Regular Submission

Generative Artificial Intelligence in Qualitative Data Analysis: Analyzing—Or Just Chatting?

Organizational Research Methods
1–37

DOI: 10.1177/10944281251377154
journals.sagepub.com/home/orm


Duc Cuong Nguyen[1] **and Catherine Welch**[2]

Abstract

Researchers, engineers, and entrepreneurs are enthusiastically exploring and promoting ways to apply generative artificial intelligence (GenAI) tools to qualitative data analysis. From promises of automated coding and thematic analysis to functioning as a virtual research assistant that supports researchers in diverse interpretive and analytical tasks, the potential applications of GenAI in qualitative research appear vast. In this paper, we take a step back and ask what sort of technological artifact is GenAI and evaluate whether it is appropriate for qualitative data analysis. We provide an accessible, technologically informed analysis of GenAI, specifically large language models (LLMs), and put to the test the claimed transformative potential of using GenAI in qualitative data analysis. Our evaluation illustrates significant shortcomings that, if the technology is adopted uncritically by management researchers, will introduce unacceptable epistemic risks. We explore these epistemic risks and emphasize that the essence of qualitative data analysis lies in the interpretation of meaning, an inherently human capability.

Keywords

qualitative data analysis, scientific tools, generative artificial intelligence, epistemic risks

Introduction

Since their release, generative artificial intelligence (GenAI) models—specifically large language models (LLMs)—have sparked vigorous debates about their potential to transform, even revolutionize, scientific research (Gatrell et al., 2024; Grimes et al., 2023; Koehler & Sauermann, 2024; Kulkarni et al., 2024; Wang et al., 2023). Central to these discussions is the capacity of LLMs to generate text that resembles human creativity and reasoning. Some voices in this debate go so far

[1]Alliance Manchester Business School, The University of Manchester, Manchester, UK
[2]Trinity Business School, Trinity College Dublin, College Green, Dublin 2, Ireland

Corresponding Author:
Duc Cuong Nguyen, Alliance Manchester Business School, The University of Manchester, Booth St W, Manchester M15 6PB, UK.
Email: duc.nguyen@manchester.ac.uk

as to claim that such models are capable of human-like reasoning (e.g., Bubeck et al., 2023; Kosinski, 2024), so they may well one day replace human researchers in the generation of scientific knowledge.

The proliferation of these claims about how such technologies might improve, augment or automate activities conducted by human researchers has led to the development of a wide range of applications that utilize and build on GenAI models. Qualitative data analysis is seen as fertile ground for such models, due to their natural language processing capabilities. The use of these models is often pitched as a way of 'scaling up' qualitative research designs (Gamieldien et al., 2023; Karjus, 2025), of making resource savings (Chubb, 2023; Xiao et al., 2023) and optimizing qualitative data analysis 'workflows' (Gao et al., 2023; Törnberg, 2024; Zhang et al., 2023). GenAI models are even positioned as autonomous digital assistants that can perform high-order analytical tasks such as identifying patterns and interpreting meaning (e.g., Koehler & Sauermann, 2024; Lixandru, 2024; Xiao et al., 2023).

Given these efforts underway to incorporate GenAI into qualitative data analysis, our aim in this paper is to investigate the following research questions: (1) what sort of technological artifact is GenAI? (2) Is it appropriate for use in qualitative data analysis? We offer a technically informed yet accessible evaluation of GenAI models and evaluate their claimed potential to contribute to qualitative data analysis. Our motivation in this paper is to go beyond prior work (e.g., Morgan, 2023), which does not inquire into the nature of the technology itself. In doing so, we provide a more informed assessment of the potential applications of such models to the diverse interpretive and reasoning processes that constitute qualitative data analysis.

This paper is structured as follows. We first turn to existing literature on scientific tools and instruments to conceptualize the role they play in knowledge production. This literature alerts us to the consequential nature of our tools; the epistemic risks they pose, in the form of introducing errors into knowledge production; and the validation process that a scholarly community should undertake before warranting their use. We then present the results of our evaluation of GenAI. We commence by addressing the first research question: what sort of technology is GenAI? To address this question, we review the computer science literature on generative artificial intelligence and transformer-based autoregressive large language models such as ChatGPT and DeepSeek. Based on the current state-of-the-art architecture, these GenAI models can be understood as a type of chatbot (Dam et al., 2024; Narayanan & Kapoor, 2024). As such, their output consists of word strings, probabilistically selected from their database of rules on word associations, that have the appearance of human-produced text (Narayanan & Kapoor, 2024).

Armed with this understanding of the nature of GenAI, we then outline the claims that are rapidly disseminating about its suitability as a tool for qualitative data analysis. We test these claims against three kinds of evidence: (1) our own analysis of a qualitative dataset; (2) the results from other published studies that explored the use of GenAI as a scientific tool for qualitative data analysis; (3) the GenAI capabilities that are now being offered by providers of qualitative data analysis software. This evidence allows us to address the second research question: whether GenAI is a suitable tool for qualitative data analysis. The conclusion we present is that GenAI—based on the current transformer architecture on which all foundational LLMs are built—is unsuited to qualitative data analysis and presents multiple epistemic risks to management research were its adoption sanctioned. We argue that these risks include, but go beyond, the production of false and untrustworthy results. In addition, the very nature of GenAI technology obfuscates our attempts to evaluate the validity of its output.

Before proceeding, it is worth clarifying our understanding of qualitative data analysis. By qualitative data analysis, we are referring to the full range of 'lower-order' tasks of exploring and organizing data (e.g., search and retrieval and thematic coding), as well as the 'higher-order' tasks of finding unexpected (even counterintuitive) patterns and connections, engaging in

theoretical abstractions and modeling, applying multiple forms of reasoning, and proposing novel, critical and disruptive problematizations and reconceptualizations (see Flick, 2014 on the scope of analysis; see also Davidson & di Gregorio, 2011; Silver & Lewins, 2014). Above all, qualitative data analysis concerns the interpretation of social meaning. This is necessarily an inter-subjective process that depends on the researcher's own experience, which is derived from deep field engagement and participation in multiple social communities, including the scholarly community. Our view on data analysis can be characterized as interpretivist: there is no objective standpoint that researchers can adopt when investigating the social world. This leads us to assert the importance of the scientific community in underpinning the credibility of research. The 'organized skepticism' (Merton, 1973, p. 277) of the scholarly community is essential, given the fallibility and fragility of the conclusions researchers draw. As we shall now discuss, this role of the scholarly community is critical when it comes to the technologies we adopt for our research.

Research Instruments and Qualitative Research

Historians, philosophers and sociologists of science[1] have pointed to the prominent, yet often overlooked, role played by technological artifacts in the generation of knowledge. By technological artifacts, we refer specifically to equipment, both physical and digital (e.g., computer databases, see Hine, 2006, and computer simulations, Alvarado, 2023), that has been made possible by scientific advances (cf., Shapin & Schaffer, 1985). Such knowledge-producing technologies are not just passive objects which we employ to enable scientific activities. More than that, they enable, distort and limit our observations of the world (Croissant, 2022), and in doing so, are constitutive of the knowledge they produce (Clarke & Fujimura, 1992)—including our understanding of what good scholarship looks like.

Technology has also impinged on qualitative research (Brinkmann et al., 2014), despite the mantra that in such studies, the researcher is the (exclusive) research instrument (e.g., Tracy, 2012). Technologies have been applied to qualitative data sources, qualitative data collection, data analysis and verification. For example, the introduction of the tape recorder, today largely taken for granted as a tool of the trade, altered the qualitative research process. The changes wrought were not confined to fieldwork interactions but extended to data analysis. The details of speech acts that a recorder can capture made new forms of analysis possible, such as conversation analysis and (forms of) discourse analysis (Jones, 2021). At the same time, recording changed how analysis of interviews was conducted. Given that researchers today typically analyze transcripts (Lee, 2004), not fieldnotes or sound recordings, they place more emphasis on data in the form of verbatim quotes from interviews, while other aspects of the interview—notably the dynamics of its production and the 'social world' from which it was sourced—are more easily overlooked. A greater demarcation between 'text' and 'context' was the result (Jones, 2021).

Given this consequential nature of scientific instruments, a crucial question is how, why and for what purpose have particular tools been adopted (or not) by a research community? As Clarke and Fujimura (1992) put it, how does a scholarly community determine that a tool is the right one for the job? Ultimately, the answer to this question is up to us: an artifact needs to be socially accepted as the appropriate tool for a specific purpose. Convincing a scholarly community to adopt a new technology as a legitimate scientific instrument should be a rigorous and protracted undertaking (Alvarado, 2023; Bechky & Davis, 2025) driven by scholarly norms (e.g., Merton, 1973).

In this process, the first step is establishing criteria for rightness, i.e., appropriate use. Evaluative criteria have been much debated in qualitative research, given the diversity of

onto-epistemological positions that researchers may adopt. Despite this, we argue that there are concerns that matter to all qualitative researchers, no matter their paradigmatic positioning. Although interpretivist researchers reject traditional positivist criteria of validity and reliability (e.g., Yin, 2014), they too care deeply about their epistemic responsibility to report participants' accounts faithfully, avoiding misrepresentation and knowingly introducing errors; reflexively question their own assumptions; report their research processes transparently (i.e., providing an audit trail, to use the term popularized by Lincoln & Guba, 1985); analyze data systematically; demonstrate deep engagement with the research setting; and reach conclusions that are supported by evidence and are logically sound (Yanow & Schwartz-Shea, 2014). All qualitative researchers need to exercise doubt and be constantly alert to how their analysis might be wrong—including taking responsibility for the tools they use.

A non-human and non-interpreting instrument must not detract from the interpretive, reasoning and conceptual activities of the human analyst who makes use of it. Does the instrument assist qualitative researchers in producing credible interpretations of their data? A tool needs to demonstrate to a scholarly community more than the improvements in efficiency and convenience that might be sufficient for adoption in other settings. As well, its processing of qualitative data needs to produce results that are trustworthy in terms of factual accuracy (i.e., not introducing errors, distortions and biases), reliability (i.e., stability, dependability), transparency (i.e., verifiability, explainability) and compatibility with ethical standards (for a discussion, see e.g., Baird & Faust, 1990).

Having established criteria for evaluating the fitness of a tool, the next step is to undertake an independent and evidence-based validation process. Without this, the results the tool produces cannot be accorded the status of warrantable knowledge. Judgement calls need to be made: for example, how high an error rate or what degree of transparency is acceptable? The benefits of the tool need to be weighed against its epistemic risks—in other words, the risk of being wrong in our conclusions (Biddle & Kukla, 2017). The scientific process is fraught with potential risks—of misconceptions, biases, poor reasoning, and observational errors—and tools may exacerbate these risks. In qualitative research, the stakes are even higher, given that the data are imbued with social meaning that is difficult to access and can easily be misunderstood (Maxwell, 1992).

As scholars, we are exposed to these same risks of error and false beliefs during the process of validating a tool—ultimately, our attitudes to the tool may lead us to overlook disconfirming evidence. A further complication is that the validation process is shaped by the power structures, interests and intellectual paradigms which dominate a research community (Casper & Clarke, 1998; Mikami, 2015). Dissemination and widespread adoption of the tool typically also involves influences from regulators, funding agencies and market actors such as instrument manufacturers, publishers and, more recently, software houses (Alvarado, 2023).

To conclude this section, we use this framing of the role of instruments in research as the conceptual orientation for our study, both in terms of its rationale and direction. It allows us to understand the consequential (but often overlooked) nature of technological artifacts—they affect the nature of the data we use, the analysis we produce, and our standards of what good research is. For this reason, they can be regarded as 'the third element of scientific inquiry' alongside theory and methodology (Alvarado, 2023: 101). Given their importance, the legitimization of artifacts for use as scientific tools requires a careful process of validation by a scholarly community. The initial introduction of a technological artifact is the critical point in time to conduct this debate, before the use of the artifact becomes normalized and taken for granted. In the case of GenAI and its proposed adoption by qualitative researchers, that time is now. Accordingly, we proceed to evaluate GenAI. We start with the nature of the artifact itself, before discussing its application to qualitative data analysis.

Generative Artificial Intelligence: What Is It?

Recent advancements in natural language processing (NLP) techniques and computational power have significantly expanded the capabilities and applications of GenAI models. These systems are now widely integrated across various professional domains, facilitating the automation of tasks such as content generation, software development, and customer support (Gozalo-Brizuela & Garrido-Mechán, 2023). Unlike other AI systems, GenAI models encode statistical patterns from data to produce novel content based on probability. Central to this progress is the emergence of autoregressive large language models (AR-LLMs), a class of GenAI models designed to generate synthetic word strings resembling human-like text (Henighan et al., 2020). The term 'autoregressive' refers to the model's statistical approach of linear prediction: generating synthetic text one token (i.e., word) at a time, using each newly generated token as additional context for predicting the next. This method enables AR-LLMs to produce seemingly coherent and contextually relevant text (Wei et al., 2022a).

Building on this foundational architecture (see Vaswani et al., 2017), developers and researchers at OpenAI, Google, Meta, and Anthropic, among many others, have introduced interactive AI systems designed to make advanced AR-LLMs more accessible to a broader audience. Commonly referred to as 'conversational agents' or 'chatbots', GenAI tools such as ChatGPT, Claude, and DeepSeek leverage AR-LLMs to generate synthetic human-like responses in real time to facilitate naturalistic conversations with users (Bommasani et al., 2022; Brown et al., 2020).

