Polski Przegląd Ewaluacyjny
Nr 1(3)/2024


ISSN 2956-5332
pte.org.pl/o-czasopismie

AI-Supported Individual Interviews: Opportunities and Threats for Evaluation

Tomasz Kupiec
University of Warsaw, Poland; ORCID: 0000-0002-6469-7746

Abstract

This article presents a comprehensive literature review of the opportunities and challenges associated with using AI to support qualitative evaluation, particularly semi-structured in-depth interviews. By analysing experiences across six stages of the interview process—planning, protocol development, sampling and recruitment, data collection, transcribing and analysis—the study explores the potential of AI in improving efficiency, scalability, and accuracy of qualitative studies. Key findings highlight AI's capacity to streamline transcription and coding processes while addressing limitations such as contextual understanding and emotional nuance. However, the reliance on human oversight, ethical concerns necessitate further research. This review underscores AI's role as a complementary tool, enhancing qualitative methods rather than replacing human expertise.

Keywords

Artificial Intelligence, AI, LLM, Qualitative Studies, In-Depth Interviews, Evaluation

Corresponding author(s):
Tomasz Kupiec, Center for European Regional and Local Studies (EUROREG), University of Warsaw, ul. Krakowskie Przedmieście 30, 00-927 Warsaw, Poland. Email: tomasz.kupiec@uw.edu.pl

Introduction

The exact timing of the onset of the AI revolution remains uncertain, yet it is evident that it is currently underway[1]. The literature highlights how AI is revolutionising or reshaping various social services, including education (Adıgüzel et al., 2023; Pratama et al., 2023; Rahiman & Kodikal, 2024), medicine (Shiwlani et al., 2023; Alowais et al., 2023; Zeb et al., 2024), and multiple branches of industry (Rane, 2023; Velarde, 2020; Javaid et al., 2022).

AI-induced transformation is obviously not confined to the private sector; it is increasingly reshaping public policy including how evidence and knowledge inform decision-making (Kuziemski & Misuraca, 2020). Traditionally, policymakers relied on historical data, expert intuition, and qualitative insights to address societal challenges. AI now enhances these processes by enabling the analysis of vast, real-time datasets, offering unprecedented precision and predictive capabilities (Benoit, 2024). These advancements allow policymakers to make evidence-based decisions that are more responsive to societal needs (Vredenburgh, 2024).

The changing practice and perception of evidence is related to how AI is transforming the social sciences by reshaping the methods and processes of knowledge production. AI-powered tools enhance the ability of researchers to analyse large datasets, test complex theories, and develop predictive models. These advancements open new possibilities for studying social phenomena with greater precision and speed (Grimes et al., 2023; Wang et al., 2023).

[1] The term "AI revolution" appears in 14,000 publications on Google Scholar, with 10,000 of these published in 2020 or later (as of December 2024).

Evaluation, both as a profession and as a tool for public policy, is no exception to the transformative influence of AI. As Nielsen (2023, 2025) observes, what is being evaluated is likely to change, along with how it will be evaluated and by whom. While examples of AI applications in evaluation can be found in the literature (e.g., Anand, 2025; Gatto & Bundi, 2025), they remain relatively scarce compared to other industries or professions. This scarcity may align with Jacob's (2024) argument that evaluators are adopting AI at a slower pace.

This article addresses the practical gap in the slow adoption of AI within evaluation practice. Its aim is to explore the potential of using AI for enhancing qualitative evaluation, particularly semi-structured in-depth interviews, drawing on current experiences from the social sciences.

The focus on interviews is justified by their prominence as the most widely used technique for gathering data in qualitative evaluation. Additionally, AI, in its widely understood contemporary form, is synonymous with large language models (LLMs), which are designed to understand and generate natural language (Roberts et al., 2024). Consequently, these models appear to be particularly well-suited for supporting the implementation of qualitative research. On one hand, they enable researchers to preserve the inherent advantages of qualitative studies, such as the ability to explore participants' emotions, complex meanings, perspectives, and points of view. On the other hand, LLMs have the potential to address some of the key limitations of qualitative research: the time-intensive nature of conducting interviews, the constraints of small sample sizes limiting generalisability, and the labour-intensive and complex process of data analysis, which often demands considerable expertise on the part of the researcher (Bailey,2008; Noble & Smith, 2014; Hennink & Kaiser, 2022).

The rest of the article is structured as follows. The Research Design section outlines the methodological approach, including the rationale behind adopting a quasi-systematic literature review. The Findings section is divided into six subsections, each corresponding to a specific stage of the interview process—planning, developing the instrument, sampling and recruitment, data collection, transcribing, and coding and analysis. Each subsection highlights AI applications, and evaluates their effectiveness and limitations. The Discussion synthesizes these findings, drawing comparisons across stages to identify overarching trends, challenges, and areas for further research.

Research design

To explore the use of AI in enhancing qualitative studies based on semi-structured in-depth interviews, a literature review was conducted. Initially, the intention was to perform a systematic literature review using Scopus. However, a preliminary comparison between the resources available in Google Scholar and Scopus revealed that a significant proportion of recent publications relevant to this topic—particularly those addressing empirical examples of using AI to conduct interviews—were not included in the latter database. This omission was primarily due to the nature of the excluded works, which often comprised grey literature, preprints, or articles from non-indexed journals.

As a result, Google Scholar was adopted as the primary source for the review, despite its recognised limitations as a standalone resource for systematic literature reviews (Shultz, 2007; Haddaway et al., 2015). Consequently, the methodological approach for this study is best described as a quasi-systematic review[2].

The aim was to explore the utility of AI at each stage of conducting interviews. To achieve this, the initial step involved identifying these stages, which subsequently guided the literature search. The process of conducting in-depth interviews was categorised into the following stages (Boyce & Neale, 2006; Knott et al., 2022):

- Planning: Generating or identifying research questions or problems. In the context of evaluation, this stage may also involve identifying the knowledge needs of evaluation users.
- Developing the Instrument: Designing the interview protocol.

[2] Any review that consciously deviates from the principles of a systematic review due to resource constraints (Olejniczak et al., in press).

