a lot of these other models where I don't know how they're taking care of the data." P18 expressed a similar strength of conviction: "I can't feed my data into ChatGPT, or something like that, because of confidentiality concerns." In the absence of clear, universal guidelines from IRBs within universities or venues providing recommendations for these issues, some researchers are creating their own strong norms: "I never put the data that is private to the model," P5 confirmed.

P7 and P13 highlighted that it is currently challenging to understand if LLM-based features are integrated into existing QDA tools in a way that protects participant data. For example, P13 noted that MAXQDA now has LLM-based features, but their team chose not to use them because they were unable to confirm the legality of doing so based on their data agreement and privacy standards. P7 echoed this, sharing their concern is not "just the researcher directly using ChatGPT, but also the researcher using tools which use

ChatGPT further." Researchers were also concerned these risks may be exacerbated for the populations of vulnerable groups qualitative research often seeks to study. Researchers in our study work with vulnerable populations, such as undocumented immigrants, those in political exile, and those with disabilities. The researchers studying these three populations shared explicit concern that using LLMs could put those participants' privacy at risk. Given the potentially grave consequences of identity exposure for individuals in those vulnerable groups, these researchers strongly felt those participants' data should not be used with LLMs readily available now.

To mitigate these risks, some researchers reported experimenting with ways to honor privacy while still using these tools. P11 shared, "Sometimes I do put some participant's quotations in, but I intentionally trim out specific keywords, for example, specific product names or functions." This participant is also an example of the multiple participants who reported believing it unethical to share participant data with third-party tools like ChatGPT, but who did not discuss potential alternatives, like open source models, that could address some of these concerns. This absence may be due to a combined lack of knowledge of privacy-preserving alternatives or the relative scarcity of mainstream, user-friendly alternatives to proprietary tools, particularly for users without technical backgrounds.

P9 said researchers should take a proactive stance towards participant privacy, particularly in the case of models that train on inputs: "Am I allowed to upload even anonymized content to an LLM? Usually you don't sign a consent specifically on that question. And maybe legally it's fine. But I feel like it shouldn't be shared with a third party that might use it as a training tool." Several proposed a stronger change to consent practices in anticipation of these concerns, like P7, who said "I think it's of the utmost importance to... let them know we would be following this method or doing this type of analysis wherein we are sharing some of their information with the LLMs." Broadly, in the absence of strong guidelines, researchers grapple with appropriate use of these tools while still following their own convictions and research norms, leading to variable perceptions of acceptable and non-acceptable use.

4.2.4 Concerns around validity and evaluation. Participants considered whether LLMs would produce valid, usable outputs for research. In particular, participants explained that the specific issues they feared or observed with LLM outputs included hallucinations, label assignment errors, or overly general responses. "LLMs can give you wrong information" said P18, and P4 agreed: "some points are just wrong." More broadly, participants often found that LLM—specifically ChatGPT— outputs were vague, odd, or missing key points: "they're getting some of these broad ideas, but maybe the variables or the concepts that are in this theory are a little off or they might be different theories," P6 shared, and P16 said "the LLMs tend to write really generically and I would have concerns." Researchers sometimes conclude these tools may not work for their qualitative research due to these limitations. Researchers shared that the practice of collecting data (like interviews) themselves gave them an intimate understanding of the data, which could help them guard against misinterpretation of the data through an LLM's hallucination or misconstrual (P16, P20).

Researchers also reported uncertainty regarding the best way to validate and evaluate LLM performance in a way that aligns with their research community. P18 shared that LLMs “are all constantly being updated [...] So the model is always changing, and that means that the output that you get is always changing.” In response to these concerns about validity, many participants emphasized that checking the LLM’s outputs is key to being confident in the use of these tools in research (P5, P7, P10, P14, P18, P19, P20) and some others also mention that it is unacceptable to not check outputs (P1, P5, P18). P10, for example, emphasized the need for researchers to take responsibility: “there needs to be some sort of oversight because there will be an error or mistake through using LLMs, and it’s on the researcher to catch those errors or to have some safeguard.”

4.2.5 Concerns about model bias. Researchers raised concerns that biases embedded in LLMs might make them unsuitable for qualitative research on populations not well-represented in training data (P6, P7, P8, P10, P12, P16, P18, P19) including non-English-speaking populations (P1, P12, P18, P19). Are LLMs “able to understand data from someone with a particular identity?” P8 wondered. Researchers reported both experiences with and fears of LLMs misinterpreting data from marginalized groups through the mistranscription of important language from that community, or by making judgments that are not faithful to the values in that community. For example, one participant noted “AI often falsely identifies queer content as hateful,” confirming that an LLM’s construals of their participant group may be at odds with faithful stewardship of that community’s interests. Another participant noted ChatGPT gave reductive information about women from different backgrounds experiencing grief. “Probably, if you’re saying, ‘what issues do women experience?’ [ChatGPT is] going to try and give you the most broad, all-encompassing answer. There’s not really any way to be sure that what it’s giving you is covering all the ground.” P19 cited concerns over known Western bias as the main reason they do not feel comfortable using LLMs in their work: “We all know that these models are pre-trained on datasets that are highly Western, mostly completely built from universities and companies that are based in developed places.” On their interest in using LLMs in qualitative research, they continued, “So for now, no, in the future I don’t know, maybe. We have a long way to go for non-English speaking languages.” Researchers worried that these LLM biases could perpetuate existing power inequalities or harms towards marginalized groups. P6 observed, “The machine is made within these power structures... so inherently it’s going to privilege some things over others because that’s how the world works,” expressing a concern also shared by P12. Noticing similar properties of LLM training data, P7 said, “Most of the literature that’s accessible is the voice of the powerful, right?” Researchers were attuned to the connection between LLM training data and model outputs, and thus using LLMs may not reflect researchers’ desire to responsibly analyze specific populations.

4.3 Qualitative Research and LLMs: Inherent Tensions?

Simultaneous with curiosity and exploration of LLM tools described above, participants grappled with potential tensions between the values of qualitative research and the capabilities LLMs may enable.

4.3.1 Emergent insights from the data, versus interpretations imposed by LLMs. Many qualitative researchers use methods where findings emerge from the data bottom-up, such as grounded theory, which encourages deep interaction with the data without imposing a view from pre-conceived theories. Researchers wondered whether bringing in the external interpretations, language, and potentially even theory from an LLM were in alignment with these qualitative approaches (P1, P3, P6, P9, P12, P16). P6 worried about letting LLMs lead theme generation: “My biggest worry is having AI code qualitative data: that feels a little bit different than engaging in a thematic analysis where we’re letting the data drive the themes and we’re recognizing that as humans, we’re bringing the meaning to those themes.” Using LLMs to create an interpretation of data could produce outright errors or produce standpoints that marginalize people. P16 said, “these people didn’t agree to editing their words or modifying their perspective” stressing the personal responsibility researchers feel in making sure participants are heard on their own terms. A few participants reflected that relying on a single language model or sole output to engage with qualitative data might reduce the impact of their own perspective. P1 shared: “There’s a little bit of like domain expertise that you’re able to apply as a researcher and spend time in the space and spend time with these individuals,” and in using LLMs to do this, P16 said, “I think we lose out on that, because now we’re going with this very centralized, generalized model.”