What distinguishes AR-LLM-based chatbots (henceforth, shortened to LLMs) from earlier language models is their training on vast datasets using what are termed 'deep learning' techniques, particularly the transformer architecture (Liu et al., 2024; Radford et al., 2018; Touvron et al., 2023). During training, transformer models encode large amounts of textual data[2] that are usually scraped from the internet (McCoy et al., 2024), allowing them to create a model (a set of rules) based on patterns, structures, and relationships within the natural language. The model then represents the words and phrases as vectors in a 'high-dimensional space'—dataset arrays that contain large amounts of features and attributes from various sources—capturing semantic similarities and correlations (Vaswani et al., 2017). When generating synthetic text, the model uses these 'learned' representations to predict the most probable next word while considering the entire context of the preceding text. This process is iterative, involving the calculation of probabilities based on the patterns it has acquired, allowing the model to generate novel synthetic texts. Moreover, as a product of their training on large datasets and the transformer architecture, these models can be applied to a range of NLP tasks such as conversations, text summarization, editing, and translation (Brown et al., 2020; Wei et al., 2022a).

Once trained to predict the next token in a sequence, LLM chatbots are periodically updated or fine-tuned with additional data. Fine-tuning involves a process called reinforcement learning from human feedback (RLHF). In RLHF, human evaluators assess the model's outputs and provide feedback (human-annotated/labeled data), which is used to build a separate 'reward' model. This reward model then guides LLM chatbots toward generating outputs that better align with human preferences, ethical standards, and contextual appropriateness. LLM chatbots are then optimized through reinforcement learning, whereby their rules are strengthened ('rewarded') or weakened ('penalized') based on the alignment of their synthetic outputs with desired outcomes as determined by their developers (Achiam et al., 2023). However, it should be noted that alignment is an ongoing challenge (Qi et al., 2023), and RLHF to improve model outputs can introduce more biases (Mu et al., 2024).

When a user interacts with an LLM chatbot—e.g., uploads documents, inputs texts, or writes a natural language instruction (prompt)—several automated pre-processing steps occur. The input is first tokenized: texts are broken down into smaller units (tokens), which are then transformed into

numerical vectors (embeddings) that capture relationships between words in high-dimensional space. These embeddings are then processed through multiple layers of the transformer architecture (Vaswani et al., 2017). The LLM chatbot then predicts the next token by sampling from a probability distribution conditioned on the input text (or data) and decodes the tokens back into human-readable text (Radford et al., 2018).

Researchers evaluating the potential of LLM chatbots in science are increasingly identifying a range of shortcomings that limit their scientific applicability (Bommasani et al., 2022; Messeri & Crockett, 2024). In discussing these challenges, we revisit the criteria for validating a scientific tool outlined in the previous section: factual accuracy, reliability, transparency, and ethical responsibility. We now discuss each criterion in turn.

Factual accuracy is a critical concern when evaluating scientific tools. The assessment often made about LLM chatbots is that they are prone to 'hallucinations', that is, produce content that is nonsensical or untruthful (Maynez et al., 2020; Phan et al., 2025). The term implies that these models produce accurate responses, with occasional errors. However, this characterization is misleading. LLM chatbots are designed to produce word strings that approximate word associations and patterns that humans have provided in an existing corpus of text. They are not designed to evaluate the truthfulness of the answer they produce (Walsh, 2023); their output 'does not have to *be* correct, it must *appear* correct' (Thornton, 2023, p. 27). As LLM chatbots are 'indifferent to the truth' of their outputs (Hicks et al., 2024; Narayanan & Kapoor, 2024), correspondence to facts is a matter of chance (Kalai et al., 2025).

Compounding this lack of truthfulness are the multiple biases inherent to each stage of the process of synthetic text generation. These include data-derived biases, ingrained in the language corpus used to train LLM chatbots (Bender et al., 2022); developer/evaluator biases, given that humans are involved throughout the training process (Mu et al., 2024); and algorithmic biases: systematic and recurring errors that occur when a model privileges certain representations or social categories over others (Arrieta et al., 2020; Eloundou et al., 2024; Vassel et al., 2024). These biases are accompanied by knowledge cut-offs—that is, the date at which LLM chatbots no longer have up to date information (Cheng et al., 2024)—and a lack of temporal awareness (Dhingra et al., 2022).

In relation to the second criterion, reliability remains a critical yet unresolved issue as GenAI models cannot generate stable and reproducible results[3] (Achiam et al., 2023; Bommasani et al., 2022). This limitation stems from their probabilistic nature, which makes their output unpredictable and beyond user control. This means that even when using the same prompt, document or input text, varying results will be generated across different interactions (Bommasani et al., 2022; Brown et al., 2020). Moreover, as there is a context window—a textual range that GenAI models can process at any given time, including both input text and output response—extended conversations are highly prone to generating incoherent results as earlier segments of the conversation fall outside the range in which tokens are processed (Achiam et al., 2023).

In domains where reliability and accuracy are paramount, this inherent variability undermines any practical utility (Kaddour et al., 2023; McCoy et al., 2024). Accordingly, the developers of ChatGPT caution that '[g]reat care should be taken when using language model outputs, particularly in high-stakes contexts, with exact protocol (such as human review, grounding with additional context, or *avoiding high-stakes uses altogether* [our italics]) matching the needs of specific applications' (Achiam et al., 2023, p. 10). This admission from OpenAI should cast doubt on the suitability of LLM chatbots for use as scientific instruments: they cannot, and are not built to, produce trustworthy results.

Moving to the third criterion, GenAI tools are inherently non-transparent, owing to both algorithmic complexity and developer opacity (Alvarado, 2023; Burrell, 2016; Marcus, 2024). Artificial neural networks, which underpin these models (Achiam et al., 2023; Liu et al., 2024;

Touvron et al., 2023), do not follow explicitly programmed rules. Instead, they 'learn' autonomously through a self-supervised process (Henighan et al., 2020; Radford et al., 2018; Touvron et al., 2023), making their internal workings complex and the output they produce difficult, if not impossible, to explain (Arrieta et al., 2020; Bommasani et al., 2022; Burrell, 2016; Mitchell & Krakauer, 2023). This algorithmic complexity is further exacerbated by developers' intentional non-disclosure of key details—including model architecture, training data and fine-tuning processes—which obstructs external evaluations and scrutiny of the technology[4] (Mitchell & Krakauer, 2023).

Last but not least, the ethics of GenAI remain an ongoing concern (Huang et al., 2025). Among a host of ethical issues, perhaps the most salient for researchers are data leakage, which refers to the risk of user inputs being incorporated into a model's training data to improve model performance, raising privacy, intellectual property, and consent concerns (Lukas et al., 2023); the immense energy consumption to train and operate GenAI models, raising concerns about environmental sustainability (Luccioni et al., 2024); and accountability, arising from questions as to who should be held responsible for the generation and utilization of GenAI outputs, especially in high-stake applications (Achiam et al., 2023; Raza et al., 2025).

In response to the growing ethical and practical concerns, researchers and developers have introduced technological innovations aimed at enhancing the factual accuracy and contextual relevance of GenAI outputs. A notable advancement is the development of vector-based retrieval-augmented generation (RAG) architectures (Ni et al., 2025; Zhao et al., 2024), which integrate an external knowledge base to improve the likelihood of a model generating more accurate and contextually relevant content (Lewis et al., 2020). Beyond architectural advancements, developers have also turned to external tools, via application programming interfaces (APIs), to mitigate some of the inherent technological limitations of AR-LLMs. For example, ChatGPT can access a dedicated Python execution environment to perform arithmetic calculations and other tasks, such as counting characters or words.

However, RAG architectures and APIs primarily function to increase the probability of generating a correct next-word answer, and they do not and cannot ensure factual correctness, accuracy, reproducibility, or ethical responsibility (Lin et al., 2021). Moreover, while continuous fine-tuning and data updates can enhance model performance, they introduce risks such as model drift (Shumailov et al., 2024) and a reduction in response diversity (Peterson, 2025; Zhang et al., 2023). Consequently, model reliability and accuracy in any given task can fluctuate or degrade over time[5] (Chen et al., 2024; McCoy et al., 2024). These limitations derive from the fundamental constraints of the transformer architecture and autoregressive method on which all current GenAI tools are built, and cannot be mitigated through technical improvements, prompt engineering, or scaling efforts; that is, increasing the size and training data of a model (Dziri et al., 2023; LeCun, 2022; Lin et al., 2021; Marcus, 2018; Sahoo et al., 2024).

Evaluation of the effectiveness of LLM chatbots is hindered not only by algorithmic complexity and inherent opacity but also by the way scholars perceive them. Despite their technical limitations, GenAI models are often misunderstood, leading to an over-estimation of their capabilities (Morris, 2023). This disconnect is not accidental but is, in part, an outcome of the power of commercial interests that are actively pushing the adoption of the technology (Widder et al., 2023). AI hype is a feature of the field since its inception—and the coining of the term 'artificial intelligence' itself (e.g., Placani, 2024), which conflates literal with metaphorical intelligence. Words associated with human intelligence infuse the AI discipline—for example, as we have discussed in this section, the technical terms of 'learning' and 'rewards.' This linguistic framing is highly misleading as it creates a false equivalence between humans and algorithms (Bender, 2024; Van Rooij et al., 2024).

Another obstacle is the technological artifact itself. As we have outlined, LLM chatbots are explicitly designed to generate synthetic text that imitates human-produced text to the greatest

extent possible. The aim is precisely to achieve text that looks plausibly human, and to mimic human utterances in a convincing way. The results of this 'faking' of human intelligence (Walsh, 2023) are that users are often misled, according intelligence to LLM chatbots (Bender, 2024; cf. Bubeck et al., 2023)—an anthropomorphizing response encouraged by the dialogical, interactive nature of the technology.[6] For these reasons, we would regard the epistemic risks of LLM chatbots as qualitatively distinct from other technological contenders for scientific use. We now examine these risks in relation to qualitative data analysis.

LLM Chatbots as a Tool for Qualitative Data Analysis? The Claims

The introduction of GenAI has been hailed as revolutionary in the development of qualitative research (Zhang et al., 2023). Its claimed applications span multiple stages of the research process, from designing interview protocols (Goyanes et al., 2025) and conducting interviews (Chopra & Haaland 2023) to performing inter-rater reliability checks[7] (Hou et al., 2024). GenAI has even been suggested as a replacement for human research participants (Dillion et al., 2023), although this use is also (deservedly) contested (Harding et al., 2024; Wang et al., 2025). In data analysis, many anticipate that GenAI tools will not only augment and automate existing analytical techniques (Lee et al., 2024; Morgan, 2023) but could also 'act' as a virtual assistant to aid in interpretive endeavours (Hitch, 2024; Lieder & Schäffer, 2024; Wheeler, 2025). Table 1 highlights illustrative claims that have been made about the use of GenAI tools in qualitative data analysis, ranging from automation to augmentation and anthropomorphism; that is, treating GenAI as a virtual research assistant.

Crucially, these claims are not only made by early adopters (many of whom are computer scientists rather than experienced qualitative researchers), but also by major computer-aided data analysis software (CADQAS) providers (e.g., ATLAS.ti, MaxQDA and NVivo), which have introduced GenAI capabilities. These incumbent firms are joined by recent start-ups (e.g., Coloop, AILYZE; QInsights) that have integrated or built on GenAI tools for qualitative data analysis. Overwhelmingly, all these 'Qual-AI'[8] software providers are promoting their solutions as providing efficiency gains, such as 'qualitative insights in minutes instead of weeks' (see Figure 1) or the ability to make qualitative projects '30x faster' (see Figure 2). It should be noted that not all CAQDAS providers have incorporated GenAI features. A notable exception is Quirkos, which has taken a more cautious approach, emphasizing the importance of transparency, researcher reflexivity, and the risk that GenAI may obscure researcher decision-making and introduce biases that may go unchecked (Quirkos, 2025).

The claims that position GenAI models as valuable tools for qualitative data analysis rest on their supposed capability to automate the labour-intensive task of 'coding' to identify themes and patterns. In expediting this process, users are promised efficiency gains (Anis & French, 2023; Huang et al., 2024) and cost savings that are associated with hiring human research assistants (Chubb, 2023; Nguyen-Trung, 2025). It has even been claimed that they can improve trustworthiness and replicability by minimizing human subjectivity and involvement in the analytical process (Abdüsselam, 2023; Theelen et al., 2024)—a fundamental misunderstanding of the nature of qualitative data analysis and its interpretive and iterative process of meaning-making. We will now turn to how we went about evaluating the claims being made about GenAI.

Methodology

To consider how the adoption of 'Qual-AI' tools might impact qualitative data analysis, we evaluated the main analytical steps of a GenAI-enabled project. When we commenced this evaluative exercise in late 2023, we (like most management researchers) were new to GenAI and were

Table 1. Illustrative Claims About AR-LLMs in Qualitative Data Analysis.

Use Cases	Claims	Misconceptions
Automation (e.g., Autocoder)	'LLMs outperform both state-of-the-art supervised models and human coders, offering higher accuracy across language and country contexts' (Törnberg, 2024, p. 11) 'ChatGPT outperforms [human coders] for several annotation tasks, including relevance, stance, topics, and frame detection' (Gilardi et al., 2023, p. 1). 'The results demonstrate the great potential of ChatGPT as a data annotation tool (i.e., autocoder) using simple prompt design' (Huang et al., 2024, p. 297) 'AI can be trained to recognize and categorize common themes or concepts within qualitative data, thus eliminating manual coding time requirements while speed up and streamlining analysis processes' (Abdüsselam, 2023, p. 3). 'automating aspects of qualitative data analysis through LLMs can enhance efficiency and scalability, thus allowing researchers to focus on the parts of the research that require deeper insights that are only currently possible with human cognition' (Gamieldien et al., 2023, p. 14)	Assumes that the results have equivalent epistemic status to that of a human-driven process, or sufficiently so to be useful.
Augmentation (Human-AI interaction)	'[ChatGPT] had features such as revealing the meaning in qualitative data analysis, understanding and interpreting the essence of the data. [ChatGPT] was able to analyze the interview text, extract code, category and themes, and include direct quotes from the text' (Sen et al., 2023, p. 14) ChatGPT's 'powerful language model and its ability to summarize, interpret, and solve various research problems make this artificial intelligence model a potential candidate to assist researchers in many phases of empirical investigation' (Goyanes et al., 2025, p. 17). 'LLMs as tools to augment, not replace, human analysis in qualitative research […] in streamlining the workflow, reducing processing costs, and ensuring a more transparent and credible analysis process' (Zhang et al., 2023, p. 14) 'LLM-based algorithms for NLP enhances [sic] the theorization process, allowing us to derive new insights from sentiments, topics, and phrases and to refine our analysis' (Garcia Quevedo et al., 2025, p. 15)	Assumes that AR-LLMs have the capability to understand human experiences to enable or enhance interpretations.

