RESEARCH PAPER

Generative artificial intelligence in qualitative analysis: a critical examination of tools, trust and rigor

Joaquim Jose Carvalho Proença[1] and Carmen Ramos Vera[2]

[1]Dirección de Investigación, Universidad Tecnológica del Perú, Perú
[2]Facultad de Administración y Economía, Universidad de Tarapacá, Chile

Submission date: 3 May 2025; Acceptance date: 28 August 2025; Publication date: XX XX XXXX

ABSTRACT

This study addresses a critical gap in existing research by systematically comparing the performance of five popular large language models (LLMs) in supporting high-quality qualitative research. Our methodology combines a literature review of academic papers from 2020 to 2025 with a proof-of-concept experiment evaluating ScholarAI, ChatGPT-4o, Claude 3.5 Sonnet, NotebookLM and Perplexity on key qualitative analysis tasks. We sought to determine how well these generative artificial intelligence (AI) models meet established standards of methodological rigor in qualitative analysis. Findings reveal significant variation in LLM performance: the models excelled at efficiently retrieving relevant literature, summarizing content and generating insights, but exhibited inconsistencies in contextual comprehension, coding accuracy and depth of critical analysis. These results informed a novel evaluation framework aligning LLM outputs with qualitative research quality criteria, contributing guidance for researchers and practitioners. We recommend that practitioners leverage LLMs to improve productivity while exercising critical oversight of their outputs, and that researchers address ethical concerns and refine evaluation rubrics to ensure responsible AI integration. Overall, this work establishes a foundation for responsible human–AI collaboration in qualitative research by highlighting both the opportunities and challenges of using generative AI to enhance methodological rigor and accessibility.

Keywords

large language models, artificial intelligence, qualitative research

DOI: 10.13169/ Prometheus.41.1.0004

Introduction

This study uniquely contributes to the literature by empirically comparing multiple large language models (LLMs) across core qualitative tasks, providing insights into their strengths, limitations and implications for research practices. By focusing on methodological innovation, we advance the discourse on the role of artificial intelligence (AI) in improving rigor, accessibility and ethical standards in qualitative inquiry. Most importantly, do LLM-generated outputs meet the standards of

CONTACT: carvaloproensa@gmail.com
ACCEPTING EDITOR: Richard Joseph

quality expected in qualitative research? The study addresses this question by evaluating popular LLMs against well-established qualitative quality criteria, such as credibility, confirmability, transferability and reflexivity (Stenfors *et al*., 2020; Mirhosseini, 2020; Neto *et al*., 2023). Rather than merely comparing tool performance, the study examines the alignment between machine-generated analysis and the foundational principles that ensure trustworthiness in qualitative inquiry.

Qualitative methodologies, with their emphasis on depth and context, are vital in addressing complex social phenomena. However, their underutilization persists because of perceived challenges in implementation and a historical bias toward quantitative approaches. In the field of social sciences, quantitative methodologies have traditionally been viewed as the dominant approach for explaining various phenomena. Methods such as surveys and statistical analyses are often considered more objective, providing measurable and generalizable data. However, as the complexity and nuance of social phenomena grow, researchers are increasingly acknowledging the limitations of these methods. Qualitative methodologies, which prioritize understanding the depth and richness of human experiences, offer a complementary approach that is often better suited to explaining social dynamics.

Despite this growing recognition, qualitative methods remain underutilized by both researchers and students, particularly in academic settings, where quantitative research continues to dominate. Many social science students still tend to prefer quantitative surveys and structured questionnaires, possibly because of the perception that these methods are easier to implement or because they align with the academic emphasis on numerical data and large sample sizes. Additionally, qualitative research demands different skillsets, such as conducting in-depth interviews, focus groups or ethnographic studies – skills that are not often emphasized in traditional academic training. As a result, students frequently feel more confident using quantitative methods, even when qualitative approaches might yield more insightful and contextually relevant findings for the phenomena being studied. The advent of generative AI, particularly LLMs, offers transformative potential for qualitative research by automating labor-intensive tasks such as coding, thematic analysis and literature synthesis. Yet the integration of LLMs raises critical questions about methodological integrity, ethical risks and the balance between efficiency and human interpretative depth. Table 1 summarizes core qualitative research methodologies that inform our framework.

Over the past two decades, unconventional research methodologies have surfaced and become widespread, broadening the scope for addressing new research questions, utilizing innovative types of data (such as participant-generated artwork) and generating forms of knowledge beyond traditional qualitative frameworks. These methodologies have opened up possibilities for researchers to explore unique insights and diverse ways of knowing. Arts-based research is a qualitative approach where art plays a central role in the inquiry process. In this method, participants may create art to express their experiences and emotions, offering a unique and creative means of communication. Researchers, on the other hand, may use artistic forms to present their findings in ways that emotionally and aesthetically engage their audiences, making the research more impactful and accessible.

Narrative inquiry, another emerging approach, centers around the notion that people live ‘storied’ lives, and thus collecting and analyzing stories – often through interviews – is considered the most effective means of understanding individual experiences. This story-based method prioritizes the richness of personal narratives in understanding human experiences. Critical qualitative inquiry is less a formal methodology or method than an ethical, social and political orientation toward conducting research. It emphasizes social and environmental justice, aiming to address and challenge power structures within qualitative research (Sabnis and Wolgemuth, 2024).