- Sampling and Recruitment: Deciding on the sampling strategy and recruiting respondents.
- Data Collection: Conducting the interviews.
- Data Analysis : For the purposes of this review, this stage was further divided into two separate steps:
    - Transcribing,
    - Coding, and analysing the data.

A separate literature search was conducted for each of these stages, utilising distinct search strings for each (Table 1). The first 30 records for each stage were then screened based on their titles and the brief descriptions provided by Google Scholar (equivalent to abstracts). Two inclusion criteria were applied:

- Does the publication relate to the relevant stage of the interview process (e.g., does it discuss generating research questions)?
- Does the publication present empirical findings?

Only records that met both criteria and were available for retrieval were included in the subsequent analysis. The full texts of the included records were read and analysed. If an individual record appeared in two searches – pertaining to two different stages – it was analysed twice, with respect to both stages. The number of records included for each stage is presented in Table 1.

Table 1. Stages of Conducting Individual Interviews and Their Corresponding Search Terms

Stage	Google Scholar search term(s)	Number of records included
Planning	developing / generating / designing “research questions” ai / chatgpt	5
Developing the Instrument	developing / generating / designing “interview protocol” ai / chatgpt	2
Sampling and Recruitment	Identifying / recruiting / sampling interview respondents / participants ai / chatgpt	0
	emulating / simulating interviews / respondents ai / chatgpt	6
Data Collection	interviewer social science / conducting interview ai	7
Transcribing	interview transcription ai human comparison	14
Coding and analysing	qualitative data coding and analysis ai / ai qualitative analysis without coding	25

Source: own elaboration

Findings

The findings are presented according to the stages outlined in the research design section. For each stage, a concise overview of the identified studies is provided, followed by a summary detailing the number of studies that deemed AI either useful or ineffective, along with the advantages and disadvantages associated with the use of AI at this stage of the process.

Identifying research questions

Researchers have explored the application of AI in generating research questions across various fields, producing a mix of promising outcomes and challenges. A study in natural language processing (NLP) revealed that large language models (LLMs) generated ideas that were notably more novel than those created by human experts, although they were often less feasible. Human re-ranking of the AI-generated ideas further enhanced their quality, indicating the potential for AI-human collaboration in research ideation (Si et al., 2024). Healthcare research demonstrated AI's capability to address real-world priorities by analysing over 600,000 patient portal messages and generating research questions that aligned with

patient concerns in oncology and dermatology. Evaluators found a third of these questions to be highly significant and novel, highlighting AI's utility in patient-centred research (Kim et al., 2024). The use of ChatGPT in gastroenterology, particularly for topics such as inflammatory bowel disease and the microbiome, produced clear and relevant questions; however, their low originality raised questions about the tool's capacity for innovation in this area (Lahat et al., 2023). In interdisciplinary academia, ChatGPT was applied to generate research questions spanning diverse fields like tax policies, war economies, and behavioural studies. While it proved useful for sparking ideas, issues such as hallucinated citations and a lack of academic rigor underscored the need for caution in its application (Chatham et al., 2023). Finally, a broader study examining AI's role in generating original research ideas across scientific disciplines found that while AI could suggest plausible questions, significant human intervention was often required to refine and assess their feasibility (Elbadawi et al., 2024).

To summarise, two studies—one in patient-centred healthcare and the other in natural language processing—demonstrated that AI support can be both useful and reliable, particularly when enhanced by human re-ranking to refine its outputs. A less definitive picture emerges in the fields of gastroenterology and interdisciplinary academia, where concerns regarding limited originality and accuracy have raised questions about AI's overall effectiveness. In the final cross-disciplinary study, AI support could be perceived as less effective, as the ideas it generated often required substantial human refinement to become actionable, thereby limiting its direct utility in standalone applications.

Based on the reviewed articles, AI demonstrates several notable advantages in generating research questions, making it a potentially transformative tool for researchers. One of its most significant benefits is efficiency, as it streamlines the process of brainstorming and reviewing existing literature, saving valuable time. Another advantage is its ability to suggest novel ideas, often introducing perspectives and approaches that human experts may not have considered. Furthermore, AI enhances accessibility to advanced research ideation tools, enabling a broader range of researchers to engage with innovative methodologies and expand their investigative capacities.

However, the articles also highlight several limitations that need to be addressed. A key concern is the lack of originality in many AI-generated questions, which often fail to provide the depth and innovation required for meaningful research. Feasibility is another issue, as some suggestions are impractical or poorly aligned with current methodologies, requiring significant refinement before they can be actionable. Accuracy poses a further challenge, with AI sometimes producing incorrect or hallucinated content that necessitates thorough human validation. Finally, the presence of bias in AI outputs, stemming from the limitations of its training data, can restrict the diversity and inclusivity of its suggestions. These limitations underscore the importance of human oversight to fully harness the potential of AI in research ideation.

Developing the protocol

The only identified contribution regarding the design of the research tool is a study from Parker et al. (2023a), who investigated the application of LLMs in supporting the development and refinement of interview protocols. Utilising Castillo-Montoya's (2016) Interview Protocol Refinement framework, the authors demonstrate how ChatGPT can generate interview questions, structure inquiry-based conversations, provide feedback on protocols, and simulate interview scenarios. The study highlights the tool's potential to enhance efficiency, particularly in reducing the time and resources required for protocol development, making it highly valuable for novice researchers and projects with limited access to participants. Moreover, the research underscores the adaptability of ChatGPT in tailoring protocols to various research contexts and cultural sensitivities, thereby expanding its utility across diverse qualitative studies.

Despite its advantages, the study emphasises the limitations of relying solely on AI for this task. ChatGPT's outputs often require iterative refinement and critical human oversight to meet the nuanced demands of qualitative research. While the tool can simulate interviews and provide generic feedback, it lacks the contextual and emotional depth essential for capturing human experiences and cultural nuances. Furthermore, ethical considerations, including data privacy and the mitigation of potential biases in AI-generated outputs, remain significant challenges. The authors argue that while LLMs can

complement traditional methods, they cannot replace the role of human intuition and expertise in ensuring the quality and ethical integrity of research protocols. This study thus positions ChatGPT as a valuable yet supplementary tool in the evolution of hybrid human-AI research methodologies. To mitigate the limitations Parker et al. (2023b) proposed guidelines for the integration of large language models in developing and refining interview protocols.