4.3.2 Close engagement with the data, versus LLMs that automate processing. Participants reported that directly collecting interviews and reviewing them manually enables sensemaking and synthesis (P4, P6, P8, P10, P16, P19, P20). P20 showcased this perspective, saying “I do think there’s something in looking at all the transcripts, having done all the interviews yourself, and having intuitions around what themes arise.” P10 also felt manual sensemaking was valuable, and tied this view to a potential tradeoff: “You also wonder: if an AI were to do it, would you lose some of that sensemaking that naturally happens when you’re individually doing it yourself? And could that lead to not a strong findings or not a strong discussion section?” Relatedly, some researchers worry that the ubiquity and ease of using LLMs in these tasks may lessen the degree of deep engagement with the data through manual sensemaking by allowing researchers to rely on “shortcuts.” P19 shared that if someone is “using [an LLM] as a shortcut for reading large-scale things [where] you can’t go through what’s generated, then I do have concerns.” P13 saw the main risks of involving LLMs in data analysis as only resolvable if researchers still take care to collect and review data themselves: “I’m sure there’s lots of ways this could get messed up, like only looking for certain words or words that go together and missing on nuance or interesting ways other people are talking about the same concept. If you’d never read it yourself... maybe something would get lost.” Researchers wrestled with whether deep knowledge of the data would be maintained if LLMs were used in the analysis process.

4.3.3 Relationship and sense of responsibility to participants, versus an emotionless machine. Deep engagement with participant data, as well as an often ongoing relationship with research subjects and their stories, means that many qualitative researchers internalize a responsibility to do justice to participants in their research process and to act in their best interests. P18 expressed a fear about how LLMs could change this relationship: “If people found out that I was

feeding data into an LLM, I think that can cause a breakdown of trust. Some of my participants are people I intend to continue working with in a more long term way.” When considering using LLMs for data analysis, P1 hesitated, saying “we’re not necessarily building empathy in the same way, for participants”, and P4 told us “it’s not doing the data justice, you know, by just having an LLM look at it.”

4.3.4 Subjectivity and multiplicity, versus one hegemonic way of seeing. Our participants valued the unique context they bring to their own qualitative work (P1, P6, P7, P8, P9, P12, P16, P19, P20). Participants acknowledged interactions with LLMs may uncover new insights about their data, but they were concerned those interactions could result in some insights being lost. For example, qualitative researchers see value in noticing things beyond the surface of transcribed interview text (P4, P6, P7, P10, P14, P16, P20), and note that LLMs do not see or know what humans do. P9 shared, “I think you do gain a lot of insights by not just the words that are typed: the way I hesitate on certain things like accents, body movement... there’s a tacit knowledge of the qualitative process that is not easily formalized and programmed.” As P7 put it, qualitative researchers “situate the text in a context,” and do not just analyze the “text as it is.” This led P7 to conclude that “in such cases, the LLMs might not be very helpful.”

Researchers note that LLMs may change what humans notice in their data, and could also potentially miss critical insights. P19 noticed: “There’s just something interesting that I see that it doesn’t give me, and the most important point of my analysis is to figure out the themes. If they’re figured out for me, how do I know that I haven’t missed something that is potentially interesting?” P3 also wondered if LLMs would focus on different insights than humans, who collaboratively notice and integrate findings together: “I think when you ask an LLM to give you a specific set of instructions, you lose some of that nuance and some of that discovery. So I definitely think certain types of information could be lost or just never analyzed.”

4.3.5 Who is in charge or creating meaning from human experience? Many researchers in our participant group discussed how qualitative methods are well-suited to understanding the *why* behind human experience, (P1, P2, P6, P10, P13, P16, P20), particularly in contrast to what quantitative approaches can tell us. One theme is resounding: qualitative researchers want to keep humans in charge and center their perspectives. Summarizing this point, P8 shared, “The whole point of doing qualitative research is to engage with people, and if we just stop doing that, then I feel like it endangers qualitative research in some way.” P15 took the view using LLMs in qualitative research may still be compatible with this goal as long as humans are the ones ultimately determining meaning from an analysis: “I think LLMs would be incredible for topical emergence. Only humans can really [tell] you what the themes are, ultimately.”

5 Discussion

Our interviews situate at least a partial answer on how qualitative researchers can leverage LLMs while staying faithful to responsible, participant-centered research. In response to RQ1, we find that many qualitative researchers are open to the conscientious use of LLMs across the research pipeline. RQ2 poses the question of which tensions might arise if they are used, and we find that qualitative

researchers harbor many valid questions and concerns about how to align LLM uses to their personal and disciplinary research values. RQ3 raises the idea of potential models for ethical use of LLMs in qualitative research, which we discuss in Section 6.

5.1 From Task-Specific to Ubiquitous AI

We build on prior work on human-AI collaboration in qualitative contexts [26, 43] by situating AI use in the current moment. Feuston and Brubaker [26] argued that successful human-AI collaboration for qualitative research requires the intentional use of AI for specific tasks, such that researchers are not “displaced as the primary analysts.” Jiang et al. [43] similarly suggest future research explore which precise sub-tasks of qualitative research are most appropriate for AI involvement. Additionally, many papers on human-AI collaborations reported that the order of AI involvement matters in qualitative research, since researchers do not want to be influenced by suggestions too early in their bottom-up sensemaking process [26, 43, 47, 60]. Aligning with this idea, our participants were comfortable with task-specific tools that use LLMs such as grammar check, speech-to-text transcription, and translation, and used them extensively.

However, our findings highlight that the ideal of intentionally using specific AI tools for clearly delineated research tasks is increasingly challenging to meet, due to the prevalence of AI through flexible, seemingly all-purpose chatbot interaction patterns popularized by ChatGPT. Participants used AI across their research pipeline, a finding also supported by Liao et al. [50]. A natural consequence of this shift is that the lines between AI and non-AI tools blur, making intentional use more challenging. Narayanan and Kapoor [54] speculate that generative AI will become “part of our digital infrastructure, instead of being a tool people use for specific purposes.” Our findings suggest that this shift is underway.

Nonetheless, this shift need not be solely negative. Our participants indicated that the “helpful assistant” nature of the most popular AI tools today collapse the boundary between seeing LLMs as a “tool” and as a “collaborator.” While there are significant risks of overly anthropomorphizing AI [52], these natural language interactions may also be the kind of human-AI collaboration that “support serendipity,” which Jiang et al. [43] champion.