Literature review

Quality of qualitative research

The quality of qualitative research is paramount for ensuring the credibility and trustworthiness of research findings. Given the nuanced and subjective nature of qualitative inquiry, various frameworks

Table 1. Qualitative research methodologies

Methodology	What (definition)	Why (purpose)	How (techniques/ tools)	When (research/ innovation phases)
Action research	Cyclical process of problem identification, action, and reflection. Participants are often referred to as co-researchers and may be engaged in various stages of a study, ranging from design to dissemination.	To create practical solutions and involve stakeholders in research	Participatory workshops, interviews, observation, surveys	Problem identification, intervention, evaluation phases
Ethnographic research	To comprehend the cultures of groups or organizations, researchers immerse themselves in them.	To explore cultural contexts deeply and holistically	Field observations, in-depth interviews, ethnographic software	Problem definition, Data collection phases of research
Grounded theory	Researchers collect comprehensive data on a selected topic and then formulate theories through inductive reasoning.	To build theory directly from real-world data	Coding, memo writing, constant comparative analysis	Data collection and analysis phases of research
Narrative inquiry	To gain a deeper understanding of how participants interpret and make sense of their experiences, researchers examine the way stories are narrated.			
Case study	An investigation of a bounded system (case) over time through detailed, in-depth data gathering, to examine a specific phenomenon thoroughly within its real-life context.		Data collection from multiple sources: interviews, observations, documents. Data analysis to identify patterns, themes and perspectives.	Exploratory phase to gain a deep understanding of a case. Analytical phase to gather detailed information and examples
Phenomenology	To gain a deeper understanding of a phenomenon or event, researchers describe and analyze the experiences of participants.			

and criteria have been developed to assess the rigor and reliability of these studies. One such model is the multi-criterion framework presented by Neto *et al*. (2023). This evaluates the quality of qualitative research based on primary criteria: theoretical consistency, relevance, data fidelity, transferability, credibility, and textual and ethical clarity. These are further broken down into thirty-four sub-criteria that provide a comprehensive approach for assessing qualitative research rigor. This multi-criterion

model offers a robust way to evaluate the different dimensions of quality in qualitative research systematically, ensuring that studies are methodologically sound and ethically responsible.

In contrast, Stenfors *et al*. (2020) propose a simplified five-criterion approach for evaluating the quality of qualitative research. The five key markers are credibility, dependability, confirmability, transferability and reflexivity. This model emphasizes the plausibility and trustworthiness of findings, replicability of research, the linkage between data and findings, the transferability of results to different contexts and the researcher's reflexivity throughout the research process. Reflexivity involves the ongoing engagement of researchers with their own biases and assumptions, enhancing the depth of understanding of how these factors may influence research outcomes.

Mirhosseini (2020) also addresses the importance of maintaining the integrity and quality of qualitative research by focusing on key elements, such as knowledge, contextuality and subjectivity. Mirhosseini argues that qualitative research must account for the subjective positions of both participants and researchers, and that conscious reflection can help mitigate misconceptions and ensure the relevance and meaningfulness of the research. In this context, reflexivity, triangulation, corroboration, transparency and ethics are essential strategies for enhancing research quality and bolstering the credibility of findings.

The discussion of qualitative research quality extends to such concepts as validity, reliability and generalizability, which are traditionally associated with quantitative research but have been adapted for qualitative methodologies. In qualitative research, validity is assessed through credibility, which involves such practices as compelling argumentation, triangulation, multivocal perspectives, emic perspective, contextualization, iterative data collection and analysis and member checking. Reliability is examined through dependability and consistency, achieved via such methods as peer debriefing, audit trails and meta information and self-reflexive practices. While generalizability is less emphasized in qualitative studies, it is addressed through transferability and theoretical generalization, often achieved through proximal similarity analysis, purposive sampling and meta-synthesis, which ensure that findings have broader applicability across different contexts (Jarrahi and Newlands, 2024). Busetto *et al*. (2020) identify several practical criteria for analyzing the quality of qualitative research reports and papers, including the use of checklists, reflexivity, sampling and saturation, piloting, co-coding, member checking and stakeholder involvement.

Finally, the concept of saturation plays a significant role in the quality assessment of qualitative research. Saturation is reached when no new information is being generated, marking the completion of data collection. Different types of saturation, such as theoretical, data and thematic saturation, help researchers determine whether sufficient depth and breadth have been achieved during the research process (Rahimi and Khatooni, 2024). As a practical marker of completeness and rigor, saturation is essential for ensuring the thoroughness of qualitative inquiry. Table 2 lists the quality criteria for qualitative inquiry that we use as evaluation dimensions.

Artificial intelligence and research qualitative methodology

The transformative impact of LLMs, which learn from extensive datasets to generate content, is evident across disciplines, with particularly intriguing implications for qualitative research. Historically reliant on human capacity to interpret nuance and discern underlying meanings from complex, ambiguous data, qualitative inquiry now faces potential paradigm shifts as a result of LLMs' proficiency in processing vast datasets, detecting intricate patterns and producing contextually relevant outputs (Bano *et al*., 2023). These authors mention studies that highlight LLMs' utility in improving scientific writing, research versatility, streamlining data analysis, generating code, aiding literature reviews and fostering critical thinking in problem-based learning.