Sampling and Recruitment

Our review did not identify any attempts to utilise AI for the purpose of identifying or recruiting interview respondents. However, there is considerable interest in both academic and applied research in exploring the potential of AI, particularly large language models such as GPT, to simulate respondents for qualitative interviews and quantitative surveys.

Ferreira et al. (2024) evaluated GPT-4's ability to simulate populations for testing the Eysenck Personality Questionnaire-Revised (EPQR-A) in three languages. While the virtual populations demonstrated specific personality traits, significant discrepancies arose when compared to real populations. The researchers highlight the potential of LLMs for pre-testing questionnaires but note that further refinements are necessary to align virtual and real populations. The usefulness of this application is unclear, showing promise for limited applications but requiring further development.

The capacity of GPT to simulate users in design research methods has also been explored. Freitas (2023) examined its performance in Card Sorting, Usability Testing, Desirability Testing, and Concept Testing. While the AI excelled in generating open-ended responses, it struggled with close-ended questions and visual-dependent tasks, demonstrating its utility for open-ended user feedback but limitations for more nuanced or visual methods.

Parker et al. (2023a) investigated the use of ChatGPT to simulate human responses and found it a valuable source of feedback but just for protocol refinement, prior to in-person piloting and but not a substitute for live interviews. Another interesting use case involves startup business model validation. Potekhin (2024) used ChatGPT to simulate customer interviews, finding notable alignment between AI and human responses regarding factual queries. However, the AI's limitations in predicting future customer behaviour suggest that it is more suited for preliminary validation rather than predictive tasks.

Gerosa et al. (2024) explored AI-generated text as an alternative to human qualitative data in software engineering. Persona-based prompting effectively simulated demographic-specific perspectives, but integrating AI with human data remains the most effective approach, making this a useful complementary tool rather than a replacement. In the same field Steinmacher et al. (2024) tested GPT's ability to replicate responses in surveys. Authors found that while the AI could emulate trends from specific demographics, its accuracy varied significantly across studies, with some results close to random baselines. This underscores its potential for demographic-specific applications but highlights its unreliability as a standalone tool.

The reviewed studies reveal varying levels of AI usefulness. Ferreira et al. (2024) and Steinmacher et al. (2024) underscore applications with unclear utility, whereas Freitas (2023), Parker et al. (2023a), and Gerosa et al. (2024) provide compelling evidence of AI's value in specific contexts. Potekhin's (2024) research bridges these categories, demonstrating utility in preliminary validation while highlighting limitations in predictive tasks. The advantages of using AI to simulate respondents include reductions in time and costs during the early stages of research, mitigation of participant recruitment challenges, and adaptability to diverse research contexts and demographics.

Although none of the studies found AI entirely without value, significant limitations persist. These include inconsistent alignment with real human behaviour, ethical concerns surrounding participant replacement, and restricted capabilities in tasks requiring visual or nuanced comprehension. Collectively, the studies illustrate that while AI offers promise for simulating respondents in qualitative and quantitative research, its applications remain highly context-dependent and necessitate considerable oversight.

Data Collection

There is a plethora of literature exploring the application of AI-based reviewers in domains such as job recruitment (e.g. Lee & Kim, 2021; Black & van Esch, 2020; Kammerer, 2021) and medicine (e.g. Hong, Smith, & Lin, 2022; Kanazawa et al., 2023; Gashi et al., 2021).

While examples from social science research practice are not that numerous, they also offer exciting insights into the potential of AI in conducting interviews as part of research in social science. Several studies highlight the diverse applications and potential of AI-driven interview systems. Cuevas et al. (2023) examined the use of large language models (LLMs) in chatbots for social data collection, involving 399 participants. Their findings demonstrated improvements in user engagement, although response richness showed limited enhancement, underscoring challenges in aligning user expectations with chatbot capabilities. Similarly, Wuttke et al. (2024) conducted a study comparing AI and human interviewers in political interviews among university students, revealing that AI achieved data quality comparable to human interviewers while offering scalability, though nuanced conversation management required refinement. Biswas et al. (2024) explored the impact of AI-powered asynchronous video interviews (AVIs) on perceptions of fairness and social presence in recruitment processes. Their study, involving 218 participants, showed no significant differences in overall experience based on AI interviewer demographics, though participant perceptions varied by their demographic attributes.

Chopra and Haaland (2024) investigated stock market non-participation through 381 AI-conducted interviews. Their findings highlighted AI's capability to uncover thematic patterns and maintain participant engagement, with many respondents preferring AI for its non-judgmental approach. Eaton et al. (2021) described NATO's use of an AI voice bot, DUCHESS, for collecting insights from over 2,000 NATO staff about collaboration tools during the COVID-19 pandemic. While the system efficiently gathered critical data, it faced limitations in replicating human-like empathy, revealing gaps in handling nuanced emotional responses. Geiecke et al. (2024) developed a versatile open-source platform for AI-led interviews, validated with 466 U.S. respondents. This platform demonstrated adaptability and scalability across topics such as political views and subjective mental states but required structured guidelines to optimize outcomes. Finally, Sparano (2022) explored social robots in educational settings, focusing on interviews with individuals affected by autism. While robots successfully conducted structured interviews, they encountered challenges with open-ended questions and detecting emotional nuances.

Across the reviewed studies, AI was found to be useful in all except one case, demonstrating its capacity to improve engagement, scalability, and data quality in various contexts. In studies such as those by Wuttke, Chopra and Haaland, and Geiecke, AI excelled in maintaining participant engagement, uncovering thematic patterns, and providing scalable solutions. However, challenges remained in areas like nuanced conversation management and emotional detection, as seen in the studies by Eaton and Sparano. The utility of AI was unclear in one study, where demographic variations influenced participant perceptions without yielding significant overall benefits. These findings highlight the predominance of useful applications while emphasizing areas for further refinement.

The use of AI in conducting interviews presents several notable advantages and limitations. One of the primary advantages is its scalability, as demonstrated in studies like those by Geiecke et al. (2024) and Chopra and Haaland (2024), where AI systems handled large sample sizes efficiently, enabling researchers to gather extensive qualitative data. AI also offers consistency and standardization in question delivery, reducing interviewer bias and ensuring uniformity across interviews. Furthermore, the ability of AI to maintain participant engagement and adapt dynamically, as seen in Wuttke et al. (2024), highlights its potential for enhancing data richness. Many participants appreciated AI's non-judgmental and anonymous nature, particularly in sensitive topics, as noted in the findings of Chopra and Haaland (2024).