5.2 The Sociotechnical Breakdown of Guidelines, Norms, and Tooling

Worries about adhering to norms and guidelines are ubiquitous, despite the active development and publication of norms by institutions and publication venues (e.g. ACM policies), ethical and supervisory boards, and academic research papers [39, 77]. In the social norm change literature, scholars posit that there is often a tipping point where social norms change rapidly [4]. Our research suggests that some norms regarding LLMs may be close to a clear tipping point, such as the understanding that identifiable participant data ought not be sent to a third-party tool that may train AI models on the data, while norms regarding some other concerns have so far eluded convergence. Our work confirms prior perspectives that AI evolution is outpacing our norm formation processes [9, 25], and we show that scholars feel pressure and uncertainty when justifying LLM-aided techniques for academic publications.

We show that LLMs are being broadly adopted in qualitative research practices despite wide-ranging concerns about privacy. Ongoing work suggests that non-computing researchers have more pronounced ethical concerns in using LLMs [50], and our findings suggest that ethics and privacy were top of mind for qualitative researchers more broadly. Nonetheless, the present research suggests that participants use or consider using LLMs *despite* these concerns, and sometimes use these tools while still being unsure about appropriate use. Kapania et al. [44] similarly find that researchers are often aware of potential ethical issues, but uncertain about how to address them. We position the current research environment as a fundamental sociotechnical breakdown: the available *tools* do not present researchers with the right options to confidently preserve privacy, the necessary *guidelines and norms* are lagging behind, and the *social context* is increasingly normalizing and even expecting the use of AI tools. This misalignment of incentives resembles Ackerman’s notion of the socio-technical gap [1], which refers to the gap between technical capabilities and social expectations.

While the imagination of the majority of researchers we spoke with was shaped by the current capabilities of ChatGPT, the most technologically savvy participants were currently leveraging LLMs to do more complex tasks including systematic qualitative coding. Several participants wanted to use LLMs to do qualitative coding, but did not know how to use current tools, indicating their lack of use now could be partly due to a skills gap. Thus, the deployment of intentionally designed tools like CollabCoder [29], SenseMate [60], and TIQA [75] is even more important, since the state of accessible tooling may be actively limiting responsible research. We also see evidence that this divide in adoption could, on the other hand, be more fundamental: some participants simply deemed AI tools for tasks like coding to be needless departures from their qualitative sensemaking process.

5.3 Fundamental Tensions in Using LLMs for Qualitative Work

A set of potentially fundamental tensions arose in considering the use of LLMs to conduct qualitative research. There has long been a tension between work seeking to combine computational power with interpretivist approaches [10, 56] and work arguing that qualitative values must be protected from quantitative evaluation standards [20, 71]. Our interviews suggest that LLMs may continue blurring the line between positivist and interpretivist approaches, as exemplified in emerging work [19, 47]. As Messeri and Crockett [52] discuss, the way some approaches are more readily aided by LLMs may change the research landscape by making some questions more tractable than others. There is a danger that the ability to analyze data at scale with less human involvement may make situated, slow, contextual and interpretivist approaches to qualitative data less attractive, potentially because of incentives to quickly publish, or because of the perceived objectivity of LLMs compared to human analysts. Mixed-method researchers, who make up roughly half of our participants, may be particularly positioned to accelerate this change. Published work uses LLMs for many types of text analysis, including in powerful new tools that are increasingly positioned to facilitate “theory-driven analysis” of textual data [47],

and these tools may be adopted by users from different epistemological traditions. As LLMs produce summaries, topics, thematic interpretations, and labels of textual data that look increasingly fluent, nuanced, and contextual, researchers may increasingly be tempted to use these outputs in analysis while decreasing the type of engagement with participants and their data that researchers say is central to qualitative research values.

6 Implications for Researchers and Designers

6.1 Considerations for researchers using LLMs in qualitative work

Thus far, we have explored how the combination of opaque tooling, lagging norms, and social context lead to an uncertain environment for qualitative researchers today. In this section, we answer RQ3 by providing an overview of key considerations and recommendations for LLM-curious qualitative researchers, or those who feel that LLM tools may interfere with their processes. Maintaining researcher agency and educating researchers on AI tools is vital as LLMs are embedded across more surfaces. We used prior work and our own expertise to reflect on the tensions raised by participants; we thus constructed Table 2 to distill some main choices researchers can make relating to the uses and potential implications of LLMs in qualitative contexts.

We now explain the three main covered areas of Table 2: *Policies, Research Values & Ethics*, and *Tool Selection*.

6.1.1 Policies. The first dimension in the table is the role of *policies*. We recommend actively keeping up to date on all policies from relevant communities, but treating these policies as an entry point rather than as a final answer on appropriate engagement with these tools. Policies may vary between venues; for example, some venues mandate specific types of disclosure [2] while others disallow the use of AI in some contexts [68].

As surfaced by our participants, existing policies and regulations may not engage deeply enough with the risks to certain participant populations. Therefore, while policies are important, they are likely insufficient considerations by themselves. The rest of the table provides a scaffold for considering the conscientious use of LLMs in qualitative work beyond established policies.

6.1.2 Research ethics and values. The second main dimension of the table encompasses *Consent, Privacy, Disclosure, Intentional task selection*, and *Validation*. Consent and privacy protection are fundamental values in qualitative research. Our participants indicated a current lack of scaffolding in current tools guiding how to use LLMs while preserving these values. Researchers should update their *consent* forms to accurately reflect intended uses of LLMs in research, as recommended by Kapania et al. [44]. These forms should include details on how participant privacy will be protected and education about potential risks.

Participants repeatedly highlighted a fear of compromising participant *privacy*. Researchers must take care not to share personally identifiable information or compromise the data of at-risk populations. These risks are particularly salient when using proprietary LLMs for any task involving confidential participant data. If using a proprietary model for such a task, carefully consider how to anonymize data, such as removing names or identifiable terms

Table 2: Table highlights the main considerations and recommendations regarding how qualitative researchers can incorporate LLMs into their research process. Considerations include a brief overview of the main tradeoffs involved in specific decisions, while Recommendations provide explicit guidelines for researchers that minimize risk to privacy, sensemaking, and research ethics.