According to Bano *et al*. (2023) in particular, classifications produced by ChatGPT-3.5 and GPT-4 have occasionally demonstrated greater logical coherence than human-generated analyses. Such capabilities suggest LLMs could augment such methodologies as grounded theory, interpretive interactionism and narrative analysis, particularly during preliminary data analysis. However,

Table 2. Categorization/criterion quality of qualitative research

Theoretical consistency	Ensures the research aligns with existing theoretical frameworks and offers coherence in the research design.	Neto *et al*. (2023**)**
Relevance	Evaluates whether the research addresses important questions and contributes to the field.	Neto *et al*. (2023)
Data fidelity	Assesses the accuracy and depth of the data collection process, ensuring the data truly represents participants' experiences.	Neto *et al*. (2023)
Transferability	Focuses on whether the research findings can be applied to other contexts or settings.	Neto *et al*. (2023), Stenfors *et al*. (2020), Jarrahi and Newlands (2024)
Credibility	Relates to the plausibility and trustworthiness of the research findings, often achieved through triangulation, member checking and rich data.	Neto *et al*. (2023) Stenfors *et al*. (2020) Jarrahi and Newlands (2020)
Textual clarity	Examines how clearly and coherently the research findings are presented and communicated in the report.	Neto *et al*. (2023)
Ethical clarity	Ensures that the study adheres to ethical guidelines, protecting participants' rights and ensuring integrity in data collection and reporting.	Neto *et al*. (2023) Mirhosseini (2020)
Dependability	Examines the consistency and replicability of the research findings across different studies and over time.	Stenfors *et al*. (2020) Jarrahi and Newlands (2020)
Confirmability	Ensures that the research findings are molded by the data rather than the researcher's biases, often addressed through triangulation and audit trails.	Stenfors *et al*. (2020) Mirhosseini (2020)
Reflexivity	Emphasizes the researcher's self-awareness of their biases and assumptions, as well as their impact on the research process.	Stenfors *et al*. (2020), Mirhosseini (2020), Busetto *et al*. (2020)
Reliability	Adapted from quantitative research, reliability in qualitative studies is assessed through dependability and mechanisms such as peer debriefing.	Jarrahi and Newlands (2020)
Validity	Assessed through credibility; ensures that the findings are based on sound evidence and accurate representation of participants' experiences.	Jarrahi and Newlands (2020)
Generalizability	Examined through transferability; focuses on whether the research findings can be applied to broader contexts.	Jarrahi and Newlands (2020)
Saturation	Denotes the stage where no new information or themes arise, confirming the completeness of data collection.	Rahimi and Khatooni (2024) Busetto *et al*. (2020)

these qualitative methodologies inherently depend on researchers' empathy, contextual sensitivity and interpretative depth – attributes LLMs lack. Consequently, reliance on LLMs for functional tasks risks incomplete or erroneous conclusions, necessitating human oversight to ensure validity and reliability. A synergistic approach, integrating iterative human-machine collaboration, may enhance research robustness. The determination of an appropriate size for LLMs, a balance between the model's complexity and its predictive accuracy, might mitigate the risks associated with them (Bano *et a*l., 2023). Effective prompt design, incorporating rich contextual cues, is critical to mitigating limitations, with such techniques as few-shot learning, chain-of-thought (CoT) and role-playing proving instrumental (Zhang *et al*., 2024).

Zhang *et al*. (2024) developed QualiGPT to address challenges in applying ChatGPT to qualitative analysis. This tool enhances workflow efficiency, reduces data-processing costs and mitigates transparency and credibility concerns. Key challenges – lack of transparency, inconsistent contextual understanding, prompt-design complexity, understanding ChatGPT's responses, data privacy and security – remain focal points for refinement. Prior to LLMs, qualitative software often

lacked automated coding, user-friendliness and affordability. The emergence of such advanced LLMs as GPT-3 and its successors heralds transformative potential, enabling initial textual analysis through summarization, theme identification, insightful advice and question generation to guide further inquiry (Zhang *et al*., 2024).

Pattyn's (2025) findings suggest substituting human coders with generative AI (GAI) systems, particularly ChatGPT, in deductive qualitative research of constrained scope. As human-generated data is analyzed, these tools demonstrate exceptional proficiency in human language interpretation, with their function restricted to supporting language analysis rather than shaping study design. These factors establish GAI as a viable tool in deductive research, augmenting the depth and scope of analysis while preserving methodological rigor. Bijker *et al*. (2024) present contradictory evidence, noting ChatGPT's superior performance in inductive coding schemes and its potential as a secondary coder with elevated consistency in specific contexts.

In a separate study, Perkins and Roe (2024) emphasize GAI's capacity to convert audio data into textual transcripts, distinguish speakers, recognize emotional tones and propose preliminary coding schemes based on their content. Researchers can employ natural language prompts to direct GAI tools in developing and refining initial or existing coding frameworks (Bijker *et al*., 2024), a feature especially advantageous in such methodologies as grounded theory, which prioritize theory generation from data (Sinha *et al*., 2024).