However, limitations remain that constrain the broader applicability of AI in interviews. Challenges in detecting and responding to emotional and nuanced cues were evident in studies like Eaton et al. (2021) and Sparano (2022), where the lack of human empathy impacted the depth of engagement. Similarly, demographic-based variations in participant perceptions, as observed by Biswas et al. (2024), suggest that AI interactions may not always achieve uniform acceptance or effectiveness. Technical dependencies, such as the need for well-designed prompts and structured guidelines, as highlighted by

Geiecke et al. (2024), also pose constraints. Ethical concerns around privacy, fairness, and the trustworthiness of AI systems further underscore the need for careful implementation. Collectively, these findings suggest that while AI holds significant promise for conducting interviews, its effectiveness depends on addressing these limitations through thoughtful design and deployment.

Transcribing

AI-driven transcription technologies have been evaluated across diverse fields and settings, reflecting their growing relevance and application in research and practice. In behavioural finance, AI transcription tools such as Whisper were employed to analyse mixed-method interviews, demonstrating efficiency and accuracy while requiring manual adjustments for nuanced data analysis (Haberl et al., 2023). Similarly, in the field of cybersecurity, Siegel et al. (2023) tested multiple transcription tools in interviews involving technical jargon, finding that while AI tools performed adequately, human intervention remained necessary to ensure precision.

In healthcare, psychiatry, and psychological research, AI transcription has shown significant promise for various applications. Seyedi et al. (2023) compared tools like Amazon Transcribe, Whisper, and Otter.ai for psychiatric interviews, demonstrating feasibility under HIPAA-compliant conditions, though minor errors persisted. Similarly, Eftekhari (2024) explored intelligent speech recognition systems for cardiology research interviews, emphasizing time efficiency while addressing ethical and bias concerns. In psychological research, Pfeifer et al. (2024) evaluated AI transcription for analysing spoken language among younger and older adults, finding word error rates between 2.5% and 3.36%. These studies collectively highlight AI transcription's utility for qualitative research and clinical applications while emphasizing the importance of manual corrections to ensure accuracy.

In forensic investigations and law enforcement, AI transcription has been applied to investigative interviews, particularly in Norwegian police reforms. Moe’s (2023) research demonstrated the efficiency of Whisper in processing investigative interviews, reducing transcription time while maintaining legal and procedural compliance. Further studies emphasized the importance of adhering to frameworks such as the EU Artificial Intelligence Act to mitigate risks and ensure reliability in evidence handling. While AI systems significantly enhanced efficiency, human oversight remained essential to uphold the integrity of legal standards (Stoykova, 2024).

In educational research, ChatGPT was tested for its ability to refine AI-generated transcripts, achieving sub-1% word error rates, making it a highly efficient tool for qualitative analysis (Taylor, 2023). Finally, in sociology, a comparative study of nine transcription tools—including Whisper—demonstrated AI's ability to perform accurately across diverse contexts, though manual review was required for high-stakes or detailed analyses (Wollin-Giering et al., 2023).

Research on conversational AI further indicated that some systems are approaching human parity in transcribing natural, informal speech, especially under controlled conditions. However, challenges persist with informal or complex linguistic contexts, such as conversations between family members or in emotionally charged settings (Mansfield et al., 2021). These studies collectively illustrate the versatility of AI transcription across domains, from behavioural finance and cybersecurity to healthcare, psychology, and law enforcement. Despite promising results, reliance on human oversight and ethical concerns around data privacy and bias remain critical considerations for integrating AI transcription technologies into practical workflows.

Several metrics are employed in the analysed studies to compare the performance of AI and human transcription. The most commonly used metric is Word Error Rate (WER). Comparisons have been conducted under various conditions, including clear, structured content, such as broadcast news, and more conversational contexts. Benchmark human performance varied from "careful transcription" to rapid transcription in challenging conditions, which also influenced the resulting scores. These benchmarks provide a reference point for evaluating AI transcription tools, which frequently approach but seldom exceed human-level accuracy, particularly in complex or informal linguistic scenarios.

In specific studies (e.g. Pfeifer et al., 2024; Taylor, 2023; Seyedi et al., 2023), AI performance was described as closely matching or nearing human-level accuracy, though not entirely equaling it. However, in one study (Mansfield et al., 2021), AI was shown to perform worse than human transcription. An additional metric was introduced by Moe, who noted that AI-generated transcripts required human correction to achieve acceptable accuracy, but, interestingly, correcting AI-generated transcripts took 3.5 to 6.5 minutes (per three minutes of audio), compared to 10.5 minutes required for full manual transcription.

AI transcription tools offer notable advantages, including significant time and cost savings compared to manual transcription. Tools like Whisper and ChatGPT have demonstrated efficiency by reducing the time required for processing audio while achieving high accuracy in structured and controlled environments. These technologies also provide accessibility benefits, with support for multiple languages and the ability to process diverse accents, making them suitable for various fields such as healthcare, education, and law enforcement. However, limitations remain. Persistent accuracy challenges, particularly in handling technical jargon, informal speech, or non-standard accents, necessitate manual review to ensure reliability. Ethical concerns, including data privacy and compliance with regulations like GDPR and HIPAA, are critical in sensitive contexts such as healthcare and forensic investigations. Additionally, studies have highlighted biases in transcription performance across demographic groups, emphasizing the need for more inclusive datasets and robust algorithms. While AI transcription is becoming an increasingly viable option, its integration into research and practice requires careful consideration of these limitations to ensure both accuracy and ethical compliance.

Coding and analysing

Numerous studies have showcased AI's potential in enhancing the analysis of interview transcripts by generating coding frameworks. For instance, ChatGPT's performance was evaluated using focus group transcripts, where its coding aligned closely with human coders, significantly reducing time while maintaining thematic accuracy (Lixandru, 2024). Similarly, interviews with maternity care providers were analysed by Qiao et al. (2024), revealing that AI achieved over 80% alignment with human coding using both deductive and inductive approaches, although concerns regarding bias were highlighted. Kirsten et al. (2024) examined the application of large language models (LLMs) to workplace ethics interview datasets, demonstrating efficiency but also underscoring the necessity of human oversight for identifying latent themes.