Category	Considerations	Recommendations
Policies		
Institutional, venue, and government guidelines and regulations	Policies and regulations can help to provide basic guidance, but may be out of date or insufficient by themselves.	Familiarize yourself with and follow institutional, publication, and federal policies, but treat these as a starting point for ethical engagement.
Research Ethics & Values		
Consent	Participants may not know that their data could be shared with a third-party company, or be aware of potential implications, such as the implication of models training on their data.	Update consent forms on details if planning to use LLMs in the research, particularly in a training capacity, and educate participants on potential risks.
Privacy	Guidelines may not strictly disallow using participant data with proprietary LLMs, even ones that train on inputs, leading to identification risk in the future. Additionally, LLMs may be embedded in the tools you currently use.	Do not use proprietary models that train on input when interacting with participant data. Host local models and use open source models when possible. Anonymize confidential participant data if using a proprietary model. Confirm data policies before using a tool like MAXQDA that integrates LLM features.
Disclosure	Venues may mandate specific forms of disclosure, but researchers may wish to disclose additional uses of LLMs for research. Reflexively engage with how LLMs may have shaped the relationship with participants, ideas in the project, and perceptions of the data.	Proactively disclose LLM uses that significantly impacted a sense-making process or analysis outputs, including in ideation. When LLMs are used in analysis, discuss validation, including, disclosing model, model version, prompts, and performance, if appropriate.
Intentional task selection	Chat-based interactions make unintentionally transitioning from one task to another easy. Tasks carry heterogeneous benefits and risks for qualitative research, and some decisions about LLM involvement or automation depend on research values. Interpretivist qualitative research tasks may be particularly at odds with LLM use.	Before using an LLM, outline all tasks under consideration for LLM use. For each one, decide if the task is appropriate for LLM involvement or automation based on risks to compliance, originality, bias, privacy, and validity according to research values.
Validation	LLMs should always be validated in task-specific ways in research. A clear way to validate the performance of an LLM may or may not exist for every task, and some tasks may be inappropriate for LLMs to perform even if they can be validated. Formal validation may be more tractable for positivist research, and a plan for validation should be made prior to engagement. Interpretivist and other research stances should consider whether any tasks are appropriate, given challenges to validation.	Decide how to validate an LLM's performance on any chosen task. This could include side by side comparison to human performance in qualitative coding, for example, or manual review of any edits to text suggested by an LLM. Only use an LLM if performance can be appropriately validated for the task.
LLM bias	Consider LLMs' cultural and political bias as you decide which tasks are appropriate for an LLM. Consider whether participant populations are aligned with LLM training data, and what the impacts of misalignments might be.	For studies on populations unlikely to be well-represented in training data of off-the-shelf LLMs, avoid using off-the-shelf models in analysis to limit misinterpretation, misrepresentation, or hallucination, or consider fine-tuning.
Tool Selection		
LLM chat interfaces	Chat-based interactions are highly interactive and can support brainstorming in a research process, but interactivity can induce slippage across tasks and be difficult to validate for every use case. As such, some researchers may decide the risks of task-agnostic chat interfaces are not worthwhile.	If using a chat interface, clearly outline the tasks for which the chat-based LLM will be used, and make a plan for validation before approaching the task. Monitor regularly for slippage across tasks.
QDA Software	Software for QDA is intentionally designed for a qualitative coding use case, so it may provide validation (like measuring IRR across collaborators) or be designed with risks to sensemaking in mind. However, highly structured tools may miss opportunities for interactivity afforded by LLMs, so appropriateness depends on the intended task.	Using structured tools reduce some but not all risks. If using LLM integrations into existing or emerging software tools, investigate the platform's privacy policies and still consider the bias an LLM could introduce into your analysis.
Task-specific tools (e.g. Microsoft Excel, Grammarly)	LLMs are increasingly embedded across the tools we use every day, even for simple tasks. Tools like Grammarly with more prescribed functionality constrained to editing may limit interference with your sensemaking process compared to open-ended tools.	Reflect on your full set of research tools. Monitor for potential compromise of participant data if you use Grammarly or other tools to process or edit participant data.
Non chat-based LLM use	Researchers may wish to use or create a research infrastructure using an LLM, either directly or accessed through an API, for tasks like procedurally annotating data. In these cases, choices between proprietary models and open source models, or locally hosted and remotely accessed models become salient, and carry different implications for privacy, transparency, and quality.	If possible, use an open source model rather than a proprietary model to promote transparency and open science. If possible, use a locally hosted model for greater control over data. If using a proprietary model, anonymize data and do not use models or API services that train on input.

(e.g. [19]). Avoid LLM interfaces which use your participants' data as training data for the model. Check whether any tools with LLM-based features (e.g. Grammarly or Atlas.ti) train on inputs, and how they handle data privacy before using them. Privacy may be better protected by hosting open source models locally, but this is not currently feasible for all researchers.

Choosing which types of *disclosure* are appropriate regarding AI use is an opportunity for qualitative reflexivity. Qualitative researchers, as reflected in our interviews, already center reflexivity and positionality in their work, and can set precedent by explicitly engaging with how an LLM shaped a project. Researchers could consider the potential benefits of additional AI disclosure beyond those required by a venue, institution, or community. However, disclosing

the use of AI may risk stigmatization, especially in non-computing fields [50], which should be balanced with the benefit additional transparency and context disclosure about AI usage could bring. Disclosures should ideally detail which models were used, and any fine-tuning, optimization, or prompt engineering performed.

Our table also outlines the value of *intentional task selection* when using LLMs in qualitative research. Intentional deployment of AI is key to maintaining researcher agency [26, 43], but our participants highlight how general purpose tools now blur boundaries between tasks. With this challenge in mind, intentional task selection is even more pressing. The appropriateness of involving an LLM in a task can depend not only on venue policies, but also on research values, like an interpretivist or positivist stance. Before using an LLM for any task in your research, researchers should outline all tasks under consideration for LLM involvement and determine whether involving an LLM is appropriate based on their research values. In particular, researchers should consider how involving an LLM in the task may shape their perspective, ideas, and findings if used in ideation, analysis, or writing. Researchers should also consider whether they may be committing accidental acts of academic plagiarism by relying on LLM-generated outputs [12] across the research pipeline, and should take responsibility for checking for copied content. If any tasks are determined to be appropriate for LLM involvement, researchers should monitor whether the planned task matches ongoing use of the LLM.

Our table also centers *validation* as a key consideration when using LLMs for qualitative research. Whether and how LLM usage can be validated appropriately depends on the task under consideration and individual research values. For example, if using an LLM to edit text, an appropriate validation might be to personally review each LLM-generated edit. If using an LLM to scale data annotation beyond possibility of human review, it may become appropriate or even necessary to use more positivist measures of validation, like inter-rater reliability, to measure accuracy compared to a human baseline. Not all qualitative research tasks have a clear appropriate validation: in particular, interpretivist researchers may find LLMs cannot be validated for tasks of interest and should be avoided. We advise making plan for validation of LLM performance for any chosen task *before* starting to interact with the tool, since anchoring bias can cause one to rationalize findings and outputs after already interacting with the LLM.