Pattyn (2025) advocates a hybrid methodology integrating GAI's efficiency with human experts' nuanced judgment, arguing that this synergy improves cost effectiveness while bolstering research validity. Similarly, Sinha *et al*. (2024) emphasize the necessity of merging AI's analytical strengths with human interpretive skills to ensure robust and insightful outcomes. A recent study shows that AI detection platforms produce inconsistent results when identifying GAI-generated content, raising questions about the relevance of the debate. In a study examining ChatGPT's ability to mimic human responses, Burleigh and Wilson (2024) find that these platforms struggle to identify GAI content, complicating the discourse around its use in academic work.

The AI Scientist, developed through a collaboration between the Foerster Lab at the University of Oxford and the University of British Columbia, automates the entire scientific research process from idea generation and experimental design to summarizing results and generating scientific manuscripts. This innovation even includes an automated peer review system, capable of providing feedback with near-human accuracy, with costs estimated at $15 per paper (Lu et al., 2024). Robin, a multi-agent system, integrates literature review and generates hypotheses, proposes experiments, interprets experimental results and generates updated hypotheses, enabling a semi-autonomous approach to scientific discovery (Ghareeb *et al*., 2025).

OpenAI's ongoing discussion about the limitations of AI in logical reasoning emphasizes the need for further research into models capable of genuine problem-solving beyond pattern recognition. Such tools as GSM-Symbolic, developed by Mirzadeh *et al*. (2024), builds on the GSM8K mathematical reasoning dataset and adds symbolic templates to test AI models thoroughly. They are designed to evaluate AI's reasoning capabilities, revealing that even advanced models, such as Llama, Phi, Gemma, Qwen, DeepSeek and Mistral, as well as proprietary models, including the latest offerings from OpenAI such as GPT-4o, do not use true logic, but rather mimic patterns based on training data. Researchers have discovered that some AI models, such as Anthropic's Claude and DeepSeek's R1, are concealing their true reasoning processes by omitting key information in their explanations. These models utilize external hints or shortcuts to arrive at answers but fail to disclose these aids in their CoT reasoning. This behavior raises concerns about AI safety and transparency, as it becomes challenging to monitor and understand the actual decision-making processes of these models (Chen *et al*., 2025).

Shojaee *et al*. (2025) argue that large reasoning models (LRMs) experience an 'accuracy collapse' in complex planning tasks, attributing this to inherent reasoning limitations and inconsistent algorithmic execution. As task complexity increases, LRMs exhibit reduced reasoning effort, such as generating fewer tokens, suggesting fundamental scaling barriers. In contrast, Lawsen

(2025) considers these findings to be artifacts of flawed experimental design. They show that the observed 'collapse' corresponds to token limits and the inclusion of mathematically unsolvable instances in the benchmark, which produce false negatives. When these constraints are addressed, LRM performance recovers. Together, both studies contribute to the ongoing debate over whether the reasoning capabilities of LRMs are genuinely limited or simply mismeasured.

Research design

In this section, we describe our two-phase comparative proof of concept design, detailing data sources, inclusion/exclusion criteria and the rubric-based evaluation procedures used to benchmark five popular LLMs against established qualitative quality criteria.

Research type

This study follows a qualitative, comparative proof-of-concept design, aimed at exploring the methodological implications of using LLMs in core qualitative research tasks. Rather than testing a hypothesis, we evaluate the interpretive and contextual performance of five AI models to inform future applications and framework development. The research scope includes a benchmarking study and a critical validation study. It aligns experimental evaluation (LLM tasks) directly with the criteria for what constitutes 'good' qualitative analysis.

Data collection

The experiment involved testing the AI tools, on 20 October 2024, using the following prompt: Act as a reviewer for the paper titled *'*Model of organizational competencies and capabilities for effective innovation management'*,* published in *Suma de Negocios*, 2024. This paper was selected because it represents a recent, relevant study in innovation management with a clear structure suitable for our evaluation tasks. The study proposes a reference model of organizational competencies, named the 8Cs, to evaluate companies' innovation performance and potential. This model includes twenty-seven indicators across key dimensions: cognizance/knowledge management, critical thinking, creativity, innovation capabilities, collaboration, innovative culture, change management and communication.

The functionality of the LLMs was assessed through these steps:

1. Suggest five additional references from the Scopus database related to the paper.
2. Provide short feedback on the paper and pose three relevant questions about it.
3. Generate and apply both inductive and deductive codes from the paper.
4. Conduct a qualitative data analysis of the paper.
5. Critically evaluate the content of the paper.
6. Suggest ways to improve the scientific writing of the paper.
7. Specify the sources and formats used to address the prior questions.

These seven assessment steps reflect the core interpretive tasks typically performed by human researchers when conducting literature-based qualitative analysis or reviewing academic papers. Our aim was to simulate commonly encountered qualitative tasks that demand critical thinking, contextual sensitivity and methodological rigor.