(continued)

Table 1. Continued.

Use Cases	Claims	Misconceptions
Anthropomorphism (e.g., research assistant)	ChatGPT is a 'trustworthy partner in the information interpretation process. The synthesis capability of ChatGPT becomes a cornerstone, providing researchers and analysts with the ability to address complex texts with increased efficiency' (Lixandru, 2024, p. 65). 'While not as good as a human research assistant, ChatPDF was a supportive friend who helped move work ahead…' (Chubb, 2023, p. 13). 'AI as a partner and a collaborator in the research process …' (Walsh & Pallas-Brink, 2023, p. 548). 'ChatGPT as an additional member of the analysis team, contributing to researcher triangulation by adding to knowledge building and sensemaking…' (Lee et al., 2024, p. 9) 'AI can evolve into a partner in qualitative research, challenging traditional notions of intelligence and interpretation and paving the way for more insightful research' (Lieder & Schäffer, 2024, p. 1) 'I now think of ChatGPT as a virtual colleague, … who nevertheless can make a useful contribution to less sophisticated aspects of analysis while they are learning and developing their skills' (Hitch, 2024, p. 604)	Assumes that AR-LLMs possess understanding, have the capability to reason and to learn from experience.

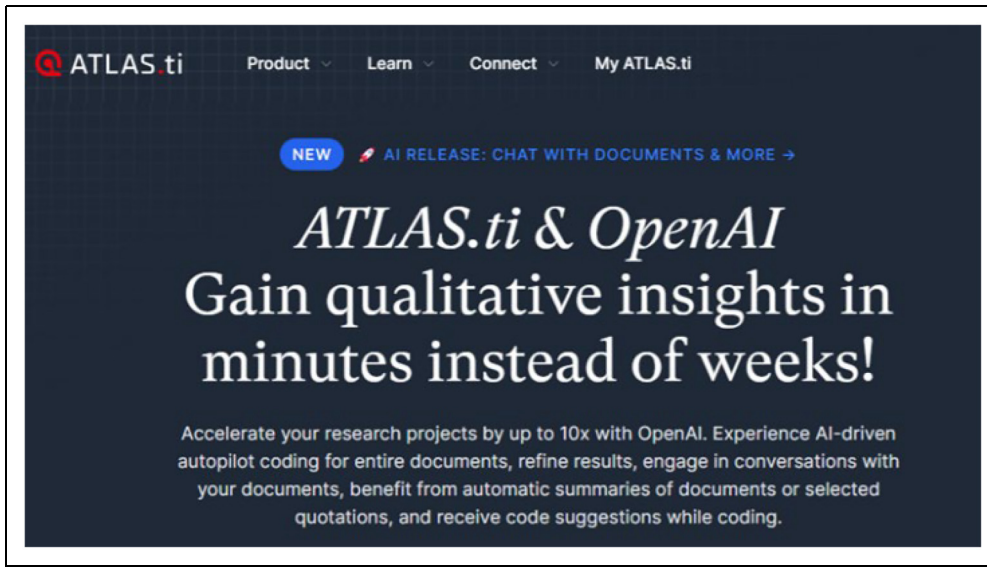


Figure 1. ATLAS.ti Marketing Material.

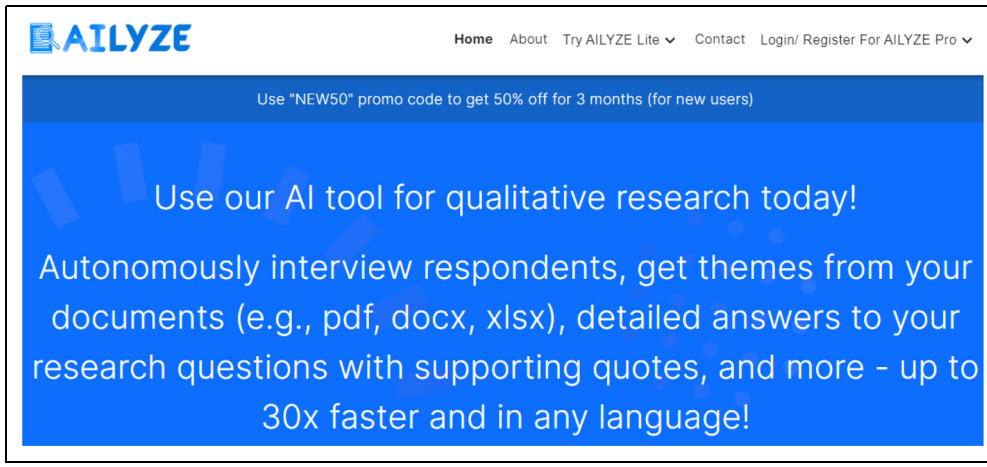


Figure 2. AILYZE Marketing Material.

agnostic as to its possible benefits. It is worth pointing out that, unlike many proponents of the tools (notably Qual-AI providers), we have no financial interest in encouraging its adoption.

To conduct our evaluation, we analyzed a collection of publicly available data on portable computing technologies from the late 1980s to 1990s (part of a larger research project currently being conducted by one of the authors). The data comprised transcripts of interviews (oral histories), blog posts and news and magazine articles from 1989 to 2016 (Table 2). The advantage of choosing this dataset is that (1) it has been assembled by one of the authors, so we have sufficient familiarity with it to judge the quality of the LLM chatbot outputs; (2) it does not involve uploading any confidential or proprietary data[9]; and (3) it includes a diverse range of sources that are commonly used in qualitative research projects (i.e., interviews and secondary data).

Table 2. Empirical Material for Our Evaluation.

Data Type	Source and Description	Word Count/Tokens
Magazine	*MIT Technology Review*: An article discussing the innovative developments in laptop computer designs.	414 (552)
	Government Computer News: An article discussing the advantages and disadvantages of GRiDPad in professional work settings.	666 (888)
	Pen Computing Magazine: Part 1: An article highlighting Jeff Hawkins's early life and inventive upbringing, tracing the origins of his success in reviving the handheld computer.	1,729 (2306)
	Pen Computing Magazine: Part 2: An article highlighting Jeff Hawkins's early life and inventive upbringing, tracing the origins of his success in reviving the handheld computer.	2,261 (3015)
Blog post	*Tedium*: A reflection on RadioShack's decline due to a shift away from its original business model for tech enthusiasts, featuring the GRiDPad.	577 (770)
	QuinStreet Enterprise: an opinion piece on how GRiDPad pushes the limits of the portable computing market.	955 (1274)
	Centre for computing history: A post on the design and features of GRiDPad.	252 (336)
	Old Computers Museum: A post on the features, functions, and innovative technology of GRiDPad.	393 (524)
	OSNews: A post that traces the origins and influence of Jeff Hawkins on the computing industry, detailing key innovations, product failures and successes.	2,925 (3900)
	FastCompany: A post that recounts the journey of Jeff Hawkins, whose perseverance and willingness to learn from mistakes led to market successes.	5,199 (6932)
News article[a]	*The Newsweekly of Information Systems Management*: A news article that discusses the innovative technological developments around GRiDPad.	406 (542)
	The New York Times: A tribute to John Ellenby, the visionary who founded Grid systems, the company behind the clamshell laptop computers designs.	1091 (1455)
Transcript	*Computer History Museum*: An oral history of Jeff Hawkins, designer of the GRiDPad's pen-based computing program.	21,648 (28864)
	Computer History Museum: a panel discussion on pioneering the laptop, featuring Glenn Edens, Carol Hankins, Crag Mathias, and Dave Paulsen.	10,928 (14571)
Total		49,444 (65926)

[a]We selected only news articles available on the Internet; articles derived from the Factiva database were excluded from our dataset.

Our trial covered the range of LLM-based tools that are currently being promoted to qualitative researchers. The first option is to use a general-purpose LLM chatbot for data analysis. ChatGPT has so far been the most common tool used in existing studies, so we also adopted it in the interests of comparability of results. We acknowledge that even research making use of ChatGPT is not directly comparable, given the regular release of new versions (we used the latest available at the time, ChatGPT-4o); however, OpenAI itself concedes that there are only incremental differences between recent versions (OpenAI, 2025).[10] Having selected a tool, we then decided on the form that the evaluation of it would take. As ChatGPT is widely purported to be capable of expediting the 'coding' process to identify themes and patterns (Table 1), we

examine its ability to perform these tasks efficiently and accurately. Specifically, we assessed whether ChatGPT could reliably code the data, recognize key themes, and provide coherent insights without compromising the depths and nuance in the data.

The first author ran multiple rounds of testing on the dataset, keeping a log of the prompts used and the outputs. To ensure comparability, we reused prompts that had been run in previous studies, so as to benchmark against them, but also experimented with a variety of qualitative data analysis prompt templates drawn from practitioner blogs and scholarly articles. In addition, we formulated prompts based on 'state-of-the-art' engineering techniques developed by Google engineers called 'chain-of-thought prompting' (Wei et al., 2022b). Chain-of-thought prompting[11] involves creating sequential prompts that instruct an LLM chatbot to complete a series of intermediate steps, where the outputs of one step become the input for the next to enhance the overall quality of the final output. This is in contrast to zero-shot prompts, where the model generates responses without prior examples, and few-shot prompts, where the model is provided with examples.[12] Appendix 1 in the supplemental material lists the different prompts that we used.

In addition to our evaluation, we compared our results with five other studies that explored the use of ChatGPT as an 'analytical tool' to augment or automate qualitative coding and thematic analysis (see Table 3 for details). We chose these studies as they represent the range of approaches that have been used, in terms of diverse prompts, datasets, ChatGPT models, and disciplinary backgrounds (i.e., political science, medicine, education, occupational therapy, sociology). This enabled us not only to benchmark the main steps in a GenAI-enabled qualitative data analysis but also to draw on and incorporate other evidence that evaluated the potential and limitations of these models. Moreover, as each study includes a reflection on challenges encountered at different stages of analysis, we were able to identify and synthesize the recurring themes and obstacles of a GenAI-enabled analysis. Table 3 summarizes each study (including our own), including the model version, type of data and analysis, prompting techniques, and the authors' reflections on leveraging GenAI for qualitative data analysis.

At the same time, we undertook an analysis of Qual-AI offerings, comparing the features on offer from the main CAQDAS providers and from emerging start-ups. Table 4 documents these features, which center on 'chatting' with documents, text summarization, and code suggestions. In addition to comparing the features of Qual-AI tools currently on the market, we also examined two products in greater depth: one from an established CAQDAS provider (NVivo) and one from a startup (QUALSOFT[13]). We selected these two because our comparative analysis identified them as the most promising for academic use. In conducting the analysis, we were careful to follow the instructions that the developers provided to obtain optimal results, which included carefully selecting data sources that are easier to analyze, such as interviews, which have a consistent question–answer format. We also compared the results we obtained against coding results produced by Custom GPTs available on OpenAI's GPT Store,[14] where we tested several purpose-built models for qualitative data analysis, including those fine-tuned for coding and thematic analysis. Overall, then, our evaluation was able to make extensive use of triangulation: multiple GenAI models, a variety of data sources, and multiple investigators.

In the course of our analysis, we realized that we needed to extend its scope to consider not just the epistemic risks posed by the tool's output but also how these outputs are interpreted by authors and Qual-AI developers. In sum, we approached our inquiry guided by the interpretive principles of qualitative research: we critically examined the nature of the phenomenon we were studying, compared within as well as across studies, and undertook a holistic investigation of the meaning that researchers ascribed to the tools, rather than confining ourselves to the outputs alone.

Table 3. Evaluation of ChatGPT as 'Scientific Tools' to Augment or Automate Qualitative Data Analysis.