The integration of semi-automated systems, such as "Cody," introduced by Rietz and Maedche (2021), combined rule-based and machine learning methods to enhance coding quality, particularly when user-defined parameters were employed. Generative AI tools were employed by Prescott et al. (2024) to code SMS text messages, achieving a 71% alignment with human-identified themes. While these tools showcased efficiency, challenges arose when dealing with nuanced themes. Additionally, Mazeikiene and Kasperiuniene (2024) used ChatGPT-4 to facilitate initial coding of TED Talks transcripts, though human intervention remained necessary for complex visual mapping.

AI's capabilities in rapid analysis were further demonstrated through the development of AQUA, an automated assistant achieving intercoder reliability comparable to human efforts, as outlined by Lennon et al. (2021). AI's effectiveness in replicating causal loop diagrams was highlighted in the work of Jalali and Akhavan (2024), who analysed interview transcripts for obesity prevention interventions. However, nuanced connections still required human refinement. Similarly, AI efficiently identified granular themes in engineering student reflections, as reported by Gamieldien et al. (2023).

Morgan (2023) assessed ChatGPT's ability to replicate descriptive themes from datasets, finding it particularly useful for simpler coding tasks. Thematic development and codebook refinement were explored by Dahal (2024), who acknowledged the ethical challenges inherent in AI-assisted processes. Collectively, these studies underscore AI's capacity to support qualitative analysis while emphasizing the importance of human oversight to address limitations and biases.

Several studies have explored how AI supports broader qualitative analysis tasks without directly generating codes. For example, Ciechanowski et al. (2020) applied AI to social media data analysis, identifying overarching themes without producing qualitative codes. Similarly, "QualiGPT," introduced by Zhang et al. (2024), improved workflows by suggesting high-level thematic groupings, although human coders were still required for detailed analysis. Katz et al. (2024) developed "GATOS," which facilitated the creation of manual codebooks by generating thematic structures for large datasets.

AI's potential for refining qualitative methods extends beyond coding. ChatPDF was used by Chubb (2023) to summarise arts-based transcripts, with manual validation necessary to ensure relevance. Lopez-Fierro and Nguyen (2024) investigated GPT-4's utility in refining codebooks and visualising knowledge construction, though human contextualisation was essential to address nuances. Feuston and Brubaker (2021) highlighted the limitations of AI in interpretive coding, despite its effectiveness in data exploration tasks such as sentiment analysis.

Innovations in AI tools have also aimed at improving collaboration and analytical rigour. Overney et al. (2024) introduced SenseMate, a tool that enhanced intercoder reliability and addressed gaps between novice and expert coders. Paulus and Marone (2024) explored the use of AI-assisted qualitative data analysis software (QDAS), identifying trade-offs between speed and interpretive depth. In health research, ChatGPT was employed by Hitch (2024) to augment reflexive thematic analysis, though human validation remained critical to ensure methodological rigor.

No study (of the analysed) to date has failed to highlight the utility of AI in qualitative data analysis, both in generating codes and in supporting broader analytical processes. The advantages of integrating AI into qualitative analysis include substantial time savings, scalability for large datasets, improved accessibility for non-experts, and the ability to facilitate iterative workflows. For instance, SenseMate effectively bridged the gap between novice and expert coders, while AQUA demonstrated reproducibility and transparency in topic extraction.

Nonetheless, significant limitations remain. AI tools often exhibit inconsistency in nuanced thematic analysis, necessitate precise prompt engineering, and lack transparency in automated decision-making processes. Additionally, human validation is frequently required to ensure methodological rigor and contextual accuracy, underscoring the continued importance of human oversight in AI-assisted qualitative research.

Discussion

The rapid advancements in AI, particularly with the rise of LLMs and generative AI, have begun to reshape the landscape of social science research in profound ways (Xu, et al., 2024). AI serves a dual role: as a tool to enhance research and evaluation methodologies, including automated content analysis, and as an object of study, where its sociotechnical dimensions and interactions with human actors are critically examined (Lindgren & Holmström, 2020). While these advancements promise innovation and new opportunities, they also present significant challenges, including risks of low-quality research proliferation (Prieto-Gutierrez et al., 2023).

The mistrust surrounding AI is often attributed to the so-called black-box problem—the lack of transparency regarding the input data and the processes by which outputs are generated. This article aims to shed light on this issue by examining current experiences in a relatively specific task: conducting interviews. By further dividing this task into distinct stages, the study seeks to understand the extent to which AI can enhance our work and identify the stages where its impact is most significant.

This review of literature highlights the nuanced role of AI in conducting interviews within social science research, with particular interest in evaluation studies, revealing both significant opportunities and persistent limitations. AI's utility varies considerably across the stages of the interview process, with some stages demonstrating extensive scholarly exploration and practical applications, while others remain underdeveloped.

AI has proven to be a valuable tool in qualitative studies, offering enhanced efficiency, scalability, and consistency. In particular, LLMs such as ChatGPT have demonstrated the ability to support tasks traditionally requiring significant time and expertise, such as designing protocols, conducting interviews, transcribing data, and coding responses. However, AI's effectiveness is often contingent on human oversight to address gaps in contextual understanding and ensure methodological rigor. This reliance underscores AI's role as a complementary, rather than autonomous, resource in evaluation studies.

AI faces significant limitations in sampling and recruitment and data collection. While AI interviewers can maintain engagement, their inability to detect and respond to nuanced emotional cues limits their effectiveness in capturing complex social phenomena. Similarly, efforts to simulate respondents have revealed inconsistencies in aligning AI outputs with real human behaviours, necessitating further development. Ethical concerns, including bias and data privacy, are particularly salient in these stages, requiring rigorous safeguards to ensure equitable and trustworthy applications.

Several avenues warrant further exploration. First, addressing the underdeveloped stage of sampling and recruitment could unlock new possibilities for improving participant identification and engagement. Research should focus on refining AI's demographic alignment capabilities and mitigating ethical risks associated with participant simulation.