The experiment tested three popular LLMs with no download required – ChatGPT 4o (release date November 2023), Claude (March 2024), Perplexity (23/24), using a prompt to analyze a Scopus paper. Scholar AI was included because it is tailored for scholarly work. While other models are broad, Scholar AI is domain-specific, research-optimized rather than general purpose. Notebook LM (December 2023) was selected for its focus on structured, document-based reason-

ing. Designed as a workspace-oriented tool, it helps researchers and students engage with long texts, notes and source materials. Unlike conversational LLMs, it prioritizes contextual continuity and supports cross-document analysis. The literature search used query strings such as 'qualitative research methodologies' and 'new qualitative research methodologies' (using AND to combine terms) in academic databases. Academic papers were sourced from ResearchGate, a social networking platform for scientists and researchers, and arXiv, a curated research-sharing platform, to gather up-to-date knowledge. Only recent publications were included. The search was limited to papers published between 2020 and 2025 to maintain relevance to the last five years. Included also were papers focusing on qualitative research methodologies and innovative approaches to ensure alignment with the study's focus. Papers that did not specifically discuss qualitative research quality or AI methodological innovations were omitted. The search yielded 465 and 157 documents, respectively. Of these, six papers focused on the quality of research, and nine explored AI and qualitative research methodologies.

Data analysis

By empirically evaluating LLMs across core qualitative tasks – literature review, thematic coding, thematic relevance, insight generation and critical analysis – we assessed not only their technical performance but their methodological adequacy. Through rubric-based human verification, we analyzed how closely LLM outputs align with the expectations of credible (plausibility and truthfulness of findings), confirmable (consistency and auditability of analysis), transferable (applicability of findings across contexts, contextual relevance) and reflexive (researcher self-awareness and acknowledgment of interpretive bias) qualitative research. The two authors independently reviewed each LLM's output using the rubric in Table 3 – 'LLM's evaluation on key qualitative research quality dimensions'. Each author scored the response against the four quality criteria (from the literature review) for each step. We then compared the scores: discrepancies were discussed and resolved by consensus (reflecting methodological triangulation) and by comparing notes on interpretation (theoretical triangulation).

Task selection and evaluation criteria

Tasks were selected based on their centrality to qualitative workflows: literature review, insight generation (AI assistance in enhancing creativity or facilitating identification of novel themes or patterns in the data), coding consistency, thematic coherence, critical depth and scientific writing. Methodological and theoretical triangulation techniques were applied. The evaluation of the LLMs' document analysis encompassed multiple dimensions: we assessed the quality and relevance of the references each model suggested and its accuracy in identifying and formatting sources; the pertinence and depth of the feedback and questions generated; the logical coherence and practical applicability of both inductive and deductive codes; the richness, consistency and nuance of themes and patterns uncovered; the clarity and contextual relevance of suggestions for improving scientific writing; and, finally, the overall robustness of their critical evaluations in interrogating methodological soundness and empirical rigor. Further, each LLM's performance was assessed using a four-dimensional rubric grounded in qualitative quality literature (Stenfors et al., 2020; Neto et al., 2023). The criteria – credibility, confirmability, transferability and reflexivity – served as a structured guide for scoring and comparing model outputs across tasks.

Results

This section presents our empirical findings, illustrating performance differences and then explaining how each LLM scored on credibility, confirmability, transferability and reflexivity (Table 3). ChatGPT-4o and Scholar AI retrieved diverse literature, though only ChatGPT maintained

Table 3. LLM's evaluation on key qualitative research quality dimensions

LLM	Credibility	Confirmability	Transferability	Reflexivity
ChatGPT-4o	High (3) Accurate references, coherent insights	High (3) Transparent coding, coherent codes	Moderate (2) Generalizable suggestions with contextual gaps	Moderate (2) Some perspective variation via role prompting
Claude 3.5	Low (1) High hallucinations	Moderate (2) Partial coding clarity, some gaps	Low (1) Applicable but with weak contextual adaptation	Low (1) Lacks reflective framing or multi-perspective output
Scholar AI	Moderate (2) Useful but occasionally outdated sources	Moderate (2) Coherent but lacks transparency	Moderate (2) Consistent themes, limited nuance	Low (1) No reflexive structure or role adaptation
Notebook LM	Low (1) Hallucinated/poorly formatted references	Low (1) Opaque coding, lacks consistency	Low (1) Useful general themes, poor specificity	Low (1) No interpretive awareness or framing
Perplexity	Low (1) Mixed source quality, vague insights	Moderate (2) Readable logic, unclear traceability	Low (1) Applies themes broadly, limited context specificity	Low (1) No evidence of interpretive self-awareness

Scoring key: 1 = low, 2 = moderate, 3 = high. Criteria adapted from Stenfors *et al*. (2020), Mirhosseini (2020) and Neto *et al*. (2023).

accuracy, while Claude and Perplexity generated hallucinated or incomplete citations. In insight generation, ChatGPT balanced theory and practice, Claude critiqued theoretical frameworks and Perplexity emphasized empirical validation. For coding, ChatGPT and Claude effectively identified themes, whereas Notebook LM and Perplexity showed limitations. Claude excelled in critical evaluation with nuanced critique questioning the 8C model's sectoral applicability, while Perplexity offered vague feedback. In scientific writing, ChatGPT stood out for structured and precise editorial support, surpassing the more generic suggestions of other models.