	GRiDPad (authors)	Study 1 (Lee et al., 2024)	Study 2 (Hitch, 2024)	Study 3 (Sen et al., 2023)	Study 4 (De Paoli, 2024)[a]	Study 5 (Turobov et al., 2024)
Model	ChatGPT-4o	ChatGPT-3.5	ChatGPT-4	ChatGPT-4	ChatGPT-3.5 Turbo - API	Custom GPT
Analysis type	Coding and thematic analysis	Coding and thematic analysis	Coding and thematic analysis	Coding and thematic analysis	Coding and thematic analysis	Coding and thematic analysis
Prompt type	Zero-shot, few-shot and Chain-of-thought	Few-shot	Zero-shot	Zero-shot	Zero-shot	Few-shot, Chain-of-thought
Step 1. Upload data	Publicly available data on GRiDpad's development, the first portable tablet computer.	Transcript of Diabetes Discussion: A Diabetes UK Podcast featuring 2 guests sharing their diabetes experiences.	Newspaper article on long COVID-19 featuring interviews between patients and health experts.	Interview transcripts of two participants from an unpublished study on university compliance.	Open access interview datasets from a gaming-horizon project and a teaching undergraduates project.	63 open access UN policy documents and press releases.
Step 2. Code generation	Random, inaccurate, and inconsistent requiring multiple rounds of conversations to achieve a closer approximation to the data.	'successfully identified multiple codes in the transcript …, and the corresponding codes match the textual content' (p. 4)	'Generally speaking, the codes identified were very similar. However, subtle but identifiable differences were also evident' (p. 599)	'Direct quotes selected for the generated codes appear to be extremely accurate. […] [ChatGPT] produces different codes from the same paragraph' (p. 6)	'the model would sometimes get hallucinated and generated new code names out of existing ones' (p. 1006).	The 'tendency to generate more descriptive than interpretive response (i.e., codes)' (p. 8).
Step 3. Iterative theme refinement	Repeated rounds of conversation leading to variations in focus and themes.	'several rounds of analysis … led to interesting variations … [that] indicates that codes were interpreted from different angles, thereby adding more layers' (pp. 5–6)	'AI developed themes are more 'literal' in comparison to the reflexive thematic approach' (p. 600).	ChatGPT 'could create meaningful code, category and themes' (p. 13).	There 'may be potentially valid themes which represent the data. However, there are a few other themes which … have been overlooked' (p. 1008).	Despite errors, 'results … are impressive … GPT can enrich thematic analysis by providing detailed, nuanced insights' (p. 7).
Challenges with ChatGPT	• Errors and hallucinations • Limited context window • Controlling model behavior	• Errors and hallucinations • Limited context window • Generates responses that are not justified by data	'not currently capable of the contextual interpretation and reflective deliberations' (p. 604)	• Errors and hallucinations • Limited context window • ChatGPT does not always 'behave' as desired and makes	• Errors and hallucinations • Limited context window • 'Different prompts (even aimed at the same output' often	• Errors and hallucinations • Descriptive outputs • The 'validity of the research might be compromised due to the … randomness …

(continued)

Table 3. Continued.

	GRiDPad (authors)	Study 1 (Lee et al., 2024)	Study 2 (Hitch, 2024)	Study 3 (Sen et al., 2023)	Study 4 (De Paoli, 2024)[a]	Study 5 (Turobov et al., 2024)
	• Repetitive and inconsistent	• 'Prompt dependent' which requires repetitive rounds of analysis (p. 5)		mistakes that require the process to be repeated (p. 13)	lead to different responses (p. 1016).	of ChatGPT's outputs' (p. 9).
Conclusion	LLM chatbots are the wrong tool for qualitative data analysis. Researchers using LLM chatbots become trapped in an infinite loop of chatbot conversation.	'ChatGPT has the potential to function as a valuable tool during analysis, enhancing the efficiency of the thematic analysis and offering additional insights into the qualitative data' (p. 9).	'AI can augment but not adequately replace human researchers' (p. 604)	ChatGPT 'can be used as an auxiliary software for researchers in qualitative research data analysis, meaning emergence, code, category and theme creation processes' (p. 14).	'the likely scenario is not one of the human analysts being replaced by AI analysts, but one of Human-AI collaboration' (p. 1016).	'AI tools can transform traditional qualitative research methods. … [with] a custom GPT model, researchers can handle larger datasets more efficiently, uncover nuanced insights more quickly' (p. 10)

Note. The five comparative studies are published in a range of less-established journals, some authored by qualitative researchers and others from different scholarly backgrounds. While journal outlet is not necessary indicative of quality, these articles have been recognized and promoted by reputable sources, including the University of Surrey's well-regarded CAQDAS networking project.

[a]In the pre-print version of this paper, the claims regarding ChatGPT's performance were stronger than in the published version.

Table 4. Qual-AI Features and Developers' Descriptions.

Provider	GenAI Features	Developer Description
NVivo	Text summarization	Select text in a document and automatically summarize the selected text in the preferred language and length.
	Child code suggestion	Select a code and generate a child code based on the coded content.
ATLAS.ti	Intentional AI coding	Provide your intention and research goals to produce highly relevant codes your research deserves—on autopilot.
	AI code suggestions	Dive deeper into your qualitative data and save time [with] AI-driven code suggestions as you code.
	Conversational AI	Chat directly with your documents and have them automatically coded based on your intentions, providing customized results.
	AI summaries	Turn your research data into summarized insights.
MAXQDA	AI coding	Automate your coding process with AI coding while maintaining complete control over your analytical work.
	AI subcode suggestion	Get new code recommendations based on the selected text passage.
	Chat with your data	Pose questions about already-coded text segments or entire documents.
	Summarize content	Generate summaries in your preferred language, length, and format.
QInsights	Theme analysis	AI assistant, Q, intelligently scans your data to identify recurring topics, phrases, and ideas … to uncover the key themes.
	Conversational analysis	Simply ask questions, and QInsights provides insights you can validate directly in the transcript.
CoLoop	AI chats	CoLoop's AI chat works just like ChatGPT, but based on your research material.
	Topline summaries	Automatically generate topline summaries for the uploaded interviews data.
AILYZE	AI chatbot	Chat with files and analyze.
	AI thematic analysis	Ask AI to generate themes or specify your own themes.
	AI summaries	Ask AI to generate an overall summary or an individual summary of each document.
	AI-assisted coding	Ask AI to generate codes.

AR-LLM Chatbots as a Tool for Qualitative Data Analysis? The Evidence

We have established that rather than directly processing raw data or input texts in a systematic way, a GenAI model reduces and transforms the data into vector embeddings—numerical representations of input text—to generate statistically plausible responses based on patterns encoded during training. We now illustrate the implications of these differences by presenting our results and a comparison of GenAI-enabled qualitative data analysis. In doing so, we detail the steps we took to account for the model's known limitations to ensure a fair and systematic evaluation while clarifying and illustrating the conclusion from our study: GenAI models are not an analytical tool suited to qualitative data analysis.

ChatGPT Results

Step 1: Upload Data

A GenAI-enabled qualitative data analysis begins with uploading documents[15] (e.g., interview transcripts, news articles) or copying and pasting text into a model like ChatGPT. For our

evaluation, we uploaded our dataset in Word format[16] to three separate ChatGPT-4o chat windows. Document 1 contained magazines, blog posts, and news articles (approximately 21,603 tokens), Document 2 contained an interview transcript (approximately 28,864 tokens), and Document 3 was a transcript of the panel discussion (approximately 14,571 tokens).[17]

In the comparison studies, Studies 1 and 2 uploaded their data—a podcast transcript on living with diabetes and a news article on long COVID—to a single chat window. Study 3, on the other hand, copied and pasted their interview transcripts into two separate chat windows. Study 4, utilizing OpenAI's API with ChatGPT-3.5 Turbo, processed their datasets by dividing them into roughly 2,500-token chunks, separating the 13 gaming interviews into 56 text chunks and the 10 teaching interviews into 35 text chunks, which were then copied and pasted into ChatGPT. Study 5 adopted a different approach and created a custom GPT model. The authors uploaded their dataset (63 open-access UN policy documents and press releases on the topic of AI) to the model to serve as its knowledge base and provided explicit instructions for the model to 'act as an academic expert in qualitative coding and thematic analysis.' Separating the data posed practical challenges, including reconciling the different conversations, codes and themes across multiple chat windows. But this approach is necessary to stay within the model's context window.

Step 2: Code Generation

Having separated the data into separate chat windows, we used multiple and different prompts to instruct ChatGPT to code the datasets as a first step in conducting a thematic analysis. Within about 10–20 seconds, ChatGPT generated synthetic word strings that linked associated phrases and terms, as well as example descriptions under themes. While the speed of automation was notable, and the model was able to reproduce the typical format found in a qualitative paper of a table of quotations, the results did not stand up to scrutiny.

Table 5 is an example of the 'coding' performed by ChatGPT. The left-hand column presents the outputs of the model, and the right-hand column shows the original text[18] along with our evaluation. Sections in the dataset that align with quotes used by ChatGPT are presented in normal font, and deviations are bolded. In this instance, ChatGPT did not code the data, in the sense of line-by-line tagging of text, but rather provided output labelled 'themes' and 'quotes'. Some of the quotes, although related to the object of the study, were not based on our uploaded dataset. The 'themes' were not themes in the sense of labels of patterns in the data; rather, at best, they approximated a summary of the corresponding quote. Moreover, they were either incorrect (e.g., there is no reference to 'evolution'), too general to be useful (e.g., 'technological challenges'), or misleading (e.g., 'standardization and collaboration' refers to the possible development of industry standards, but without any actual collaboration having taken place). Although these results are expected given what we know about the technology, for illustrative and comparison purposes, a new chat log was created, the dataset re-uploaded, and the model was prompted again to code the dataset to conduct a thematic analysis (see Appendix 2 in the supplemental material for different iterations and variations in results).

The randomness and unstructured nature of ChatGPT's qualitative 'coding' was also explicitly reported in Studies 2, 4, and 5. Generated 'codes' were repetitive, inconsistently applied across similar data segments, or overly descriptive. Moreover, as reported in Studies 1, 3, 4, and 5, while some codes appeared to be plausible, the corresponding quotes were often truncated, fabricated, or drawn from other sources. These errors necessitated repeated cycles of conversation and manual human intervention, involving the creation of a new chat window, re-uploading the data, and re-prompting the model, which typically resulted in variations or new codes, themes and quotes. Despite this repeated effort, the results did not improve.

Table 5. Excerpts of ChatGPT coding for qualitative thematic analysis

Zero-shot prompt 1 (see Appendix 1)

Excerpt of themes and quotes generated by ChatGPT	**Quotes: comparison with dataset**	**Evaluation of 'themes'**
	The stylus allows us to interact more naturally with the computer. It is better at manipulating text and graphics on a computer display. ***Source*:** not from our dataset, but can be found on Factiva.	The theme is wrong: there is no evolution referred to
	'handwriting recognition' is a phrase mentioned in our dataset. ***Source*:** unknown, possibly fabricated	Very general theme, yet still wrong: the quote mainly refers to a technological advance that still has some constraints
	Market research company Infocorp, **based in California**, predicts that at least 3.4 million units of pen interface notebook systems will be sold every year every year by 1995, with 7 million units in use. ***Source*:** not from our dataset, but can be found in Factiva.	The theme is very general: 'sales forecast' would be more accurate
	Grid **organised** the event to initiate industry standards for pen interface computers and to encourage software development. ***Source*:** not from our dataset, but can be found in Factiva, change one word from British to American English.	The theme is misleading, as the standardization referred to is the potential development of industry standards, and potential not actual collaboration
	The San Francisco Examiner has broken new ground by conducting the first ever election-day exit poll using GRiDPAD, **the only computer in America that recognizes handwriting and uses an electronic pen interface instead of a keyboard.** ***Source*:** not in our dataset, but can be found on Factiva.	It is a stretch to denote the one-off trial deployment of a product as a commercial use case

Step 3. Iterative Refinement of Themes

Given the variation in accuracy and quality, including coverage of coded data across the different chatlog conversations, we proceeded with the chat windows across our three datasets that showed fewer errors. We then used a variety of prompts to instruct ChatGPT to identify the patterns and relationships between the coded segments of data in order to develop higher-order themes. After eight cycles of conversations, the model's behavior began conforming to our expectations; that is, it began producing the appearance of the kind of results that we were explicitly soliciting. In this iteration (Table 6), ChatGPT generated 'codes' and 'themes' that were based on our dataset. The left-hand column presents the outputs, and the right-hand column shows the original text, along with our evaluation. As before, sections in the dataset that align with quotes are presented in normal font, and deviations are bolded.

We compared the generated themes against the original data and found that, as in the previous step, outputs focused on specific parts of the document, to the exclusion of the rest; and 'codes' and 'themes' remained at best vague terms and phrases that related to the corresponding quote. The two 'higher-order' themes (each comprising two codes, making for the appearance of a very orderly data display) were an amalgamation of the codes. Despite the face validity of the output (Table 6), the results were inconsistent and random (e.g., the code of 'corporate history and evolution' under the theme of 'market strategy and business evolution'). This necessitated more rounds of conversations, leading to subtle variations and, at times, significant deviations, demanding constant human intervention for verification.

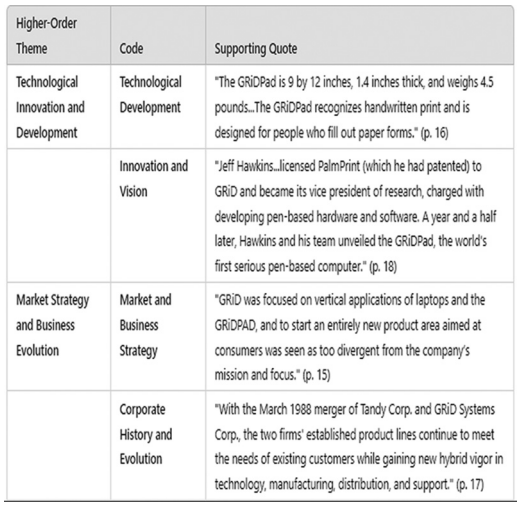
The demand for manual oversight not only negates any supposed efficiency gains but highlights the outright impossibility of scaling up qualitative designs and establishing trustworthiness using GenAI models. Because of the random nature of the text generation, a GenAI-enabled qualitative data analysis is an open-ended process, with the researcher trapped in what is effectively an infinite loop of chatbot conversations. As each new interaction generates a variation in response, it also produces an ever-expanding body of content that needs to be interpreted and evaluated, without the researcher achieving a stable set of insights (see, for example, the variations and randomness across Tables 5, 6 and Appendix 2 in the supplemental material). This infinite loop means that rather than facilitating analysis in a transparent and trustworthy manner, researchers become content moderators for each chatbot output, validating its accuracy, contextual relevance, and connection, if any, to their data. This iterative process is ethically concerning, as not only is it resource-intensive—in terms of energy consumption (Luccioni et al., 2024)—but also exploitative, as researchers effectively train and refine GenAI models instead of engaging in analysis.