Additionally, when considering LLM appropriateness for any task, we recommend considering *LLM training data, current performance, and biases*. For example, LLMs may be useful for editing English text written by non-native English speakers, which appears to be a primary emerging use case from our participants and in other research [50]. However, LLMs may not yet be advanced enough in other languages to be used for the same purpose. Model performance on tasks like transcription and analysis often directly reflects training data [45], so alignment between participant interests and a model's training data should be key considerations. Using a mainstream LLM to analyze themes in a study of non-Western, non-English speaking participants may be inappropriate, since mainstream LLMs likely have poorer representations of those groups and viewpoints and could cause misinterpretation or harm. If LLM capabilities and bias cannot be considered and confidently addressed for a potential task, the tool should not be used.

6.1.3 Tool Selection. In the table, we outline the four types of LLM-aided tools that came up in interviews: *LLM chat interfaces*, *QDA software*, *task-specific tools*, and *non-chat based LLM use*.

Participants largely reported experiences with ChatGPT, which has interactivity and flexibility researchers enjoy but does not scaffold appropriate use for research purposes. Conversely, structured software solutions like MAXQDA for data analysis or Grammarly for text editing have LLM-based features but are designed to support users in a specific task-based use. Researchers should consider if a dedicated tool exists for their use case that guards against some risks to the user that open-ended chat interfaces pose. Chat is a useful way to interact with LLMs, but researchers may instead wish to integrate LLMs into software pipelines to procedurally generate outputs like data annotations.

The way an LLM is hosted and accessed also has implications for privacy and transparency. Proprietary models from companies like OpenAI can be accessed through a web browser or API, meaning data shared through those access points is shared with the company. Models from AI companies are often also closed source, meaning the model's training data and weights are private, and understanding model behavior can be difficult. If using a proprietary model for research purposes, consider providing explicit justification as recommended by Palmer et al. [61]. Open source models with known model weights and transparent training data are preferable for transparency, data privacy, and control, but currently, hosting them locally can be difficult and time-consuming without resources and technical skills.

6.2 Design principles for qualitative research tools incorporating LLMs

We have identified the urgent need for transparent tools that center qualitative research use cases. Based on our findings from qualitative researchers interested in engaging with LLMs, tool builders could emphasize 1) interactivity, 2) data exploration, and 3) support for qualitative coding in future qualitative research tools. Participants consistently mentioned the useful nature of the interactivity LLMs afford through a chat-based interface: they could ideate, obtain feedback, and potentially make requests for data retrieval or formatting in natural language (e.g. [31, 40]). There is a gap in tools that facilitate these use cases in addition to other, more structured uses. Second, LLMs could help researchers see their data in a new way, ideally widening rather than homogenizing the nature of their insights. Lastly, qualitative researchers are still interested in using LLMs to conscientiously assist in the coding process. Below are principles for tool designers when considering how to create tools that responsibly address these opportunity areas.

6.2.1 Design for participant privacy. Proprietary LLMs have sparked concerns regarding participant privacy. Tools that incorporate LLMs should be explicit about if, when, and how they call external APIs, and whether LLMs used are proprietary or open source. They should also explain to users before data is uploaded how they handle privacy, how they expect users to deanonimize data before using tools, how they scrub or plan to delete data, and whether they use inputs for training models. Second, some researchers mention open source models as a potential solution to privacy concerns, but many are unable to implement solutions

to this problem themselves. As such, there is an opportunity area to support researchers in having greater control and transparency over data privacy. More tools that make it easy to locally host, steer, and fine-tune open source models for qualitative research functions could empower researchers.

6.2.2 Design for intentional use. The lines between structured tools built for a single task and open-ended chat-based tools are blurring. Chat-based LLMs allow flexibility and interactivity, but do not currently support intentional use for specific tasks in the research process, making slippage into unintentional uses of LLMs easy. Researchers want to maintain agency over if, when, and how they are assisted or influenced by AI tools in their research process. Tools could be designed to guide users to intentionally select appropriate uses of LLM-based tools, and also offer features that allow researchers agency to develop their own ideas before opting into LLM assistance in any task.

6.2.3 Design for transparency and validation. Tools should give researchers ways to transparently evaluate performance. Appropriate validation varies depending on research values, and on the specific research task. For brainstorming, designing for validation could involve designing features to understand or probe outcomes and suggestions. For positivist qualitative coding tasks, features designing for reproducibility of outcomes, interrater reliability, and error analysis from LLM annotation compared to human coding may be useful. Features that require researchers to review outputs may help accomplish this goal, depending on the task.

6.2.4 Design for researcher context. Researchers often wish to bring in their own context when performing qualitative analysis, including their past experiences, the texts that influence their work, and the theories in which they ground their work. Tools could embed unique researcher perspectives with models that are trained or fine-tuned according to particular data sets or perspectives of interest to the researcher.

6.2.5 Design for deep engagement with data. Qualitative researchers worry that LLM-based tools could distance them from data rather than deepen their relationship with it, but tools can be explicitly designed to foster ‘closeness’ with data [33]. Many qualitative researchers expressed interest in using LLMs to see their data through another lens, but found that insights and themes produced on their data are generic or repetitive. Tools can explore LLMs as a way to produce more unique ways of understanding data, drawing researchers’ attention to *less* predictable features of the data, as suggested by Jiang et al. [43], or helping researchers examine or challenge their existing theories with direct evidence from the data. These features may even deepen researchers’ understanding of the data.

6.2.6 Design to consider participant perspectives and interests. Researchers are concerned about the variable performance LLMs have across contexts, knowledge domains, cultures, and languages. Tools should work to facilitate better alignment between LLMs and diverse subject populations. When possible, tool builders could give the option to use models that better reflect a target, like models that specialize in a particular language or have specialized training data, and facilitate investigating where there might be errors or bias.

7 Limitations

While we recruited participants across disciplines, our participant group does not cover all perspectives on qualitative research and its values. Disciplines that are even more situated than those represented in our participant group, such as ethnography, may be at greater risk of diminishment by LLM involvement, and those perspectives may not be represented by the voices highlighted here. The fact that our participants were mostly PhD students also likely shapes the findings, since younger people have more positive views of AI [73], and senior researchers are more likely to have ethical concerns about AI research use [50]. While our participants represent a diverse group of people studying various populations, there was a heavy bias for American universities, which may miss key perspectives from non-Western institutions. Despite these limitations, our study highlights an important population wrestling with incorporating transformative technologies into their research processes for the first time.