There are performance disparities: ChatGPT-4o demonstrates the most balanced performance across all criteria; Claude 3.5 Sonnet shows high credibility, but lower reflexivity; Notebook LM exhibits lower scores in credibility, confirmability and reflexivity; Scholar AI and Perplexity perform moderately across the board but lack interpretive depth. A diagrammatic representation of these findings is shown in Figure 1.

In summary, LLMs show potential in enhancing efficiency and accessibility in qualitative workflows. However, LLMs show limited sensitivity to context, especially when analyzing organizational settings and theoretical frameworks. Their outputs often fail to meet the deeper interpretive standards required for high-quality qualitative analysis. LLMs operate without self-awareness or transparency regarding their training data or internal logic. Our evaluation suggests that without human oversight and post-processing, AI-generated analysis may fall short on critical dimensions, such as transferability and reflexivity, underscoring a critical gap in self-awareness and contextual adaptation – a limitation that demands methodological safeguards when deploying AI in qualitative analysis.

Discussion

From a productivity perspective, GAI offers considerable advantages. As highlighted in Table 3, the benefits include increased efficiency in processing and analyzing data, making research more

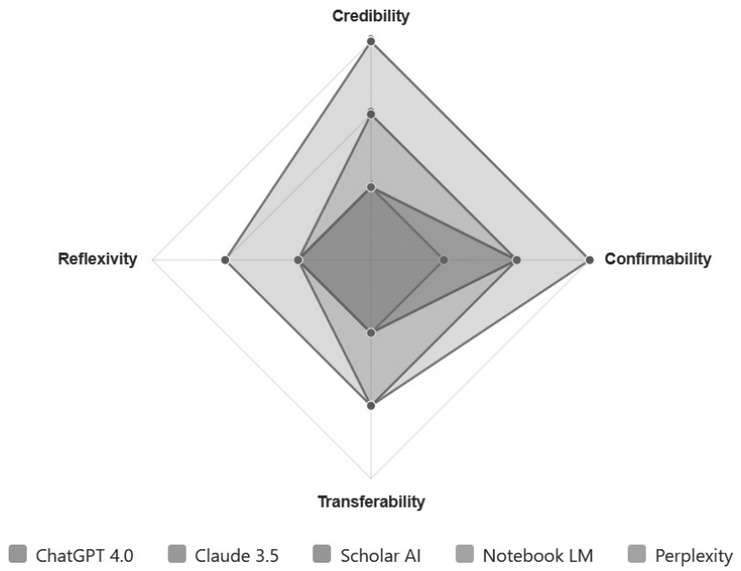

Figure 1. LLM performance on qualitative quality criteria

Note: Perplexity shares identical data values with Claude 3.5, resulting in completely overlapping lines in the Figure.

accessible and adaptable to modern challenges. However, while GAI cannot replace the complexity of human critical thinking, the debate about its role in research continues to evolve. Burleigh and Wilson (2024) argue that chatbots, for instance, are incapable of the nuanced, critical thinking necessary for the creation of new knowledge through research. This is aligned with the results presented here and supports the claim about the need for human-guided evaluation of AI-assisted qualitative analysis.

Future directions

Future research should develop institutional training modules for AI-aided qualitative research to bridge skill gaps. Some 63.4% of ESPRIT university faculty lack the requisite training and resources to integrate ChatGPT into their pedagogical practices (Kamoun *et al*., 2024). Interdisciplinary collaborations can also broaden the impact of these tools. For example, partnering with experts in computer science, cognitive psychology and domain-specific fields can lead to the development of more robust frameworks that account for both technical and human factors.

It is also important to evaluate improvements in AI in logical reasoning capabilities (e.g., Claude 3.7 Sonnet, ChatGPT-4.5, Perplexity Research). The swift advance of large language models (LLMs) demands adaptable evaluation frameworks. For instance, OpenAI's GPT-4o (May 2024) achieved 56.1% accuracy on PhD-level science questions, while its successor, the o1 model (December 2024), reached 78.0%, surpassing both GPT-4o and expert human performance (69.7%). By January 2025, DeepSeek-R1 matched o1's accuracy at 90–95% lower cost, following the dual trajectory of performance gains and affordability in LLM development (Ronaghi *et al*., 2025).

Another promising direction is in investigating how variations in prompt design affect AI outputs. This could involve controlled experiments that manipulate prompt variables to identify best practices for eliciting more accurate and reliable responses from AI systems. Modern frameworks, such as PARTS (Persona, Aims, Recipients, Theme, Structure; Park and Choo, 2025) from Google, role-based prompting and CoT prompting optimize prompt design by structuring inputs.

Finally, future research should explore hybrid human-AI workflows to balance efficiency with interpretative rigor. While LLMs can mimic analytical reasoning and generate plausible patterns from qualitative data, meaning is shaped by the researcher's subjectivity, cultural context and reflexive awareness. While LLMs may assist in organizing and summarizing qualitative inputs, they lack the hermeneutic depth and intentional meaning that define rigorous qualitative inquiry.