The inadequacies that we encountered were confirmed by our cross-study comparison. Similar to our observations, the authors of Study 4 report that ChatGPT overlooked significant parts of the texts while incorrectly assigning codes and hallucinating. However, in stark contrast to our conclusion, they suggest that 'the likely scenario is not one of the human analysts being replaced by AI analysts, but one of a Human-AI collaboration' (De Paoli, 2024, p. 1016). Similarly, although confronted with unstable, inconsistent and inaccurate results, the authors of Studies 2 and 3 conclude that the model can conduct a thematic analysis, albeit with limitations. These include errors and hallucinations, and an inability to recognize context or subtle nuances, underscoring the need for human oversight (Study 2) or the use of other tools to ensure accuracy and trustworthiness (Study 3).

Studies 1 and 5 also had disappointing results, yet reported them in generally favorable terms. For Study 1, variation in themes across different chat windows 'indicated' to the authors that ChatGPT interpreted the data from 'different angles, thereby adding more layers to the overall analysis' (Lee et al., 2024, p. 6). This led the authors to conclude that ChatGPT is a 'valuable tool' for qualitative data analysis. In Study 5, despite recognizing that 'the validity of the research

Table 6. Excerpts of ChatGPT coding for qualitative thematic analysis

Few Shots Prompt (see Appendix 1).

Excerpt of high-order themes, codes and quotes generated by ChatGPT	Original text	Evaluation of 'codes' and 'higher-order themes'
	The GridPad is 9 by 12 inches, 1.4 inches thikc and weighs 4.5 pounds. **Grid calls it "handheld." I would not go that far, but it is at least an "arm-held." I tended to rest it in the crook of an arm like a clipboard. It does not come with a keyboard.** The GridPad recognizes handwritten print and is designed for people who fill out paper forms. ***Source***: Dataset 1, truncated text. Page number do not match.	The 'higher-order theme' is an amalgamation of the two 'codes'. This theme is too general to be useful
	Hawkins left Berkeley jazzed and confused – jazzed about what he had discovered, confused about how he would earn a living. So he re-upped with GRiD, this time on different terms. He licensed PalmPrint (which he had patented) to GRiD and became its vice president of research, charged with developing pen-based hardware and software. **Later that year, GRiD was purchased by Tandy, an electronics manufacturer and the parent company of Radio Shack.** ***Source***: Dataset 1, incorrectly cited. Page number do not match.	
	"They saw it as just too big of a risk," noted Geoff Walker, GRiD's director of OEM Products, in comments to Pen Computing. "GRiD was focused on vertical applications of laptops and the GRiDPAD, and to start an entirely new product area aimed at consumers was seen as too divergent from the company's mission and focus."	The 'higher-order theme' is an amalgamation of the two 'codes'. This theme is too general to be useful

Higher-Order Theme	Code	Supporting Quote
Technological Innovation and Development	Technological Development	"The GRiDPad is 9 by 12 inches, 1.4 inches thick, and weighs 4.5 pounds...The GRiDPad recognizes handwritten print and is designed for people who fill out paper forms." (p. 16)
	Innovation and Vision	"Jeff Hawkins...licensed PalmPrint (which he had patented) to GRiD and became its vice president of research, charged with developing pen-based hardware and software. A year and a half later, Hawkins and his team unveiled the GRiDPad, the world's first serious pen-based computer." (p. 18)
Market Strategy and Business Evolution	Market and Business Strategy	"GRiD was focused on vertical applications of laptops and the GRiDPAD, and to start an entirely new product area aimed at consumers was seen as too divergent from the company's mission and focus." (p. 15)
	Corporate History and Evolution	"With the March 1988 merger of Tandy Corp. and GRiD Systems Corp., the two firms' established product lines continue to meet the needs of existing customers while gaining new hybrid vigor in technology, manufacturing, distribution, and support." (p. 17)

(continued)

Table 6. Continued.

Few Shots Prompt (see Appendix 1).		
Excerpt of high-order themes, codes and quotes generated by ChatGPT	**Original text**	**Evaluation of 'codes' and 'higher-order themes'**
	Source: Dataset 1, truncated text. Page number do not match. With the March 1988 merger of Tandy Corp. and GRiD Systems Corp., the two firms' established product lines continue to meet the needs of existing customers while gaining new hybrid vigor in technology, manufacturing, distribution and support. ***Source:*** Dataset 1. Page numbers do not match.	

might be compromised due to accuracy, randomness, or unstructured nature of ChatGPT's outputs,' the authors concluded that 'AI tools can transform traditional qualitative research methods … [as] researchers can handle larger datasets more efficiently, [and] uncover nuanced insights more quickly' (Turobov et al., 2024, pp. 9–10).

Across the five studies, the authors acknowledged the challenges in using ChatGPT as a tool for qualitative data analysis, in particular, its tendency to produce erroneous and false results. However, they argue that despite these errors, efficiency gains are significant. These include faster processing times and reduced manual effort in coding and theme identification. Moreover, while these studies concede that human oversight is needed to ensure accuracy and validity, they suggest that with the supposedly rapid advancement of AI, the need for this oversight may decrease over time as these models 'learn the craft' of qualitative data analysis (Study 2). In sum, they obtained comparable results to ours but drew opposing conclusions.

Qual-AI Results

The shortcomings we identified in the previous section are not unique to ChatGPT but extend to the specialized Qual-AI offerings currently on the market, as they all rely on OpenAI's foundational ChatGPT model and therefore inherit its defects, limitations, and shortcomings (Bommasani et al., 2022). Although these tools are advertised as being 'purpose built' for qualitative researchers and claim to mitigate hallucinations through the integration of RAG architectures, this integration does not eliminate the risk of fabricated or misleading outputs. This is because while RAG architectures may improve the output by retrieving from relevant external documents, the model can still misrepresent this information, and retrieval quality itself is highly variable. As such, although it could be expected that the results for these tools would be better, they do not resolve the reliability problems produced by general-purpose LLM chatbots.

We illustrate these problems with the results from our tests using a frontrunner product, QUALSOFT, but note they were found across all the Qual-AI offerings we analyzed (Appendix 3 in the supplemental material). To do so, we concentrate on the scenario with the best performance that we obtained, which was to examine a single interview transcript containing 21,648 words (28864 tokens), rather than compare across multiple documents, whereupon the results rapidly deteriorated. Certainly, if used within the limits that developers advise—i.e., using the tools on small datasets with similar types of documents—we found that the tool can obtain results mostly from the researcher's own data. However, in many other ways, the results were as unsatisfactory as when using ChatGPT.

We commenced with a thematic analysis of the transcript, the first step recommended by the QUALSOFT developers. This process is automated. The researcher is not able to customize the prompts or calibrate them to specify the analytical purpose or the desired granularity of the analysis. We reran the thematic analysis (see Appendix 4 in the supplemental material) and on each occasion, the tool produced a list of five or six themes with accompanying summaries. While all chatbot-produced themes are mentioned in the transcript, many are not central, some themes overlap, or are the same theme by a different name, and some themes misrepresent what the interviewee said. Other key themes in the document (e.g., convergence of two technologies) are not mentioned. The summaries are banal and omit detail, for example, 'simplicity' is repeatedly mentioned, but not the contextual conditions that made this important.

The themes and summaries are accompanied by the quotes from the document on which they are supposedly based. These quotes match the original text, but it is not clear how they are connected to the themes, nor do they represent all the mentions of the theme in the document. The extracts are of very different lengths—some are several paragraphs long, some are an excerpt

from a sentence—but it is not clear why the quotes start or stop where they do. In some instances, the extracts run answer and question together. The Qual-AI output drew heavily from some parts of the interview and ignored others, with no apparent reason. It was also unable to deal with the temporality of the interview, which was chronological in structure, starting from the interviewee's childhood.

We also used the software's querying function, i.e, asking questions of the chatbot. Some portray querying as a new form of analysis ('conversation analysis') that amounts to a paradigm shift: a replacement for the traditional labor-intensive analytical practice of coding (Friese, 2025; Morgan, 2023). Unlike automated thematic analysis, this form of analysis promises interactive engagement with data by allowing the researcher to pose questions, probe themes, and identify patterns. To frame the questions, we followed the advice in the user guide, using the recommended prompts to optimize performance. However, the LLM chatbot's outputs (which, like the themes, took the form of summaries) were routinely incomplete.

To illustrate: when prompted to list the challenges mentioned by the interviewee, several important challenges were omitted. Beyond omissions, the chatbot also generated incorrect responses, especially when queried about information that was not answerable from the document. It was also prone to wrapping the text it generated in quotation marks, as if it were actually quoting the original. It was not reliably able to distinguish between different people referred to in the text, different companies, or even between people and companies. It would produce vague generalities even when asked specific questions that were answerable from the text. Some of the errors were obvious upon a first read, but others could only be detected following our own in-depth analysis of the transcript (see Appendix 5 in the supplemental material for more detailed evidence). All these errors are to be expected, given the limitations of the technology discussed above.

For this reason, the outputs had to be cross-checked against the document and corrected. Overall, the chatbot output did not deepen our understanding of the transcript. Its main function (irrespective of whether it was termed theme or conversation analysis) was in the form of synthetic summaries, which have limited use for any in-depth qualitative analysis, and which were not accurate enough to be dependable. The summaries were displayed on a separate screen from the interview transcript, which drew our attention away from the actual content of the interview, rather than increasing our familiarity with it.

Beyond these content-related issues, the lack of an audit trail makes it difficult to trace how outputs are generated or to assess their reliability and connection, if any, to the data. This opacity is compounded by the fact that Qual-AI developers do not disclose their prompting techniques, design choices or RAG architecture. These persistent shortcomings raise critical questions about the trustworthiness of (Gen)AI-assisted qualitative data analysis. We will now consider the broader implications of these findings.

Discussion

The widespread accessibility of GenAI models, which are designed to be used by non-experts in an intuitive conversational way, has been positioned as a compelling alternative to more traditional CAQDAS software and NLP approaches for qualitative data analysis (Sen et al., 2023; Turobov et al., 2024). However, in our evaluation, LLM chatbots consistently failed to code the data efficiently, accurately, reliably and comprehensively, or identify meaningful themes. ChatGPT, and the Qual-AI products built on it, are simply not designed nor fine-tuned to perform even the lower-order analytical tasks of searching and coding text—as we have shown, this is beyond the capabilities of autoregressive LLM models, even when augmented to provide Qual-AI solutions. These automated processes and features of

GenAI models should make it apparent that they cannot function as a scientific tool for qualitative data analysis.

Obtaining inaccurate, opaque and unstable results is not the only epistemic risk we identified. In this section, we also address the less expected finding from our study: the logical errors that researchers committed—in the form of anthropomorphic fallacies, causal misattributions, and apologetics—when confronted with disappointing chatbot outputs. Authors attributed poor results to factors—such as users' lack of prompt engineering skills, or the immature state of the technology, which they assume will be corrected in future releases—other than the inherent limitations of LLM chatbots themselves. We now discuss these epistemic risks of chatbot use in qualitative data analysis, and how commonly suggested remedies do not address them—in fact, they only exacerbate the potential harms.

Epistemic Risk No 1: Category Error

Believing that it is possible to use an LLM chatbot for qualitative data analysis commits what we would term a category error—it mistakes a synthetic predictive next-word text generator for an analytical aid. As we have made clear, LLM chatbots replicate word patterns contained in their sets of rules. The word strings that are generated may have the appearance of analytical findings: they use the same structure, layout, and academic rhetoric. But to assume that this implies actual analysis has been conducted is wrong. In the case of qualitative data analysis, the ubiquity of 'templates' (Köhler et al., 2022) for presenting findings (e.g., Gioia's data structure) means that the chatbot can conform to our expectations of what analytical findings should look like (e.g., Table 6). The sheer speed and volume of AI-generated outputs can also reinforce this appearance of a comprehensive analysis. But this falsely equates computational efficiency with human reasoning and interpretation. It also ignores the interpretive nature of qualitative data analysis, which relies on the researcher's contextual understanding, subjectivity, embodied experience, and reflexivity.

LLM chatbots are also not doing systematic search and retrieval of the input data that the user has submitted, as would traditional qualitative data analysis software such as Quirkos. In processing the input, LLM chatbots match word patterns against those found in the textual corpus used to generate their ruleset and produce output that follows those rules. Anything falling outside their ruleset is either ignored or force-fitted into the rules. The output generated therefore cannot be reliably faithful to the text that the user has submitted. Because of this matching process, LLM chatbots cannot confine themselves to the data that the researcher has uploaded, and, although less likely to occur in purpose-built Qual-AI platforms, this is why chatbot output routinely contains text not in the input data (as per Table 5). The synthetic text generated thus cannot be considered an honest or accurate reflection of the researcher's data. This divergence raises significant ethical and validity concerns about the provenance of the output and the link between evidence, ideas, and conclusions that can be reached.

Epistemic Risk No. 2: Producing Unreliable Outputs

Chatbots are inherently unstable and unreliable—and, as we have outlined, this cannot be remedied without a fundamental shift in the underlying architecture and method deployed to process and generate synthetic text (Dziri et al., 2023; LeCun, 2022; Lin et al., 2021). The randomness and variability of ChatGPT's outputs become apparent when instructed to carry out complex analytical tasks like qualitative coding and thematic analysis. In our evaluation of ChatGPT, as well as in our five comparative studies, controlling model behavior was an ongoing challenge. When ChatGPT was explicitly tasked to code data, the model sometimes generated not only codes but

also themes, while in other instances, it either annotated text segments according to their literal content or generated 'example', 'supporting' or 'illustrative' quotes that, like other outputs from an LLM chatbot, were often fabricated. In our evaluation of Qual-AI, outputs were more stable when chatting was confined to descriptive queries. However, this apparent stability should not be mistaken for increased reliability; it is a consequence of restricted interaction parameters that developers have employed to reduce variability in outputs.