Second, advancing the emotional intelligence of AI systems used in data collection could enhance their ability to capture nuanced responses. Integrating affective computing technologies and cultural sensitivity into AI models may help bridge this gap.

Finally, future studies should explore hybrid human-AI workflows that maximise the strengths of both. For instance, combining AI's efficiency in transcription and initial coding with human expertise in thematic refinement could enhance the quality and reliability of qualitative research outputs. Ethical frameworks and guidelines for implementing AI in social science research also require further development to ensure responsible adoption.

Declaration of conflicting interests: The author declares no potential conflicts of interest with respect to the research, authorship, or publication of this article.

Use of Generative-AI tools declaration: The author declare he has used ChatGPT 4o for the initial analysis and summary of identified articles relevant for each stage of process of conducting in-depth interviews. The results of analysis were validated, and check for the accuracy by the author. The same tool was used for polishing the text.

Funding: This work was supported by the National Science Centre, Poland, grant number 2019/33/B/HS5/01336

References

Adıgüzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology*.

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., ... & Albekairy, A. M. (2023). Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC medical education*, *23*(1), 689.

Anand A., Batra, G., & Uitto, J.I. (2025). Harnessing Geospatial Approaches to Strengthen Evaluative Evidence. In S.B. Nielsen, F. Mazzeo Rinaldi and G.J. Petersson (eds.). Artificial Intelligence and Evaluation . Emerging Technologies and Their Implications for Evaluation (pp. 196–218). London: Routledge. https://doi .org /10 .4324 /9781003512493

Bailey, K. D. (2008). *Methods of social research*. Simon and Schuster.

Biswas, S., Jung, J.-Y., Unnam, A., Yadav, K., Gupta, S., & Gadiraju, U. (2024). "Hi, I'm Molly, Your Virtual Interviewer!" Exploring the impact of race and gender in AI-powered virtual interview experiences. AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2024).

Black, J. S., & van Esch, P. (2020). AI-enabled recruiting: What is it and how should a manager use it?. *Business Horizons*, *63*(2), 215-226.

Boyce, C., & Neale, P. (2006). *Conducting in-depth interviews: A guide for designing and conducting in-depth interviews for evaluation input* (Vol. 2). Watertown, MA: Pathfinder international.

Castillo-Montoya, M. (2016). Preparing for interview research: The interview protocol refinement framework. The Qualitative Report, 27(5), 811-831. https://doi.org/10.46743/2160-3715/2016.2337

Chatham, M. D., Duncan, T. K., & Li, Y. (2023). Employing AI in Academia: The Role of ChatGPT in Generating Research Questions . SSRN. Retrieved from https://ssrn.com/abstract=4721270.

Chopra, F., & Haaland, I. (2024). Conducting qualitative interviews with AI. CESifo Working Papers, 10666.

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*, 16094069231193593.

Ciechanowski, L., Jemielniak, D., & Gloor, P. A. (2020). TUTORIAL: AI research without coding: The art of fighting without fighting: Data science for qualitative researchers. *Journal of Business Research*, *117*, 322-330.

Cuevas, A., Brown, E. M., Scurrell, J. V., Entenmann, J., & Daepp, M. I. G. (2023). Automated interviewer or augmented survey? Collecting social data with large language models. arXiv preprint arXiv:2309.10187.

Dahal, N. (2024). How Can Generative AI (GenAI) Enhance or Hinder Qualitative Studies? A Critical Appraisal from South Asia, Nepal. *The Qualitative Report*, *29*(3), 722-733.

Eaton, J., & Olaru, S. (2021). Does artificial intelligence conduct better research interviews than you? NATO Joint Analysis and Lessons Learned Centre.

Eftekhari, H. (2024). Transcribing in the digital age: qualitative research practice utilizing intelligent speech recognition technology. *European Journal of Cardiovascular Nursing*, zvae013.

Elbadawi, M., Li, H., Basit, A. W., & Gaisford, S. (2024). The role of artificial intelligence in generating original scientific research. International Journal of Pharmaceutics, 652, 123741.

Ferreira, G., Amidei, J., Nieto, R., & Kaltenbrunner, A. (2024). How well do simulated populations with GPT-4 align with real ones in clinical trials? The case of the EPQR-A personality test. In Proceedings of Artifi cial Intelligence and Data Science for Healthcare (AIDSH-KDD'24).ACM, New York, NY, USA, 7 pages.

Feuston, J. L., & Brubaker, J. R. (2021). Putting tools in their place: The role of time and perspective in human-AI collaboration for qualitative analysis. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1-25.

Freitas, G. (2023). Synthetic user testing: meaningful user emulation or stochastic parroting? Master Thesis, Politecnico Milano.

Gamieldien, Y., Case, J. M., & Katz, A. (2023). Advancing qualitative analysis: An exploration of the potential of generative AI and NLP in thematic coding. *Available at SSRN 4487768*.

Gashi, F., Regli, S. F., May, R., Tschopp, P., & Denecke, K. (2021). Developing intelligent interviewers to collect the medical history: lessons learned and guidelines. In *Navigating healthcare through challenging times* (pp. 18-25). IOS Press.

Gatto, L., & Bundi, P. (2025). The Use of Quantitative Text Analysis in Evaluations. In S. B. Nielsen, F. Mazzeo Rinaldi and G. J. Petersson (eds.). Artificial Intelligence and Evaluation. Emerging Technologies and Their Implications for Evaluation (pp. 144–167). London: Routledge. https://doi .org /10 .4324 /9781003512493.

Geiecke, F., & Jaravel, X. (2024). Conversations at scale: Robust AI-led interviews with a simple open-source platform. London School of Economics.

Gerosa, M., Trinkenreich, B., Steinmacher, I., & Sarma, A. (2024). Can AI serve as a substitute for human subjects in software engineering research?. *Automated Software Engineering*, *31*(1), 13.

Haberl, A., Fleiß, J., Kowald, D., & Thalmann, S. (2024). Take the aTrain. Introducing an interface for the accessible transcription of interviews. *Journal of Behavioral and Experimental Finance, 41,* 100891. https://doi.org/10.1016/j.jbef.2024.100891.