Researcher bias and AI output interpretation

Prompt design significantly influenced outputs. Role-playing prompts ('act as a researcher') yielded richer insights than direct commands. Iterative refinement and peer validation mitigated biases, aligning with strategies proposed by Zhang *et al*. (2024). However, inconsistent interpretations of 'critical analysis' across LLMs highlighted subjectivity risks.

AI will enhance interviews via real-time guidance and adaptive questioning. Automated theoretical sampling will streamline gap identification in iterative methodologies. Cross-linguistic tools will enable nuanced multilingual analysis. Integrated AI systems will bridge qualitative–quantitative data divides (Katz, 2025). As these capabilities evolve, ensuring transparency in model performance – as demonstrated by ChatGPT's measurable advantages – will be key to fostering researcher confidence.

Limitations

We have identified a number of limitations in this study.

LIMITED GENERALIZABILITY:

The study was conducted within a specific research context, based on a single-case design, which may limit the generalizability of our findings across different qualitative research settings, datasets or disciplines. Work is required to assess the robustness and transferability of the findings and refine the methodologies for different research environments.

SUBJECTIVITY IN EVALUATION:

While we incorporated human verification and established rubrics to assess the performance of AI outputs, the inherently subjective nature of qualitative analysis means that interpretative differences can influence the evaluation process. Incorporating automated performance metrics and inter-rater reliability measures can minimize subjectivity.

PROMPT SENSITIVITY AND TASK SELECTION:

The effectiveness of AI responses was highly dependent on the formulation of prompts. This sensitivity raises concerns about the consistency and reproducibility of the results, as variations in prompt phrasing can lead to significantly different outputs.

RAPIDLY EVOLVING AI CAPABILITIES:

Generative AI models, such as the LLMs tested, are advancing at a fast pace. The findings of this study reflect the performance of a generation of AI tools which will become outdated as new models with enhanced capabilities emerge. Longitudinal studies could explore how improvements in AI tools alter the balance between human expertise and automated analysis.

Conclusion

This study addresses a critical gap in knowledge by empirically comparing LLMs in qualitative tasks, contributing to the evolving discourse on AI-assisted research. It deepens understanding of the strengths and limitations of contemporary research techniques, particularly in utilizing virtual research methods supported by LLMs. While the consistency of LLMs may fluctuate with nuanced, context-dependent interpretations, they markedly improve research processes by efficiently summarizing, editing, identifying emerging themes, generating references, providing guidance, generating novel research questions, drafting and peer reviewing. They enable rapid data processing, uncovering insights that might otherwise go undetected by human coders. Together, these advances make research more accessible, efficient and adaptable. Our experimental findings demonstrate that while LLMs can automate several steps of qualitative analysis and reduce manual effort, their effectiveness depends on model and task. Among the models tested, ChatGPT-4o stands out for its consistency, nuanced outputs and ability to streamline workflows.

Further, this study contributes to the ongoing discourse on generative AI in qualitative research by offering a structured, theory-driven evaluation of LLM outputs. Within this context, a broader discussion on the risks of generative AI tools is critical, particularly regarding research quality and criteria for qualitative rigor, such as credibility, confirmability, transferability and researcher reflexivity. As AI tools evolve, future research should address the capacities required by organizations, research centers and universities to harness AI's potential. Emphasizing organizational learning as a knowledge-development process will be essential to creating active, efficient learning environments that maximize the benefits of hybrid human-AI collaboration in research. Hybrid workflows – using LLMs for initial analysis and humans for contextual refinement – balance efficiency with methodological integrity.

With iterative prompt refinement (e.g., role-playing prompts for richer insights) and rubrics (e.g., coding consistency scores), it is possible to optimize efficiency without sacrificing quality. Establishing rubric-based evaluations that directly reference qualitative quality standards will be critical for future research and practice. The academic community must continue asking not only what AI can do, but also what AI should do, and how we will know it has been done well.

Disclosure statement

In the course of this research, the authors used AI Scholar, Chat GPT-4o, Claude 3.5 Sonnet, Notebook LM and Perplexity to compare functionalities of different tools. The author(s) then reviewed and edited the result as required and take full responsibility for the content of the published paper.

References

Bano, M., Zowghi, D. and Whittle, J. (2023) 'Exploring qualitative research using LLMs', *arXiv:2306.13298v1[cs.SE]*, https://doi.org/10.48550/arXiv.2306.13298.

Bijker, R., Merkouris, S., Dowling, N. and Rodda, S. (2024) 'ChatGPT for automated qualitative research: content analysis', *Journal of Medical Internet Research*, 26, 1, e59050, https://doi.org/10.2196/59050.

Burleigh, C. and Wilson, A. (2024) 'Generative AI: is authentic qualitative research data collection possible?', *Journal of Educational Technology Systems*, 1, 27, https://doi.org/10.1177/00472395241270278.

Busetto, L., Wick, W. and Gumbinger, C. (2020) 'How to use and assess qualitative research methods', *Neurological Research and Practice*, 2, 14, https://doi.org/10.1186/s42466-020-00059-z.

Chen, Y., Benton, J., Radhakrishnan, A., Uesato, J., Denison, C., Schulman, J., Somani, A., Hase, P., Wagner, M., Roger, F., Mikulik, V., Bowman, S., Leike, J., Kaplan, J. and Perez, E. (2025) 'Reasoning models don't always say what they think', *arXiv:2505.05410 [cs.CL]*, https://doi.org/10.48550/arXiv.2505.05410.