8 Conclusion

LLMs are poised to continue proliferating and permeating research. While LLMs present potential opportunities for assistance in some research tasks, they also surface potential tensions with the values of qualitative research. In this climate, we document a sociotechnical breakdown: the functionality of *tooling* is insufficient to reliably preserve participant privacy, the *guidelines and norms* lag behind, and the *social context* increasingly incentivizes the broad use of AI across surfaces. Active, intentional consideration from researchers on issues like privacy and model bias in advance of task selection, coupled with thoughtful tool design, can help to maintain agency as researchers decide appropriate uses for LLMs. We synthesized current uses, tensions, and ethical challenges into a set of choices for the intentional incorporation of LLMs in qualitative research. We hope that this paper will empower qualitative researchers to leverage LLMs confidently, and even creatively, for their work if they choose.

Acknowledgments

We thank James Eschrich for their formative contributions to the scope, literature, and analysis in this piece. We also thank Mor Naaman, Emily Tseng, Louise Barkhuus, and students and staff at the MIT Center for Constructive Communication for their perspectives on this work. We thank the participants of the “LLMs as Research Tools” workshop at CHI 2024 for their contributions that informed the direction of this work. We acknowledge that we used ChatGPT and Elicit to find specific works of related literature, which we then verified from source documents.

References

- [1] Mark S Ackerman. 2000. The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human-Computer Interaction* 15, 2-3 (2000), 179–203.
- [2] ACM. 2024. *Frequently Asked Questions: ACM Policy on Authorship*. ACM. <https://www.acm.org/publications/policies/frequently-asked-questions>
- [3] Cimmieron F Alvarez, Kristina M Scharp, and Amanda M Friz. 2023. “I am a binary trans man and I love being pregnant”: Making meaning of pregnancy in seahorse dad narratives. *Journal of Social and Personal Relationships* 40, 9 (2023), 3028–3050.

- [4] Giulia Andrighetto and Eva Vriens. 2022. A research agenda for the study of social norm change. *Philosophical Transactions of the Royal Society A* 380, 2227 (2022), 20200411.
- [5] Marianne Aubin Le Quéré, Hope Schroeder, Casey Randazzo, Jie Gao, Ziv Epstein, Simon Tangi Perrault, David Mimmo, Louise Barkhuus, and Hanlin Li. 2024. LLMs as Research Tools: Applications and Evaluations in HCI Data Work. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.
- [6] Christopher A. Bail. 2024. Can Generative AI Improve Social Science? (Jan. 2024). <https://doi.org/10.31235/osf.io/rwtzs> Publisher: OSF.
- [7] Jeffrey Bardzell and Shaowen Bardzell. 2011. Pleasure Is Your Birthright: Digitally Enabled Designer Sex Toys as a Case of Third-Wave HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Vancouver BC Canada, 257–266. <https://doi.org/10.1145/1978942.1978979>
- [8] Shaowen Bardzell and Jeffrey Bardzell. 2011. Towards a Feminist HCI Methodology: Social Science, Feminism, and HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 675–684. <https://doi.org/10.1145/1978942.1979041>
- [9] Andrea Baronchelli. 2024. Shaping new norms for AI. *Philosophical Transactions of the Royal Society B* 379, 1897 (2024), 20230028.
- [10] Eric P. S. Baumer, David Mimmo, Shion Guha, Emily Quan, and Geri K. Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology* 68, 6 (2017), 1397–1410. <https://doi.org/10.1002/asi.23786> _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23786>.
- [11] Peter Berger and Thomas Luckmann. 1966. The reality of everyday life. *The social construction of reality: A treatise in the sociology of knowledge* (1966), 33–42.
- [12] Saumyamani Bhardwaz and Jitender Kumar. 2023. An extensive comparative analysis of chatbot technologies-ChatGPT, Google BARD and Microsoft Bing. In *2023 2nd international conference on applied artificial intelligence and computing (ICAAIC)*. IEEE, 673–679.
- [13] Nick G Blismas and Andrew RJ Dainty. 2003. Computer-aided qualitative data analysis: panacea or paradox? *Building research & information* 31, 6 (2003), 455–463.
- [14] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
- [15] Virginia Braun and Victoria Clarke. 2013. *Successful qualitative research: A practical guide for beginners*. Sage publications Ltd.
- [16] David Brown et al. 2002. Going digital and staying qualitative: Some alternative strategies for digitizing the qualitative research process. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, Vol. 3.
- [17] Grzegorz Bryda and António Pedro Costa. 2023. Qualitative research in digital era: innovations, methodologies and collaborations. *Social Sciences* 12, 10 (2023), 570.
- [18] Kathy Charmaz and Robert Thornberg. 2021. The Pursuit of Quality in Grounded Theory. *Qualitative Research in Psychology* 18, 3 (July 2021), 305–327. <https://doi.org/10.1080/14780887.2020.1780357>
- [19] Madiha Zahrah Choksi, Marianne Aubin Le Quéré, Travis Lloyd, Ruojia Tao, James Grimmelmann, and Mor Naaman. 2024. Under the (neighbor) hood: Hyper-local Surveillance on Nextdoor. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–22.
- [20] Andy Crabtree. 2024. H is for Human and How (Not) To Evaluate Qualitative Research in HCI. *arXiv preprint arXiv:2409.01302* (2024).
- [21] Norman K. Denzin, Yvonna S. Lincoln, Michael D. Giardina, and Gaile Sloan Cannella. 2024. *The SAGE Handbook of Qualitative Research*. SAGE.
- [22] Nicole M. Deterding and Mary C. Waters. 2021. Flexible Coding of In-depth Interviews: A Twenty-first-century Approach. *Sociological Methods & Research* 50, 2 (May 2021), 708–739. <https://doi.org/10.1177/0049124118799377>
- [23] Zackary Okun Dunivin. 2024. Scalable Qualitative Coding with LLMs: Chain-of-Thought Reasoning Matches Human Performance in Some Hermeneutic Tasks. *arXiv preprint arXiv:2401.15170* (2024).
- [24] Upol Ehsan, Koustuv Saha, Munmun De Choudhury, and Mark O Riedl. 2023. Charting the sociotechnical gap in explainable ai: A framework to address the gap in xai. *Proceedings of the ACM on human-computer interaction* 7, CSCW1 (2023), 1–32.
- [25] Steven Feldstein. 2024. Evaluating Europe’s push to enact AI regulations: how will this influence global norms? *Democratization* 31, 5 (2024), 1049–1066.
- [26] Jessica L. Feuston and Jed R. Brubaker. 2021. Putting Tools in Their Place: The Role of Time and Perspective in Human-AI Collaboration for Qualitative Analysis. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 469 (oct 2021), 25 pages. <https://doi.org/10.1145/3479856>
- [27] Guo Freeman and Donghee Yvette Wohn. 2020. Streaming Your Identity: Navigating the Presentation of Gender and Sexuality through Live Streaming. *Computer Supported Cooperative Work* 29, 6 (Dec. 2020), 795–825. <https://doi.org/10.1007/s10606-020-09386-w>
- [28] Carli Friedman, Aleksa Owen, and Laura VanPuymbrouck. 2024. Should ChatGPT help with my research? A caution against artificial intelligence in qualitative analysis. *Qualitative Research* (2024), 14687941241297375.
- [29] Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2023. CollabCoder: A GPT-Powered Workflow for Collaborative Qualitative Analysis. <https://doi.org/10.48550/arXiv.2304.07366> [cs].
- [30] Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2024. CollabCoder: a lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–29.
- [31] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* (2023).
- [32] Simret Araya Gebreegziabher, Zheng Zhang, Xiaohang Tang, Yihao Meng, Elena L Glassman, and Toby Jia-Jun Li. 2023. Patat: Human-ai collaborative qualitative coding with explainable interactive rule synthesis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [33] Linda S Gilbert. 2002. Going the distance: ‘closeness’ in qualitative data analysis software. *International journal of social research methodology* 5, 3 (2002), 215–228.
- [34] Linda S. Gilbert, Kristi Jackson, and Silvana di Gregorio. 2014. Tools for Analyzing Qualitative Data: The History and Relevance of Qualitative Data Analysis Software. In *Handbook of Research on Educational Communications and Technology*, J. Michael Spector, M. David Merrill, Jan Elen, and M. J. Bishop (Eds.). Springer, New York, NY, 221–236. https://doi.org/10.1007/978-1-4614-3185-5_18
- [35] Barney Glaser and Anselm Strauss. 2017. *Discovery of grounded theory: Strategies for qualitative research*. Routledge.
- [36] Igor Grossmann, Matthew Feinberg, Dawn C Parker, Nicholas A Christakis, Philip E Tetlock, and William A Cunningham. 2023. AI and the transformation of social science research. *Science* 380, 6650 (2023), 1108–1109.
- [37] Martyn Hammersley. 2013. *What Is Qualitative Research?* Continuum/Bloomsbury, London.
- [38] Martyn Hammersley and J.T.E. Richardson. 1996. The Relationship between Qualitative and Quantitative Research: Paradigm Loyalty versus Methodological Eclecticism. *Handbook of qualitative research methods for psychology and the social sciences* (Jan. 1996), 159–174.
- [39] Mohammad Hosseini and Serge PJM Horbach. 2023. Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Research integrity and peer review* 8, 1 (2023), 4.
- [40] Gautier Izcard and Édouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 874–880.
- [41] James Jaccard and Jacob Jacoby. 2020. *Theory construction and model-building skills: A practical guide for social scientists, 2nd ed.* The Guilford Press.
- [42] Mohammad S Jalali and Ali Akhavan. 2024. Integrating AI language models in qualitative research: Replicating interview data analysis with ChatGPT. *System Dynamics Review* (2024).
- [43] Jialun Aaron Jiang, Kandrea Wade, Casey Fiesler, and Jed R Brubaker. 2021. Supporting serendipity: Opportunities and challenges for Human-AI Collaboration in qualitative analysis. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.
- [44] Shivani Kapania, Ruiyi Wang, Toby Jia-Jun Li, Tianshi Li, and Hong Shen. 2024. “I’m categorizing LLM as a productivity tool”: Examining ethics of LLM use in HCI research practices. *arXiv preprint arXiv:2403.19876* (2024).
- [45] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences* 117, 14 (2020), 7684–7689.
- [46] Steinar Kvale. 1995. The Social Construction of Validity. *Qualitative Inquiry* 1, 1 (March 1995), 19–40. <https://doi.org/10.1177/107780049500100103>
- [47] Michelle S Lam, Janice Teoh, James A Landay, Jeffrey Heer, and Michael S Bernstein. 2024. Concept Induction: Analyzing Unstructured Text with High-Level Concepts Using LLoM. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–28.
- [48] Kevin Danis Li, Adrian M Fernandez, Rachel Schwartz, Natalie Rios, Marvin Nathaniel Carlisle, Gregory M Amend, Hiren V Patel, and Benjamin N Breyer. 2024. Comparing GPT-4 and Human Researchers in Health Care Data Analysis: Qualitative Description Study. *Journal of Medical Internet Research* 26 (2024), e56500.
- [49] Q Vera Liao and Ziang Xiao. 2023. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100* (2023).
- [50] Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X Zhang. 2024. LLMs as Research Tools: A Large Scale Survey of Researchers’ Usage and Perceptions. *arXiv preprint arXiv:2411.05025* (2024).
- [51] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW

- and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 72:1–72:23. <https://doi.org/10.1145/3359174>
- [52] Lisa Messeri and MJ Crockett. 2024. Artificial intelligence and illusions of understanding in scientific research. *Nature* 627, 8002 (2024), 49–58.
- [53] David L. Morgan. 2023. Exploring the Use of Artificial Intelligence for Qualitative Data Analysis: The Case of ChatGPT. *International Journal of Qualitative Methods* 22 (2023), 16094069231211248. <https://doi.org/10.1177/16094069231211248> arXiv:<https://doi.org/10.1177/16094069231211248>
- [54] Arvind Narayanan and Sayash Kapoor. 2024. *AI snake oil: What artificial intelligence can do, what it can't, and how to tell the difference*. Princeton University Press.
- [55] Dennis C Neale and John M Carroll. 1997. The role of metaphors in user interface design. In *Handbook of human-computer interaction*. Elsevier, 441–462.
- [56] Laura K Nelson. 2020. Computational grounded theory: A methodological framework. *Sociological Methods & Research* 49, 1 (2020), 3–42.
- [57] Haley Nolan-Cody and Kristina M Scharp. 2024. (Mal) adaptive sibling self and other communicative resilience in the context of parental substance use. *Journal of Communication* 74, 2 (2024), 145–159.
- [58] Étienne Ollion, Rubing Shen, Ana Macanovic, and Arnault Chatelain. 2024. The dangers of using proprietary LLMs for research. *Nature Machine Intelligence* 6, 1 (2024), 4–5.
- [59] OpenAI. 2024. ChatGPT. <https://chatgpt.com/>. Accessed: 2024-12-09.
- [60] Cassandra Overney, Belén Saldías, Dimitra Dimitrakopoulou, and Deb Roy. 2024. SenseMate: An Accessible and Beginner-Friendly Human-AI Platform for Qualitative Data Analysis. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 922–939.
- [61] Alexis Palmer, Noah A. Smith, and Arthur Spirling. 2024. Using proprietary language models in academic research requires explicit justification. *Nature Computational Science* 4, 1 (Jan. 2024), 2–3. <https://doi.org/10.1038/s43588-023-00585-1> Publisher: Nature Publishing Group.
- [62] M. Pentucci, G. Cioci, and C. Laici. 2024. Hybrid Analysis in Education: Comparing ChatGPT and Human Thematic Analysis of Teachers' Responses on Technology Use in Schools. In *ICERI2024 Proceedings* (Seville, Spain) (17th annual International Conference of Education, Research and Innovation). IATED, 3969–3977. <https://doi.org/10.21125/iceri.2024.1006>
- [63] Jon Porter. 2023. ChatGPT continues to be one of the fastest-growing services ever. *The Verge* (2023). <https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference#>
- [64] Casey Randazzo and Tawfiq Ammari. 2023. “If Someone Downvoted My Posts—That’d Be the End of the World”: Designing Safer Online Spaces for Trauma Survivors. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (CHI '23). Association for Computing Machinery, New York, NY, USA, 1–18. <https://doi.org/10.1145/3544548.3581453>
- [65] Casey Randazzo, Carol F. Scott, Rosanna Bellini, Tawfiq Ammari, Michael Ann Devito, Bryan Semaan, and Nazanin Andalibi. 2023. Trauma-Informed Design: A Collaborative Approach to Building Safer Online Spaces. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (Minneapolis, MN, USA) (CSCW '23 Companion). Association for Computing Machinery, New York, NY, USA, 470–475. <https://doi.org/10.1145/3584931.3611277>
- [66] Yolanda A. Rankin and Na-eun Han. 2019. Exploring the Plurality of Black Women's Gameplay Experiences. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3290605.3300369>
- [67] Gary Rolfe. 2006. Validity, Trustworthiness and Rigour: Quality and the Idea of Qualitative Research. *Journal of Advanced Nursing* 53, 3 (2006), 304–310. <https://doi.org/10.1111/j.1365-2648.2006.03727.x>
- [68] Sage. 2024. Artificial Intelligence Policy. <https://us.sagepub.com/en-us/nam/chatgpt-and-generative-ai>. Accessed: 2024-12-09.
- [69] Kristina M Scharp. 2021. Thematic co-occurrence analysis: Advancing a theory and qualitative method to illuminate ambivalent experiences. *Journal of Communication* 71, 4 (2021), 545–571.
- [70] 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.
- [71] Robert Soden, Austin Toombs, and Michaelanne Thomas. 2024. Evaluating interpretive research in HCI. *Interactions* 31, 1 (2024), 38–42.
- [72] Scott Spangler, Shannon Beasley, and Kembly Lingelbach. 2024. Analyzing artificial intelligence frameworks: Phenomenologists against machines. *Issues in Information Systems* 25, 4 (2024).
- [73] Jan-Philipp Stein, Tanja Messingschlager, Timo Gnams, Fabian Huttmacher, and Markus Appel. 2024. Attitudes towards AI: measurement and associations with personality. *Scientific Reports* 14, 1 (2024), 2909.
- [74] Sarah J Tracy. 2024. *Qualitative research methods: Collecting evidence, crafting analysis, communicating impact*. John Wiley & Sons.
- [75] Emily Tseng, Thomas Ristenpart, and Nicola Dell. 2025. Mitigating Trauma in Qualitative Research Infrastructure: Roles for Machine Assistance and Trauma-Informed Design. *Proceedings of the ACM on Human-Computer Interaction CSCW* (2025). To appear.
- [76] Jonas Wachinger, Kate Bärnighausen, Louis N Schäfer, Kerry Scott, and Shannon A McMahon. 2024. Prompts, Pearls, Imperfections: Comparing ChatGPT and a Human Researcher in Qualitative Data Analysis. *Qualitative Health Research* (2024), 10497323241244669.
- [77] Ryan Watkins. 2023. Guidance for researchers and peer-reviewers on the ethical use of Large Language Models (LLMs) in scientific research workflows. *AI and Ethics* (2023), 1–6.
- [78] Darin Weinberg. 2009. Social constructionism. *The new Blackwell companion to social theory* (2009), 281–299.
- [79] Lixiang Yan, Vanessa Echeverria, Gloria Milena Fernandez-Nieto, Yueqiao Jin, Zachari Swiecki, Linxuan Zhao, Dragan Gašević, and Roberto Martinez-Maldonado. 2024. Human-AI Collaboration in Thematic Analysis using ChatGPT: A User Study and Design Recommendations. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI EA '24). Association for Computing Machinery, New York, NY, USA, Article 191, 7 pages. <https://doi.org/10.1145/3613905.3650732>
- [80] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics* 50, 1 (2024), 237–291.