Haddaway, N. R., Collins, A. M., Coughlin, D., & Kirk, S. (2015). The role of Google Scholar in evidence reviews and its applicability to grey literature searching. *PloS one*, *10*(9), e0138237.

Hennink, M., & Kaiser, B. N. (2022). Sample sizes for saturation in qualitative research: A systematic review of empirical tests. *Social science & medicine*, *292*, 114523.

Hitch, D. (2024). Artificial Intelligence Augmented Qualitative Analysis: The Way of the Future?. *Qualitative Health Research*, *34*(7), 595-606.

Hong, G., Smith, M., & Lin, S. (2022). The AI will see you now: feasibility and acceptability of a conversational AI medical interviewing system. *JMIR Formative Research*, *6*(6), e37028.

Jacob, S. (2024). Artificial Intelligence and the Future of Evaluation: from Augmented to Automated Evaluation. *Digital Government: Research and Practice*.

Jalali, M. S., & Akhavan, A. (2024). Integrating AI language models in qualitative research: Replicating interview data analysis with ChatGPT. *System Dynamics Review*.

Javaid, M., Haleem, A., Singh, R. P., & Suman, R. (2022). Artificial intelligence applications for industry 4.0: A literature-based study. *Journal of Industrial Integration and Management*, *7*(01), 83-111.

Kammerer, B. (2021). Hired by a Robot: The Legal Implications of Artificial Intelligence Video Interviews and Advocating for Greater Protection of Job Applicants. *Iowa L. Rev.*, *107*, 817.

Kanazawa, A., Fujibayashi, K., Watanabe, Y., Kushiro, S., Yanagisawa, N., Fukataki, Y., ... & Naito, T. (2023). Evaluation of a medical interview-assistance system using artificial intelligence for resident physicians interviewing simulated patients: a crossover, randomized, controlled trial. *International Journal of Environmental Research and Public Health*, *20*(12), 6176.

Katz, A., Fleming, G. C., & Main, J. (2024). Thematic Analysis with Open-Source Generative AI and Machine Learning: A New Method for Inductive Qualitative Codebook Development. *arXiv preprint arXiv:2410.03721*.

Kim, J., Chen, M. L., Rezaei, S. J., Ramirez-Posada, M., Caswell-Jin, J. L., Kurian, A. W., ... & Linos, E. (2024). Can Artificial Intelligence Generate Quality Research Topics Reflecting Patients' Concerns? Stanford University. Retrieved from https://arxiv.org/abs/2411.14456.

Kirsten, E., Buckmann, A., Mhaidli, A., & Becker, S. (2024). Decoding Complexity: Exploring Human-AI Concordance in Qualitative Coding. *arXiv preprint arXiv:2403.06607*.

Knott, E., Rao, A. H., Summers, K., & Teeger, C. (2022). Interviews in the social sciences. *Nature Reviews Methods Primers*, *2*(1), 73.

Kuziemski, M., & Misuraca, G. (2020). AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications policy*, *44*(6), 101976.

Lahat, A., Shachar, E., Avidan, B., Shatz, Z., Glicksberg, B. S., & Klang, E. (2023). Evaluating the use of large language model in identifying top research questions in gastroenterology. Scientific Reports, 13 (4164). https://doi.org/10.1038/s41598-023-31412-2.

Lee, B. C., & Kim, B. Y. (2021). Development of an AI-based interview system for remote hiring. *International Journal of Advanced Research in Engineering and Technology (IJARET)*, *12*(3), 654-663.

Lennon, R. P., Fraleigh, R., Van Scoy, L. J., Keshaviah, A., Hu, X. C., Snyder, B. L., ... & Griffin, C. (2021). Developing and testing an automated qualitative assistant (AQUA) to support qualitative analysis. *Family medicine and community health*, *9*(Suppl 1).

Lindgren, S., & Holmström, J. (2020). A social science perspective on artificial intelligence: Building blocks for a research agenda. *Journal of Digital Social Research (JDSR)*, *2*(3), 1-15.

Lixandru, D. (2024). The Use of Artificial Intelligence for Qualitative Data Analysis: ChatGPT. *Informatica Economica*, *28*(1).

Lopez-Fierro, S., & Nguyen, H. (2024). Making Human-AI Contributions Transparent in Qualitative Coding. In *Proceedings of the 17th International Conference on Computer-Supported Collaborative Learning-CSCL 2024, pp. 3-10*. International Society of the Learning Sciences.

Mansfield, C., Ng, S., Levow, G. A., Wright, R. A., & Ostendorf, M. (2021). Revisiting Parity of Human vs. Machine Conversational Speech Transcription}}. *Proc. Interspeech 2021*, 1997-2001.

Mazeikiene, N., & Kasperiuniene, J. (2024). AI-Enhanced Qualitative Research: Insights from Adele Clarke's Situational Analysis of TED Talks. *The Qualitative Report*, *29*(9), 2502-2526.

Moe, M. K. (2023). *Post-processing automatic speech recognition transcriptions: A study for investigative interviews* (Master's thesis, NTNU).

Morgan, D. L. (2023). Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. *International journal of qualitative methods*, *22*, 16094069231211248.

Nielsen, S.B. (2023). Disrupting evaluation? Emerging technologies and their implications for the evaluation industry. New Directions for Evaluation, 178–179, 47–57. https://doi .org /10 .1002 /ev .20558

Nielsen, S.B. (2025). The Evaluation Industry and Emerging Technologies. In S.B. Nielsen, F. Mazzeo Rinaldi and G.J. Petersson (eds.). Artificial Intelligence and Evaluation. Emerging Technologies and Their Implications for Evaluation (pp. 266–286). London: Routledge. https://doi .org /10 .4324 /9781003512493

Noble, H., & Smith, J. (2014). Qualitative data analysis: a practical example. *Evidence-Based Nursing*, *17*(1), 2-3.

Olejniczak, K., Kupiec T., Wojtowicz D., (in press). Pozyskiwanie wiedzy – przeglądy źródeł wspierane Gen AI. In D. Batorski, K. Olejniczak, J. Pokorski, Wprowadzenie do zagadnień zastosowania gen AI w ewaluacji. PARP

Overney, C., Saldías, B., Dimitrakopoulou, D., & Roy, D. (2024). SenseMate: An Accessible and Beginner-Friendly Human-AI Platform for Qualitative Data Analysis. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (pp. 922-939).