The instability and unreliability of results also mean that the process becomes time-consuming. As LLM chatbots cannot consistently produce stable or reliable outputs, qualitative analysts using ChatGPT or other tools that build on them, including Custom GTPs (Study 5), need to engage in multiple rounds of prompting. The much-touted efficiency of the process is erased through the need to prepare the data for uploading, engage in multiple rounds of prompting, checks, and cleansing the output of errors, without any endpoint in sight. This was also the case with Qual-AI software, whose developers do provide disclaimers about the accuracy of GenAI outputs.[19] Because the output cannot be error-free, researchers need to keep returning to their dataset to use other means to check, re-code, and re-analyze it—meaning that the traditional process of qualitative data analysis remains essential. For this reason, posing questions in an LLM chatbot window constitutes a considerable distraction from the interpretive process of familiarizing ourselves with, and deepening our understanding of, our data.

Epistemic Risk No. 3: Anthropomorphic Fallacies

The process of multiple rounds of prompting has been construed as 'iterative dialogue' (Friese, 2025), with the risk that the model's coherent, convincing, and seemingly insightful responses lead to the anthropomorphization of chatbots as research collaborators (Studies 1 and 4), virtual colleagues (Study 2), or research assistants (Study 5). As we have discussed, the ability to 'chat' *with* our data has been promoted as a new GenAI-enabled method that could expand and deepen our understanding of qualitative data. Used in this way, GenAI tools are positioned as 'collaborators and assistants' (Friese, 2025) interacting with the data alongside researchers, who are able to probe for deeper meaning.

Although appealing, this claimed conversational method ignores the technological limitations of LLM chatbots. Although these models are, by design, capable of 'chatting' with users, this should not be mistaken for analysis; that is, systematically searching and probing input text. When researchers are 'chatting' with LLM chatbots, they are not engaging with their data: they are eliciting synthetic algorithmic outputs that lack the meaning and communicative intent that characterizes human language (Bender et al., 2022). For qualitative data analysis, this means that these tools are *not* searching, coding, or undertaking interpretive or theory-based (thematic) analysis with the user's data. Instead, as we know from the discussion above, LLM chatbots generate synthetic text strings probabilistically, assembling words and phrases based on statistical patterns to generate responses for users. Put simply, they are mimicking forms of human text that are commonly associated with the process of analysis.

Epistemic Risk No. 4: Causal Misattribution by Blaming the User, Not the Tool

Advocates argue that ChatGPT can unlock unparalleled productivity gains, provided that users can formulate effective prompts. This includes formulating clear and specific instructions, offering sufficient contextual information, and occasionally guiding the model with examples. Some enthusiasts take it a step further and suggest that users need to motivate the LLM chatbot by convincing it, and

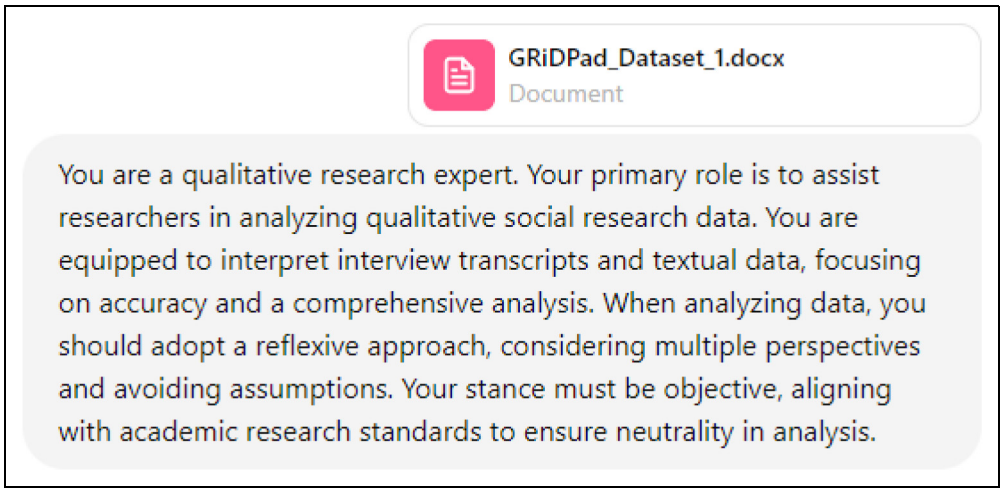


Figure 3. Example Prompt to Motivate ChatGPT to Conduct Qualitative Data Analysis. Source: Friese (2024; see also Turobov et al., 2024).

perhaps even themselves, that it can perform qualitative data analysis according to the standards of the social sciences (see Figure 3 for an illustration). By formulating more effective and convincing prompts, enthusiasts argue that users can enhance the relevance and accuracy of the model's outputs, thus increasing its utility in qualitative analysis (De Paoli, 2024; Tabone & de Winter, 2023).

While it is true that the wording of input prompts influences the outputs of GenAI models (Wei et al., 2022b), offering prompting as a fix not only overstates the capabilities of these models and contributes to overhyped narratives about their potential, but it also obscures their limitations. Moreover, it redirects undue accountability onto users, suggesting that poor results stem from their inability to craft prompts, rather than the technology's inherent limitations (Studies 1, 4, and 5; see also Dziri et al., 2023; Lin et al., 2021). Recent expert evaluations on the effectiveness of diverse prompting techniques demonstrate that '[d]espite their remarkable success [to steer model behavior and output], challenges persist, including biases, factual inaccuracies, and interpretability gaps, necessitating further investigation and mitigation strategies' (Sahoo et al., 2024, p. 8). For these reasons, experts and developers of GenAI models, including OpenAI, have issued a caution to users 'to remain skeptical of claims about [prompting] method performance' (Schulhoff et al., 2024, p. 43). Given that errors, biases, and failures of reproducibility and explainability are inherent to GenAI models, they cannot be mitigated with more 'effective' or 'convincing' prompts.

When proponents claim that improved results stem from reworking and refining prompts, it is misleading, as these better outcomes are the result of their priming activities aimed at influencing a model's outputs. Users consciously and unconsciously prime LLM chatbots through several methods. These include providing direct feedback by rating responses with thumbs up or thumbs down; using prompts that offer examples to guide a model's focus (e.g., few-shot prompts; Study 1); repeatedly selecting 'Regenerate response' until a desired result is observed (Study 3); and providing instructions to guide and steer its behavior (Study 5). While these actions may appear trivial, they create the illusion that these models are 'learning' and 'improving' over time (Study 2) when, in fact, they are not.

What LLM chatbots are doing is incorporating user data and conversations into subsequent responses, tailoring outputs to match user preferences, much like how social media algorithms curate content recommendations based on user activities. This priming effect can be seen in

our illustration of a GenAI-enabled qualitative data analysis (Tables 5 and 6) where the model progressively produces results that begin to resemble popular qualitative data structure presentations, as per our prior instructions/prompts. Thus, the notion that *better* or *more convincing* prompts lead to better results is erroneous; an example of the 'magical thinking' (Morris, 2023) about the potential of LLM chatbots in science.

Epistemic Risk No. 5: The Oracle Effect

Machine-generated outputs are often mistakenly viewed as neutral and unbiased due to their technological origin (Broussard, 2018)—thus, as more reliable than human analysts prone to error. This misplaced trust creates a false sense of analytical depth, masking the models' limitations. Yet the only meaning in the synthetic text is what we, human researchers, ascribe to it. As Alvarado (2023, p. 6) cautions, accepting output at face value and assuming there is meaning to it is tantamount to treating a technological artifact as an 'electronic oracle.'

When confronted with clearly sub-standard outputs, the response among the authors of the comparative studies in our evaluation was to rationalize or downplay the deficiencies, framing them instead as opportunities to deepen their respective analyses. For instance, although Studies 1 and 3 identified various errors and inconsistencies, these are deemed acceptable as 'humans would also generate different results' (Lee et al., 2024, p. 8; see also Sen et al., 2023, p. 13). In this view, errors, 'hallucinations,' and inconsistencies are not deterrents to LLM chatbot adoption but rather means to 'enrich […] the analytical capabilities of researchers' (Study 5; Turobov et al., 2024, p. 10). Studies 2, 4, and 5 went a step further, suggesting that such errors could instead 'reduce the risk of [human] misinterpretation by providing an alternative analysis against which researchers can test, interrogate, and critique their own analysis' (Study 2: Hitch, 2024, p. 602; see also Study 4: De Paoli, 2024, p. 1014). This treats technology as superior to human reasoning—a fundamentally dehumanizing and ethically concerning move (Bender, 2024) that also misunderstands the nature of human interpretation.

In promoting LLM chatbots as a 'research instrument' for qualitative data analysis, Karjus (2025, p. 6) argues that their limitations 'are not categorically unique to machines and also apply to human analysts.' While this comparison draws attention to the challenges in qualitative data analysis, it is a false equivalence, given that the types of errors produced are different in nature. Human analysts, though subject to biases, possess self-awareness and an ability for critical introspection and reflection, enabling them to learn from experiences and correct initial conclusions. This subjectivity enables human analysts to develop nuanced, contextualized interpretations of the social world, experiences that LLM chatbots can neither imitate nor enhance.

Conclusion

We have shown that LLM chatbots are autoregressive models that generate word strings based on probability and 'learned' representations from their training data. These models excel at generating coherent synthetic text by predicting the next likely word based on statistical methods. As a result, while generated text may seem plausible, it lacks meaning, nuance and accuracy. Applied to qualitative data analysis, this leads to superficial outputs that imitate the form of established methodological practices and reporting templates—such as coding, theme generation and data structure tables—but that are neither grounded in nor justified by data. The superficial appearance of codes and themes not only creates a false sense of validity and depth but also conceals the absence of analytical engagement, while distracting and distorting the researcher's ability to uncover meaningful insights.

Yet, despite considerable evidence, including our own, demonstrating the unsuitability of LLM chatbots as a scientific tool, there is real potential that the widespread enthusiasm for their use will persist, if not increase. This enthusiasm presents a profound epistemic risk to qualitative data analysis—not simply because of the inherent limitations of LLM chatbots, but because of the unwarranted trust placed in their capabilities to act as neutral or objective research instruments. The risk lies in the assumptions, expectations and misplaced confidence we attribute to them, based on unfulfilled promises of automation and efficiency. A contribution of this study is to alert us as a research community to the role that technological tools play in qualitative research—not only in relation to our data, our theorizing and our roles as research instruments, but also how we use and ascribe meaning to our tools.

Research tools are not passive intermediaries; they are constitutive of the knowledge they help produce, shaping both the process and the outcomes of our inquiries. They co-construct the reality we study, becoming part of the epistemological scaffolding that structures our interpretation of findings. As the uptake of GenAI in qualitative data analysis is dependent on the scholarly community's acceptance of these tools, we must critically reflect on the narratives and assumptions surrounding their supposed value. This is about upholding the trustworthiness and integrity of our research.

As we have demonstrated, claims about GenAI's potential to revolutionize qualitative research —by scaling up designs, efficiently and reliably coding data to unlock productivity gains—are flawed, as are assertions that GenAI can lead to expedited insights in a transparent and trustworthy manner. We acknowledge that this conclusion is based on the current state of GenAI technologies, recognizing that its development is ongoing and that future iterations may introduce new challenges and opportunities. However, the flaws that we have identified relate to the fundamental architecture of this technology, so any advancements will need to go beyond the incremental improvements we have seen so far. If we accept these claims without critical evaluation, we mistake the form of commonly used coding templates, which LLM chatbots are able to mimic, for substantive content. These claims also ignore the essence of qualitative research, namely the human act of interpretation, an 'ill-structured activity for which no algorithm can be provided' (Gläser & Laudel, 2013, p. 2). In this paper, we have shown the fundamental differences between 'chatting'—in other words, producing synthetic responses from a chatbot—and systematic analysis of the meaning of our data.

As GenAI and related applications that build on them are increasingly anthropomorphized and hyped for their efficiency, the potential for their widespread adoption may not just threaten the integrity of qualitative research but also the diversity of research approaches, leading to a homogenization in knowledge production. Specifically, the widespread imitation of popular templates risks marginalizing already neglected qualitative traditions such as critical, interpretive, and reflexive methodologies, which are vital for preserving the richness of qualitative insights (Köhler et al., 2022; Mees-Buss et al., 2022). In the debate about GenAI, the very future of qualitative research is at stake.[20]

Acknowledgements

The authors gratefully acknowledge Rebecca Piekkari, Tine Köhler, Roberta Aguzzoli, Ben Aveling, Heikki Mannila, and Steve Pettifer for helping us develop this paper. We also extend our thanks to the participants of the BI/NHH Qualitative Research Day, 'Artificial Intelligence in Qualitative Research Methods' held at BI Norwegian Business School, Oslo Friday 26th April 2024 for their valuable feedback. Additionally, we are grateful to the University of New South Wales School of Management and Governance, Sydney for hosting a research seminar on Wednesday 24 July 2024, where we received constructive feedback on an earlier draft of this paper.

ORCID iDs

Duc Cuong Nguyen https://orcid.org/0009-0003-1383-4565
Catherine Welch https://orcid.org/0000-0003-4389-6612

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Supplemental Material

The supplemental material for this article is available online.