9 Appendix

9.1 Interview Ethics

We took several ethical concerns into consideration when conducting the interviews. All transcriptions were done through Zoom or Adobe Premiere Pro and edited manually by the research team to ensure fidelity to the speaker's words. Both Zoom and Adobe Premiere Pro encrypt user data and delete transcriptions. Direct participant quotes were never entered into ChatGPT or any other generative AI tool. Participants were permitted to use video or audio only for the interviews. Participants were compensated \$15 for each interview. During each interview, we obtained verbal consent to record and participants could stop the interview or refuse to answer any question. We informed participants that their anonymity would be preserved, not only by protecting their personally identifiable information, but also by obscuring or altering any highly specific research studies they mentioned.

9.2 Interview Protocol

Section 1: Situating the researcher

- Can you give me a high-level overview of your work?
- Can you describe your relationship with qualitative and human-centered research?
 - How would you define qualitative research?
 - How would you define human-centered research?
- What field are you in or where do you publish?
- What would you consider your relationship with Large Language Models to be?

Section 2: LLMs in the Research Process

These questions will each be asked for the steps *Prep and Brainstorming* and *Data Analysis*. For each stage, researchers will be asked "Have you ever used LLMs for this stage of research?" and if the answer is yes, they are asked the following:

- What motivated you to use an LLM for that purpose?
- How well did the LLM perform and how did you evaluate its performance?
- Did you have any ethical concerns with using LLMs in this stage of the research process?

- If yes, how did you try to mitigate those concerns?
- Would you use LLMs for this purpose again? How would/wouldn't you change your use of LLMs for this task going forward?

If the answer to the initial question was no, the following questions will be asked:

- Why do you think you haven't used LLMs for this stage of research?
- What are your main concerns with others using LLMs in this stage of research?

Section 3: Overall thoughts

Next, we transition to a higher level discussion of thoughts on LLMs in research.

- Overall, what do you think are the promising use cases for using LLMs in interview research?

- On the other side, if anything, what do you fear might be lost from incorporating LLMs into qualitative research? What are your main concerns?
- Generally speaking, what role do you perceive the LLM as occupying? Annotator? Co-researcher? Tool?

Section 4: Wrap up

Finally, we wrap up.

- Is there anything else that you'd like me to know or you think is interesting about your usage of LLMs for interview research?
- Do you have any questions for me?
- Is there anyone else in your circle who is a qualitative researcher that you might be willing to put me in touch with? If yes, I can send you a follow up email to contact.