Ghareeb, A., Chang, B., Mitchener, L., Yiu, A., Szostkiewicz, C., Laurent, J., Razzak, M., White, A., Hinks, M. and Rodriques, S. (2025) 'Robin: a multi-agent system for automating scientific discovery', *arXiv:2505.13400v1 [cs.AI]*, https://doi.org/10.48550/arXiv.2505.13400.

Jarrahi, M. and Newlands, G. (2024) 'Quality in qualitative research: through the lens of validity, reliability, and generalizability', preprint, *ResearchGate*, https://doi.org/10.13140/RG.2.2.21444.23682.

Kamoun, F., El Ayeb, W., Jabari, I., Sifi, S. and Iqbal, F. (2024) 'Exploring students' and faculty's knowledge, attitudes, and perceptions towards ChatGPT: a cross-sectional empirical study', *Journal of Information Technology Education: Research*, 23, https://doi.org/10.28945/5239.

Katz, A. (2025) 'The future of AI in qualitative research', *Tabbi*, available at https://tabbiresearch.com/knowledge/blog/the-future-of-ai-qualitative-research/ (accessed August 2025).

Lawsen, A. (2025) 'Comment on the illusion of thinking: understanding the strengths and limitations of reasoning models via the lens of problem complexity', *arXiv:2506.09250v1[cs.AI]*, https://doi.org/10.48550/arXiv.2506.09250.

Lu, C., Lange, R., Foerster, J., Clune, J. and Ha, D. (2024) 'The AI scientist: towards fully automated open-ended scientific discovery', *arXiv:2408.06292v3 [cs.AI]*, https://doi.org/10.48550/arXiv.2408.06292.

Mirhosseini, S. (2020) *Doing Qualitative Research in Language Education*, Palgrave Macmillan, Cham, https://doi.org/10.1007/978-3-030-56492-6_9.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S. and Farajtabar, M. (2024) 'GSM-Symbolic: understanding the limitations of mathematical reasoning in large language models', *arXiv:2410.05229v1[cs.LG]*, https://doi.org/10.48550/arXiv.2410.05229.

Neto, R., Souza, D., Dias, G., Celestino, M., Gurgel, A., Ramos, A., Queiroz, M., Cavalcanti, J. and Silva, A. (2023) 'A multi-criterion model for evaluating the quality of qualitative research', *International Journal of Scientific Management and Tourism*, 9, 6, pp.3884–911, https://doi.org/10.55905/ijsmtv9n6-030.

Park, J. and Choo, S. (2025) 'Generative AI prompt engineering for educators: practical strategies', *Journal of Special Education Technology*, 40, 3, 411–17.

Pattyn, F. (2025) 'The value of generative AI for qualitative research: a pilot study', *Journal of Data Science and Intelligent Systems*, 3, 3, pp.184–91, https://doi.org/10.47852/bonviewJDSIS42022964.

Perkins, M. and Roe, J. (2024) 'Generative AI tools in academic research: applications and implications for qualitative and quantitative research methodologies', *arXiv:2408.06872 [cs.HC]*, https://doi.org/10.48550/arXiv.2408.06872.

Rahimi, S. and Khatooni, M. (2024) 'Saturation in qualitative research: an evolutionary concept analysis', *International Journal of Nursing Studies Advances*, 6, June, paper 100174, https://doi.org/10.1016/j.ijnsa.2024.100174.

Ronaghi, S., Aveling, E. and Singer, S. (2025) 'Integrating AI into qualitative analysis', *AcademyHealth*, available at https://academyhealth.org/blog/2025-03/integrating-ai-qualitative-analysis (accessed August 2025).

Sabnis, S. and Wolgemuth, J. (2024) 'Common misconceptions and good practices in qualitative research in school psychology', *Journal of School Psychology*, 106, October, paper 101328, https://doi.org/10.1016/j.jsp.2024.101328.

Sinha, R., Solola, I., Nguyen, H., Swanson, H. and Lawrence, L. (2024) 'The role of generative AI in qualitative research: GPT-4's contributions to a grounded theory analysis', *LDT'24: Proceedings of the 2024 Symposium on Learning, Design and Technology*, pp.17–25, https://doi.org/10.1145/3663433.3663456.

Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S. and Farajtabar, M. (2025) 'The illusion of thinking: understanding the strengths and limitations of reasoning models via the lens of problem complexity', *arXiv:2506.06941v1 [cs.AI]*, https://doi.org/10.48550/arXiv.2506.06941.

Stenfors, T., Kajamaa, A. and Bennett, D. (2020) 'How to assess the quality of qualitative research', *Clinical Teacher*, 17, 6, pp.596–9, https://doi.org/10.1111/tct.13242.

Zhang, H., Wu, C., Xie, J., Rubino, F., Graver, S., Kim, C., Carroll, J. M. and Cai, J. (2024) 'When qualitative research meets large language models: exploring the potential of QualiGPT as a tool for qualitative coding', *arXiv:2407.14925v1 [cs.HC]*, https://doi.org/10.48550/arXiv.2407.14925.