Notes

1. In this paper, we use 'science' in broad terms to include the social and not just natural sciences.
2. In this paper we delimit our focus to text, excluding images and video. This is because 1) qualitative research is mostly text based and 2) there are technical differences between LLMs and diffusion models which generate images/video that fall outside the scope of the current discussion.
3. Beyond input prompts, several technical factors contribute to output variability across different models (e.g. ChatGPT, Gemini, and DeepSeek). These include differences in model parameters, fine-tuning techniques, training data composition and updates, as well as variations in probabilistic sampling methods, and tokenization and computational processes. Collectively, these factors create a non-deterministic system in which model outputs vary unpredictably, even under controlled conditions. As a result, reliability and reproducibility are unattainable, as outputs cannot be reproduced across different interactions, sessions, or models.
4. For example, because of the 'competitive landscape' (i.e., commercial interest), OpenAI does not disclose details of their training processes (i.e., architecture, model size, hardware, training compute, dataset construction, or training method) (Achiam et al., 2023, p. 2).
5. AR-LLMs are updated over time based on data and feedback from users as well as design changes. However, as developers do not disclose such information, it is opaque as to when and how these models are updated, and how, each update affects model behaviour.
6. In writing this paper, we had to go through an explicit process of de-anthropomorphizing the terms we used; it requires conscious effort to avoid such slippage.
7. Empirical evaluations show that GenAI achieves 'poor to moderate and statistically nonsignificant' inter-rater reliability when compared with human coding, particularly in context-dependent tasks (Khademi, 2023: 6; see also Theelen et al., 2024). While fine-tuning prompts and providing expert labelled examples can improve alignment, the inherent variability across GenAI outputs pose challenges for achieving consistent inter-rater reliability results. We would also emphasize that inter-rater reliability is itself a problematic concept when applied to human interpretation and is rejected by many qualitative scholars.
8. 'Qual-AI' is a term currently being used by a leading knowledge hub on computer-aided qualitative data analysis, the University of Surrey's CAQDAS Networking Project. While we conform to its usage in this paper, we note that it blurs the distinction between GenAI and other forms of AI.
9. Uploading confidential or proprietary data to freely available GenAI models poses significant ethical risks related to data privacy, security and the ownership of intellectual property. In addition, it raises legal concerns under regulations such as GDPR (general data protection regulation) that mandate strict protections for sensitive information. The law surrounding AI use is evolving rapidly, exposing adopters to future regulatory risks.

10. At the time of finalizing this manuscript, OpenAI has just released ChatGPT-5, a unified model that replaces earlier versions. Its performance is mixed, with improvements in some areas but declines in others (OpenAI, 2025).
11. In the latest iteration of ChatGPT (i.e., GPT-5), the developers have integrated what they characterize as 'reasoning and thinking' capabilities, based on a pre-trained chain-of-thought model that restructures user input prompts. OpenAI does not disclose how its models edit user prompts, treating this as commercial in confidence. Despite this intervention, outputs are still prone to hallucinations and factual inaccuracies (e.g., OpenAI, 2025; Phan et al., 2025).
12. Computer scientists and LLM chatbot developers have identified over 200 different prompting techniques that can be deployed for different natural language and multimodal tasks (Schulhoff et al., 2024). In this paper, we use the three most common techniques: zero-shot, few-shot, and chain-of-thought, with the last touted as the 'state-of-the-art' technique (Sahoo et al., 2024).
13. We have anonymised the name of the product, because we do not want to damage the commercial prospects of any specific company, when the issues we identify apply across all GenAI-based offerings.
14. There is a misconception that CustomGPTs—a specialized version of ChatGPT that integrates user data—are more reliable and accurate. However, the same limitations of the foundation model (i.e., ChatGPT) are inherited by all other models that build on them (Bommasani et al., 2022).
15. While the process often commences with uploading data, it can also start with data preparation (cleaning, structuring, and organizing) prior to upload, a process that, as in our study, can be more time intensive.
16. We opted for Word to overcome the difficulties known to the community of developers building on ChatGPT when trying to process, for example, PDFs. See: https://community.openai.com/t/gpts-unable-to-read-and-process-pdfs-in-messages-anymore/653669
17. At the time the tests were undertaken, files uploaded to ChatGPT or a custom GPT have a limit of 512MB per file up to 20 files and are capped at 2 million tokens per file.
18. As a part of a larger project, the author who compiled the portable computing dataset also captured news articles on Factiva. These were excluded from the uploaded dataset. When cross-verifying ChatGPT's output, we used the original full dataset as a knowledge base, as per OpenAI developers' recommendations (Achiam et al., 2023), and therefore, were able to identify possible sources from the Internet.
19. For example, NVivo's disclaimer emphasizes that 'Lumivero makes no representation, warranty, or guarantee as to the accuracy, reliability, or error-free performance of the A.I. services'.
20. A note on our use of arXiv sources: arXiv.org is a major online, open-access repository for the latest AI research, but does not itself arrange for papers to be peer reviewed. The platform does distinguish between papers that have been peer reviewed by another publication outlet (i.e., conferences/journals) and those that have not (yet) undergone any peer review (which are labelled as pre-prints). Readers therefore need to approach platform content with caution. Our quality control was to prefer peer reviewed sources, unless (a) we had considerable trust in the authority of the authorship team (e.g., Bommasani et al., 2022; Phan et al., 2025), or (b) we wished to reflect the views of commercial AI developers, whose work typically is not peer reviewed (e.g., Achiam et al., 2023; Bubeck et al., 2023; Liu et al., 2024; Mu et al., 2024; Radford et al., 2018; Touvron et al., 2023).

References

Abdüsselam, M. S. (2023). Qualitative data analysis in the age of artificial general intelligence. *International Journal of Advanced Natural Sciences and Engineering Researches*, *7*(4), 1–5. https://doi.org/(https://doi.org/10.59287/ijanser.2023.7.4.454

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H.,

Bavarian, M., Belgum, J., & Bello, I., …, B. Zoph (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774. https://arxiv.org/abs/2303.08774v6.

Alvarado, R. (2023). *Simulating science: Computer simulations as scientific instruments (*Vol. 479*)*. Springer Nature.

Anis, S., & French, J. A. (2023). Efficient, explicatory, and equitable: Why qualitative researchers should embrace AI, but cautiously. *Business & Society*, *62*(6), 1139–1144. https://doi.org/10.1177/00076503231163286

Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Baird, D., & Faust, T. (1990). Scientific instruments, scientific progress and the cyclotron. *The British Journal for the Philosophy of Science*, *41*(2), 147–175.

Bechky, B. A., & Davis, G. F. (2025). Resisting the algorithmic management of science: Craft and community after generative AI. *Administrative Science Quarterly*, *70*(1), 1–22.

Bender, E. M. (2024). Resisting dehumanization in the age of "AI". *Current Directions in Psychological Science*, *33*(2), 114–120.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2022). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Biddle, J. B., & Kukla, R. (2017). The geography of epistemic risk. In K. C. Elliot & T. Richards (Eds.), *Exploring inductive risk: Case studies of values in science* (pp. 215–237). Oxford University Press.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., & Liang, P. (2022). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. http://arxiv.org/abs/2108.07258

Brinkmann, S., Jacobsen, M. H., & Kristiansen, S. (2014). Historical overview of qualitative research in the social sciences. In P. Leavy (Ed.), *The Oxford handbook of qualitative research* (pp. 17–42). Oxford University Press.

Broussard, M. (2018). *Artificial unintelligence: How computers misunderstand the world*. MIT Press.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712. http://arxiv.org/abs/2303.12712

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), Article 2053951715622512. https://doi.org/10.1177/2053951715622512

Casper, M. J., & Clarke, A. E. (1998). Making the pap smear into the 'right tool' for the job: Cervical cancer screening in the USA, circa 1940–95. *Social Studies of Science*, *28*(2), 255–290.

Chen, L., Zaharia, M., & Zou, J. (2024). How is ChatGPT's behavior changing over time? *Harvard Data Science Review*, *6*(2). https://doi.org/10.1162/99608f92.5317da47

Cheng, J., Marone, M., Weller, O., Lawrie, D., Khashabi, D., & Van Durme, B. (2024). Dated data: Tracing knowledge cutoffs in large language models. arXiv preprint arXiv:2403.12958. https://arxiv.org/abs/2403.12958v2

Chopra, , F., & Haaland, I. (2023). Conducting qualitative interviews with AI. *SSRN Scholarly Paper No. 4583756. Social Science Research Network*. https://doi.org/10.2139/ssrn.4583756

Chubb, L. A. (2023). Me and the machines: Possibilities and pitfalls of using artificial intelligence for qualitative data analysis. *International Journal of Qualitative Methods*, *22*, Article 16094069231193593. https://doi.org/10.1177/16094069231193593

Clarke, A. E., & Fujimura, J. H. (eds) (1992). *The right tools for the job: At work in twentieth-century life sciences*. Princeton University Press.

Croissant, J. L. (2022). Science and instrumentation. In T. W. Kneeland (Ed.), *The Routledge history of American science* (pp. 173–181). Routledge.

Dam, S. K., Hong, C. S., Qiao, Y., & Zhang, C. (2024). A complete survey on LLM-based AI chatbots. arXiv.Org. preprint arXiv:2406.16937. https://doi.org/10.48550/arXiv.2406.16937

Davidson, J., & di Gregorio, S. (2011). Qualitative research, technology, and global change. In *Qualitative inquiry and global crises* (pp. 79–96). Routledge.

De Paoli, S. (2024). Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review*, *42*(4), 997–1019. https://doi.org/10.1177/08944393231220483

Dhingra, B., Cole, J. R., Eisenschlos, J. M., Gillick, D., Eisenstein, J., & Cohen, W. W. (2022). Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, *10*, 257–273. https://doi.org/10.1162/tacl_a_00459

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, *27*(7), 597–600.

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., & Choi, Y. (2023). Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, *36*, 70293–70332.

Eloundou, T., Beutel, A., Robinson, D. G., Gu, K., Brakman, A. L., Mishkin, P., & Kalai, A. T. (2024). First-person fairness in chatbots. *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=TlAdgeoDTo

Flick, U. (ed.) (2014). *The SAGE handbook of qualitative data analysis*. Sage Publications.

Friese, S. (2024). Prompts for qualitative research. *Qualitative Insights Hub*. https://community.qeludra.com/spaces/13996487

Friese, S. (2025). Conversational Analysis with AI - CA to the Power of AI: Rethinking Coding in Qualitative Analysis. SSRN Scholarly Paper No. 5232579. *Social Science Research Network*. https://doi.org/10.2139/ssrn.5232579

Gamieldien, Y., Case, J., & Katz, A. (2023). Advancing qualitative analysis: An exploration of the potential of Generative AI and NLP in thematic coding. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4487768

Gao, J., Guo, Y., Li, T. J., Perrault, J., & T, S. (2023). Collabcoder: A GPT-powered workFlow for collaborative qualitative analysis. *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, 354–357. https://doi.org/10.1145/3584931.3607500

Garcia Quevedo, D., Glaser, A., & Verzat, C. (2025). Enhancing Theorization Using Artificial Intelligence: Leveraging Large Language Models for Qualitative Analysis of Online Data. *Organizational Research Methods*, 1–21. https://doi.org/10.1177/10944281251339144

Gatrell, C., Muzio, D., Post, C., & Wickert, C. (2024). Here, there and everywhere: On the responsible use of artificial intelligence (AI) in management research and the peer-review process. *Journal of Management Studies*, *61*(3), 739–751. https://doi.org/10.1111/joms.v61.3

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30), Article e2305016120. https://doi.org/10.1073/pnas.2305016120

Gläser, J., & Laudel, G. (2013). Life with and without coding: Two methods for early-stage data analysis in qualitative research aiming at causal explanations. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *14*(2), Article 2. https://doi.org/10.17169/fqs-14.2.1886

Goyanes, M., Lopezosa, C, & Jordá, B (2025) Thematic analysis of interview data with ChatGPT: Designing and testing a reliable research protocol for qualitative research. *Quality & Quantity*, 1–20. https://doi.org/10.1007/s11135-025-02199-3

Gozalo-Brizuela, R., & Garrido-Merchán, E. C. (2023). *A survey of generative AI applications*. arXiv preprint arXiv:2306.02781. https://doi.org/10.48550/arXiv.2306.02781

Grimes, M., von Krogh, G., Feuerriegel, S., Rink, F., & Gruber, M. (2023). From scarcity to abundance: Scholars and scholarship in an age of generative artificial intelligence. *Academy of Management Journal*, *66*(6), 1617–1624. https://doi.org/10.5465/amj.2023.4006

Harding, J., D'Alessandro, W., Laskowski, N. G., & Long, R. (2024). AI language models cannot replace human research participants. *AI & Society*, *39*(5), 2603–2605.

Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T. B., Dhariwal, P., Gray, S., Hallacy, C., Mann, B., Radford, A., Ramesh, A., Ryder, N., Ziegler, D. M., Schulman, J., Amodei, D., & McCandlish, S. (2020). Scaling laws for autoregressive generative modeling. arXiv preprint arXiv:2010.14701. https://arxiv.org/abs/2010.14701v2

Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology*, *26*(2), 38.

Hine, C. (2006). Databases as scientific instruments and their role in the ordering of scientific work. *Social Studies of Science*, *36*(2), 269–298.

Hitch, D. (2024). Artificial intelligence augmented qualitative analysis: The way of the future? *Qualitative Health Research*, *34*(7), 595–606. https://doi.org/10.1177/10497323231217392

Hou, C., Zhu, G., Zheng, J., Zhang, L., Huang, X., Zhong, T., & Ker, C. L. (2024). Prompt-based and fine-tuned GPT models for context-dependent and-independent deductive coding in social annotation. *Proceedings of the 14th Learning Analytics and Knowledge Conference*, 518–528.