Parker, J. L., Richard, V.M., & Becker, K. (2023a). Flexibility & iteration: Exploring the potential of large language models in developing and refining interview protocols. *The qualitative report*, *28*(9), 2772-2790.

Parker, J. L., Richard, V. M., & Becker, K. (2023b). Guidelines for the Integration of Large Language Models in Developing and Refining Interview Protocols. *The Qualitative Report*, *28*(12), 3460-3474.

Paulus, T. M., & Marone, V. (2024). "In Minutes Instead of Weeks": Discursive Constructions of Generative AI and Qualitative Data Analysis. *Qualitative Inquiry*, 10778004241250065.

Pfeifer, V. A., Chilton, T. D., Grilli, M. D., & Mehl, M. R. (2024). How ready is speech-to-text for psychological language research? Evaluating the validity of AI-generated English transcripts for analyzing free-spoken responses in younger and older adults. *Behavior Research Methods*, 1-11.

Potekhin, A. (2024). Harnessing ChatGPT for business model validation via AI-simulated interviews. Bachelor Thesis. Karelia University of Applied Sciences.

Pratama, M. P., Sampelolo, R., & Lura, H. (2023). Revolutionizing education: harnessing the power of artificial intelligence for personalized learning. *Klasikal: Journal of education, language teaching and science*, *5*(2), 350-357.

Prescott, M. R., Yeager, S., Ham, L., Rivera Saldana, C. D., Serrano, V., Narez, J., ... & Montoya, J. (2024). Comparing the efficacy and efficiency of human and generative AI: Qualitative thematic analyses. *JMIR AI*, *3*, e54482.

Prieto-Gutierrez, J. J., Segado-Boj, F., & França, F. D. S. (2023). Artificial intelligence in social science: A study based on bibliometrics analysis. *arXiv preprint arXiv:2312.10077*.

Qiao, S., Fang, X., Garrett, C., Zhang, R., Li, X., & Kang, Y. (2024). Generative AI for Qualitative Analysis in a Maternal Health Study: Coding In-depth Interviews using Large Language Models (LLMs). *medRxiv*, 2024-09.

Rahiman, H. U., & Kodikal, R. (2024). Revolutionizing education: Artificial intelligence empowered learning in higher education. *Cogent Education*, *11*(1), 2293431.

Rane, N. (2023). ChatGPT and Similar Generative Artificial Intelligence (AI) for Smart Industry: role, challenges and opportunities for industry 4.0, industry 5.0 and society 5.0. *Challenges and Opportunities for Industry*, *4*.

Rietz, T., & Maedche, A. (2021). Cody: An AI-based system to semi-automate coding for qualitative research. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1-14).

Roberts, J., Baker, M., & Andrew, J. (2024). Artificial intelligence and qualitative research: The promise and perils of large language model (LLM)'assistance'. *Critical Perspectives on Accounting*, *99*, 102722.

Seyedi, S., Griner, E., Corbin, L., Jiang, Z., Roberts, K., Iacobelli, L., ... & Clifford, G. D. (2023). Using HIPAA (health insurance portability and accountability act)–compliant transcription services for virtual psychiatric interviews: Pilot comparison study. *JMIR Mental Health*, *10*, e48517.

Shiwlani, A., Khan, M., Sherani, A. M. K., & Qayyum, M. U. (2023). Synergies of AI and Smart Technology: Revolutionizing Cancer Medicine, Vaccine Development, and Patient Care. *International Journal of Social, Humanities and Life Sciences*, *1*(1), 10-18.

Shultz, M. (2007). Comparing test searches in PubMed and Google Scholar. *Journal of the Medical Library Association: JMLA*, *95*(4), 442.

Si, C., Yang, D., & Hashimoto, T. (2024). Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers . arXiv preprint. Retrieved from https://arxiv.org/abs/2409.04109.

Siegel, R., Mrowczynski, R., Hellenthal, M., & Schilling, M. (2023). Poster: From hashes to ashes – A comparison of transcription services. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)* (pp. 3). ACM. https://doi.org/10.1145/3576915.3624380

Sparano, E. (2022). Robots in social research: Can social robots conduct an interview? Italian Sociological Review, 12(3), 1209–1228.

Steinmacher, I., Penney, J. M., Felizardo, K. R., Garcia, A. F., & Gerosa, M. A. (2024). Can ChatGPT emulate humans in software engineering surveys?. In *Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement* (pp. 414-419).

Stoykova, R., Porter, K., & Beka, T. The Ai Act in a Law Enforcement Context: The Case of Automatic Speech Recognition for Transcribing Investigative Interviews. *Available at SSRN 4913090*.

Taylor, Z. W. (2023). Using Chat GPT to clean interview transcriptions: a usability and feasibility analysis. *Available at SSRN 4437272*.

Velarde, G. (2020). Artificial intelligence and its impact on the fourth industrial revolution: a review. *arXiv preprint arXiv:2011.03044*.

Wollin-Giering, S., Hoffmann, M., Höfting, J., & Ventzke, C. (2024, January). Automatic Transcription of English and German Qualitative Interviews. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research* (Vol. 25, No. 1).

Wuttke, A., Aßenmacher, M., Klamm, C., Lang, M. M., Würschinger, Q., & Kreuter, F. (2024). AI conversational interviewing: Transforming surveys with LLMs as adaptive interviewers. arXiv preprint arXiv:2410.01824.

Xu, R., Sun, Y., Ren, M., Guo, S., Pan, R., Lin, H., ... & Han, X. (2024). AI for social science and social science of AI: A survey. *Information Processing & Management*, *61*(3), 103665.

Zeb, S., Nizamullah, F. N. U., Abbasi, N., & Fahad, M. (2024). AI in Healthcare: Revolutionizing Diagnosis and Therapy. *International Journal of Multidisciplinary Sciences and Arts*, *3*(3), 118-128.

Zhang, H., Wu, C., Xie, J., Rubino, F., Graver, S., Kim, C., ... & Cai, J. (2024). When Qualitative Research Meets Large Language Model: Exploring the Potential of QualiGPT as a Tool for Qualitative Coding. *arXiv preprint arXiv:2407.14925*.