Huang, J., Yang, D. M., Rong, R., Nezafati, K., Treager, C., Chi, Z., Wang, S., Cheng, X., Guo, Y., Klesse, L. J., Xiao, G., Peterson, E. D., Zhan, X., & Xie, Y. (2024). A critical assessment of using ChatGPT for extracting structured data from clinical notes. *NPJ Digital Medicine*, *7*(1), 106. https://doi.org/10.1038/s41746-024-01079-8

Huang, Y., Arora, C., Houng, W. C., Kanij, T., Madulgalla, A., & Grundy, J. (2025). Ethical concerns of Generative AI and mitigation strategies: A systematic mapping Study. arXiv preprint arXiv:2502.00015. https://doi.org/10.48550/arXiv.2502.00015

Jones, R. H. (2021). Data collection and transcription in discourse analysis: A technological history. In K. Hyland, B. Paltridge, & L. C. Lillian (Eds.), *The Bloomsbury handbook of discourse analysis* (pp. 9–20). Bloomsbury Academic.

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). Challenges and applications of large language models. arXiv preprint arXiv:2307.10169. http://arxiv.org/abs/2307.10169

Kalai, A. T., Nachum, O., & Zhang, E. (2025). Why language models Hallucinate. arXiv preprint arXiv:2509.04664. https://doi.org/10.48550/arXiv.2509.04664.

Karjus, A. (2025). Machine-assisted quantitizing designs: Augmenting humanities and social sciences with artificial intelligence. *Humanities and Social Sciences Communications*, *12*(1), 1–18. https://doi.org/10.1057/s41599-025-04503-w

Khademi, A. (2023). Can ChatGPT and Bard generate aligned assessment items? A reliability analysis against human performance. *Journal of Applied Learning and Teaching*, *6*(1), 75–80. https://doi.org/10.37074/jalt.2023.6.1.28

Koehler, M., & Sauermann, H. (2024). Algorithmic management in scientific research. *Research Policy*, *53*(4), Article 104985. https://doi.org/10.1016/j.respol.2024.104985

Köhler, T., Smith, A., & Bhakoo, V. (2022). Templates in qualitative research methods: Origins, limitations, and new directions. *Organizational Research Methods*, *25*(2), 183–210.

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, *121*(45), Article e2405460121.

Kulkarni, M., Mantere, S., Vaara, E., van den Broek, E., Pachidi, S., Glaser, V. L., Gehman, J., Petriglieri, G., Lindebaum, D., Cameron, L. D., Rahman, H. A., Islam, G., & Greenwood, M. (2024). The future of research in an artificial intelligence-driven world. *Journal of Management Inquiry*, *33*(3), 207–229. https://doi.org/10.1177/10564926231219622

LeCun, Y. (2022). A path towards autonomous machine intelligence version 0.9.2, 2022-06-27. *Open Review*, *62*(1), 1–62.

Lee, R. M. (2004). Recording technologies and the interview in sociology, 1920–2000. *Sociology*, *38*(5), 869–889.

Lee, V. V., Van Der Lubbe, S. C. C., Goh, L. H., & Valderas, J. M. (2024). Harnessing ChatGPT for thematic analysis: Are we ready? *Journal of Medical Internet Research*, *26*, Article e54974. https://doi.org/10.2196/54974

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Lieder, F. R., & Schäffer, B. (2024). *Reconstructive social research prompting (RSRP). Distributed interpretation between AI and researchers in qualitative research*. SocArXiv. https://doi.org/10.31235/osf.io/d6e9m

Lin, C.-C., Jaech, A., Li, X., Gormley, M. R., & Eisner, J. (2021). Limitations of autoregressive models and their alternatives. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5147–5173. https://doi.org/10.18653/v1/2021.naacl-main.405

Lincoln, Y.S., & Guba, E.G. (1985). *Naturalistic inquiry.* Sage.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., & Piao, Y. (2024). Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437. https://doi.org/10.48550/arXiv.2412.19437

Lixandru, D. (2024). The use of artificial intelligence for qualitative data analysis: ChatGPT. *Informatica Economica*, *28*(1), 57–67. https://doi.org/10.24818/issn14531305/28.1.2024.05

Luccioni, S., Jernite, Y., & Strubell, E. (2024, June). Power hungry processing: Watts driving the cost of AI deployment? *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 85–99.

Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., & Zanella-Béguelin, S. (2023). Analyzing leakage of personally identifiable information in language models. IEEE Symposium on Security and Privacy (SP), 346–363. https://doi.org/10.1109/SP46215.2023.10179300

Marcus, G. F. (2018). Deep learning: A critical appraisal. arXiv preprint arXiv:1801.00631. https://doi.org/10.48550/arXiv.1801.00631

Marcus, G. F. (2024). *Taming Silicon Valley: How we can ensure that AI works for us*. MIT Press.

Maxwell, J. (1992). Understanding and validity in qualitative research. *Harvard Educational Review*, *62*(3), 279–301.

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1906). Association for Computational Linguistics.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, *121*(41), Article e2322420121. https://doi.org/10.1073/pnas.2322420121

Mees-Buss, J., Welch, C., & Piekkari, R. (2022). From templates to heuristics: How and why to move beyond the Gioia methodology. *Organizational Research Methods*, *25*(2), 405–429.

Merton, R. (1973). The normative structure of science. In R. Merton (Ed.), *The sociology of science: Theoretical and empirical investigations* (pp. 267–278). University of Chicago Press.

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, *627*(8002), 49–58. https://doi.org/10.1038/s41586-024-07146-0

Mikami, K. (2015). Adoptable packages and the cost of their adoption: The craftwork of making the right cells for regenerative medicine in Japan. *New Genetics and Society*, *34*(4), 377–397.

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, *120*(13), Article e2215907120.

Morgan, D. L. (2023). Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. *International Journal of Qualitative Methods*, *22*, Article 1609406 9231211248.

Morris, D. (2023). Magical thinking and the test of humanity: We have seen the danger of AI and it is us. *AI & Society*, *39*(6), 3047–3049. https://doi.org/10.1007/s00146-023-01775-1

Mu, T., Helyar, A., Heidecke, J., Achiam, J., Vallone, A., Kivlichan, I., Lin, M., & Beutel, A. (2024). Rule based rewards for language model safety. OpenAI: https://cdn.openai.com/rule-based-rewards-for-language-model-safety.pdf

Narayanan, A., & Kapoor, S. (2024). *AI snake oil: What artificial intelligence can do, what it can't, and how to tell the difference*. Princeton University Press.

Nguyen-Trung, K. (2025). ChatGPT in thematic analysis: Can AI become a research assistant in qualitative research? *Quality & Quantity*, 1–34. https://doi.org/10.1007/s11135-025-02165-z

Ni, B., Liu, Z., Wang, L., Lei, Y., Zhao, Y., Cheng, X., & Derr, T. (2025). Towards trustworthy retrieval augmented generation for large language models: A survey. arXiv preprint arXiv:2502.06872. https://doi.org/10.48550/arXiv.2502.06872

OpenAI. (2025). GPT-5 System Card. https://cdn.openai.com/pdf/8124a3ce-ab78-4f06-96eb-49ea29ffb52f/gpt5-system-card-aug7.pdf

Peterson, A. J. (2025). AI and the problem of knowledge collapse. *AI & Society*, *40*(5), 3249–3269. https://doi.org/10.1007/s00146-024-02173-x

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., & Schut, L. (2025). *Humanity's last exam*. arXiv preprint arXiv:2501.14249. https://doi.org/10.48550/arXiv.2501.14249

Placani, A. (2024). Anthropomorphism in AI: Hype and fallacy. *AI and Ethics*, *4*(3), 691–698. https://doi.org/10.1007/s43681-024-00419-4

Qi, X., Zeng, Y., Xie, T., Chen, P. Y., Jia, R., Mittal, P., & Henderson, P. (2023). Fine-tuning aligned language models compromises safety, even when users do not intend to!. *The Twelfth International Conference on Learning Representations*.

Quirkos. (2025). Outsourcing decision making: AI, ethics, and qualitative research. https://www.quirkos.com/blog/post/outsourcing-decision-making-ai-ethics-and-qualitative-research/

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI Technical Report. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Raza, S., Qureshi, R., Zahid, A., Fioresi, J., Sadak, F., Saeed, M., & Shoman, M. (2025). *Who is responsible? The data, models, users or regulations? Responsible Generative AI for a sustainable future*. arXiv preprint arXiv:2502.08650. https://doi.org/10.48550/arXiv.2502.08650

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. arXiv preprint arXiv:2402.07927. https://doi.org/10.48550/arXiv.2402.07927

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., & Resnik, P. (2024). The prompt report: A systematic survey of prompting techniques. arXiv preprint arXiv:2406.06608. https://doi.org/10.48550/arXiv.2406.06608

Sen, M., Sen, S. N., & Sahin, T. G. (2023). A new era for data analysis in qualitative research: ChatGPT!. *Shanlax International Journal of Education*, *11*, 1–15.

Shapin, S., & Schaffer, S. (1985). *Leviathan and the air-pump: Hobbes, Boyle, and the experimental life*. Princeton University Press.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI Models collapse when trained on recursively generated data. *Nature*, *631*(8022), 755–759. https://doi.org/10.1038/s41586-024-07566-y

Silver, C., & Lewins, A. (2014). *Using software in qualitative research: A step-by-step guide* (2nd edn). Sage Publications Ltd.

Tabone, W., & de Winter, J. (2023). Using ChatGPT for human–computer interaction research: A primer. *Royal Society Open Science*, *10*(9), Article 231053. https://doi.org/10.1098/rsos.231053

Theelen, H., Vreuls, J., & Rutten, J. (2024). Doing research with help from ChatGPT: Promising examples for coding and inter-rater reliability. *International Journal of Technology in Education*, *7*(1), 1–18.

Thornton, I. (2023). A special delivery by a fork: Where does artificial intelligence come from? *New Directions for Evaluation*, 23–32. https://doi.org/10.1002/ev.20560

Törnberg, P. (2024). Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, Article 08944393241286471. https://doi.org/10.1177/08944393241286471

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2405.08828. https://arxiv.org/abs/2302.13971v1

Tracy, S. J. (2012). *Qualitative research methods: Collecting evidence, crafting analysis, communicating impact.* John Wiley & Sons.

Turobov, A., Coyle, D., & Harding, V. (2024). Using ChatGPT for thematic analysis. arXiv. arXiv:2405.08828 http://arxiv.org/abs/2405.08828

Van Rooij, I., Guest, O., Adolfi, F., de Haan, R., Kolokolova, A., & Rich, P. (2024). Reclaiming AI as a theoretical tool for cognitive science. *Computational Brain & Behavior*, 1–21.

Vassel, F. M., Shieh, E., Sugimoto, C. R., & Monroe-White, T. (2024). The psychosocial impacts of generative AI harms. *Proceedings of the AAAI Symposium Series*, *3*(1), 440–447.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Walsh, S., & Pallas-Brink, J. (2023). The ethnographer in the machine: Everyday experiences with AI-enabled data analysis. *EPIC Proceedings*, 538–554.

Walsh, T. (2023). *Faking it: Artificial intelligence in a human world*. FLINT.

Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large language models cannot replace human participants because they cannot portray identity groups. *Nature Machine Intelligence*, *7*(3), 100–411. https://doi.org/10.1038/s42256-025-00986-z

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., Anandkumar, A., Bergen, K., Gomes, C. P., Ho, S., Kohli, P., Lasenby, J., Leskovec, J., Liu, T.-Y., Manrai, A., & Zitnik, M. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, *620*(7972), 47–60. https://doi.org/10.1038/s41586-023-06221-2

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022a). Emergent abilities of large language models. arXiv preprint arXiv:2206.07682. https://arxiv.org/abs/2206.07682v2

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022b). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

Wheeler, K. (2025). *How to use generative AI to assist the analysis of qualitative data*. *Sage Research Methods How to Guides*.

Widder, D. G., West, S., & Whittaker, M. (2023). Open (for business): Big Tech, concentrated power, and the political economy of Open AI (SSRN Scholarly Paper 4543807). https://doi.org/10.2139/ssrn.4543807

Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., & Oudeyer, P.-Y. (2023). Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. *Companion*

Proceedings of the 28th International Conference on Intelligent User Interfaces, 75–78. https://doi.org/10.1145/3581754.3584136

Yanow, D., & Schwartz-Shea, P. (2014). *Interpretation and method: Empirical research methods and the interpretive turn* (2nd ed). Routledge.

Yin, R. (2014). *Case study research: Design and methods* (5th ed.). Sage Publications, Inc.

Zhang, H., Wu, C., Xie, J., Lyu, Y., Cai, J., & Carroll, J. M. (2023). Redefining qualitative analysis in the AI era: Utilizing ChatGPT for efficient thematic analysis. arXiv preprint arXiv:2309.10771. https://arxiv.org/abs/2309.10771v1

Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., & Cui, B. (2024). Retrieval-augmented generation for AI-generated content: A survey. arXiv preprint arXiv:2402.19473. https://doi.org/10.48550/arXiv.2402.19473

Author Biographies

Duc Cuong Nguyen is a Lecturer in International Business at the Alliance Manchester Business School, the University of Manchester, United Kingdom. Duc obtained his PhD in Business at the University of Sydney, Australia in 2023. His research interests include qualitative research methodologies and cross-sector social partnerships.

Catherine Welch is the chair of Strategic Management at Trinity College Dublin, Ireland. She is also a distinguished visiting professor at Aalto University, Finland. Catherine's research has concentrated on approaches to context in international business research, particularly the use of qualitative research methodology and process approaches to studying firm internationalization. Her work has appeared in leading journals in international business and management. She has a track record of launching new disciplinary conversations and advocating methodological and theoretical pluralism. Catherine is the current Book Review Editor of the Journal of International Business Studies and is a member of the journal's Research Methods Advisory Committee 2023–2025. She is an associate editor of Organizational Research Methods. She was a founding member of the Academy of International Business (AIB) Research Methods Shared Interest Group (RM-SIG). From 2022 to 2025, she served on the AIB's board as Vice President of Programs.