

## Article

# Can an AI Chatbot Reliably Assess the Reporting Quality of Health-Related Qualitative Research? A Pilot Study

Shan Jin <sup>1</sup>, Abhinand Thaivalappil <sup>2</sup>, Ian Young <sup>2</sup> and Melissa MacKay <sup>1,\*</sup><sup>1</sup> Population Medicine, University of Guelph, Guelph, ON N1G 2W1, Canada; sjin13@uoguelph.ca<sup>2</sup> School of Occupational and Public Health, Toronto Metropolitan University, Toronto, ON M5B 2K3, Canada; athaival@torontomu.ca (A.T.); iyoung@torontomu.ca (I.Y.)

\* Correspondence: melissam@uoguelph.ca

## Abstract

Generative AI (genAI) is increasingly being integrated into critical appraisal and knowledge synthesis of research, with growing discourse on its value for qualitative research. Moreover, ChatGPT covers a major portion of the market share and is widely used by researchers and academics. However, a major gap remains in genAI's ability to assess the reporting quality of qualitative research studies. This pilot study aimed to determine whether GPT-5 could reliably assess the reporting quality of health-related qualitative research studies. The Standards for Reporting Qualitative Research (SRQR) checklist was used as a framework for completing both human and AI chatbot assessments. The SRQR is a widely used validated checklist for enhancing the reporting of qualitative research. Overall, 68 peer-reviewed studies were selected for testing, comparing human SRQR assessments with GPT-5 assessments. Findings revealed a wide variability in agreement between human and GPT-5 assessments, with crude agreement ranging between 60 and 100% and interrater reliability ranging between 0 and 0.85 across checklist items. Simpler, objective checklist items such as study purpose were often correctly identified and had stronger agreement compared to nuanced items. The model struggled with the following SRQR items leading to an "Unclear" response: research design/paradigm (41%), researcher reflexivity (41%), data processing (27%), and trustworthiness (46%). These preliminary results suggest that GPT-5 may struggle with tasks requiring nuance and interpretation of qualitative research, which is rich in complex methods and differing worldviews. Currently, GPT-5 and other AI chatbots should be used with caution and alongside human oversight when assessing the reporting quality of qualitative research within public health and healthcare.

**Keywords:** artificial intelligence; generative AI; quality of reporting; qualitative methods; reporting checklist; research transparency



Academic Editor: Jose María Merigo

Received: 20 July 2026

Revised: 24 August 2026

Accepted: 9 September 2026

Published: 11 September 2026

**Copyright:** © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

## 1. Introduction

### 1.1. The Rise of Generative Artificial Intelligence (AI)

Artificial intelligence (AI) encompasses computer algorithms that can simulate human intelligence and behaviour to perform tasks that typically require cognition [1]. Since its inception in the 1950s, AI has gone through multiple stages of development, with early success in programmes such as General Problem Solver and ELIZA [2]. However, limitations in processing capacity and reduced spending stagnated development [2]. In 2022, generative AI (genAI) systems, which are capable of creating “new” content (e.g., text, images, and audio) based on user prompts, attracted global attention with the release

of ChatGPT (GPT-3.5) [2,3]. ChatGPT harnesses large language models (LLMs) and deep learning techniques to learn using large volumes of data to predict, comprehend, and generate context-relevant language [2,3].

Currently, AI is a rapidly emerging field, piquing the interest of various sectors, including academia, industry, and government [1,4]. The public sector, which includes government organizations and public health agencies, is particularly interested in genAI due to its ability to increase operational efficiency, support evidence-informed decision-making, and assist with strategic resource management [5]. A 2024 survey indicated that 67% of public sector organizations were actively engaged in exploring genAI technologies [6].

### *1.2. Applications of genAI in Research and Concerns Around the Quality of Reporting*

Beyond operational applications in public and private sectors, genAI is increasingly being integrated into both quantitative and qualitative research workflows to improve efficiency, reduce human workload, and maintain rigour [1,7–9]. Examples of applications in quantitative research include automated study screening and classification, and automated data extraction [7,10,11], whereas applications in qualitative research focus more on data analysis methods, such as thematic analysis [8,9,12]. Despite these applications, genAI's capacity to automate research processes, support transparency in reporting, and uphold methodological rigour remains unclear [7–12]. This gap is particularly pertinent to qualitative research due to ongoing concerns around the quality of reporting [13,14].

### *1.3. Importance of Transparent and Rigorous Reporting in Qualitative Research*

A reporting guideline is a structured tool that contains a list of reporting items to be included in a research article [15]. These guidelines are typically used by researchers while preparing manuscripts to ensure a minimal set of essential information is presented [15]. When used appropriately, such guidelines ensure authors report all relevant information consistently, resulting in standardized and high-quality reporting of research [15]. Among the two most widely used reporting guidelines for qualitative studies are the (i) Standards for Reporting Qualitative Research (SRQR) and (ii) Consolidated criteria for REporting Qualitative research (COREQ) [16,17]. COREQ is a checklist designed for studies that utilize in-depth interviews and focus groups [17]. Subsequently, the SRQR guidelines built upon COREQ to provide a more general tool that can be applied to all types of primary qualitative research [16,17].

### *1.4. Rationale and Objective*

Qualitative research is used to explore and understand the meanings individuals assign to a complex, context-rich social or human problem [18,19]. Qualitative research methods have historically been used in the social sciences and have become increasingly adopted in public health and the medical sciences [20]. Such inquiries help explain how people interpret health and disease in their lives and disentangle the interplay between social, cultural, and systemic factors contributing to health [20]. To aid in conducting high-quality research, reporting checklists such as SRQR can help investigators think about which elements are needed during the conduct and reporting of their qualitative inquiries [20]. However, manually assessing the reporting quality can increase human workload and be time intensive. As such, the use of genAI to assist in evaluating the reporting quality provides an opportunity to greatly reduce the time needed while promoting greater adherence to reporting guidelines.

There has been increased use and interest in integrating AI in qualitative research [21]. ChatGPT makes up the largest proportion of the market share for AI chatbots being used by consumers [22]. Moreover, GPT-5 is easily accessible and smoothly integrates into academic and professional workflows [10,11,23]. Previous research has used ChatGPT in

other aspects of the qualitative research process [8,21], and there has been demonstrated potential in automating evidence synthesis processes [10,11,23]. Its conversational interface and wide popularity make it a suitable candidate for a pilot assessment.

No study has yet examined the ability of genAI to assess the reporting quality of qualitative research. Therefore, the aim of this pilot study was to determine whether an AI chatbot, specifically GPT-5, could reliably assess the reporting quality of health-related qualitative research studies. The SRQR checklist was used to assess areas of agreement and differences between human reviewers and GPT-5.

Outcomes of this pilot can help elucidate the potential for AI chatbots to be used in rapidly determining the reporting quality of health research and provide future research directions for scaling up future iterations of this inquiry.

## 2. Materials and Methods

### 2.1. Data Preparation

Figure 1 depicts the methodological framework followed for this pilot study. A preliminary scan of the literature was conducted in Google Scholar (Google LLC, California, United States of America) using keywords, including “Artificial Intelligence”, “Generative AI”, “Healthcare”, and “qualitative research” to identify recently published systematic reviews that synthesized qualitative research in fields such as healthcare, public health, and AI. Ten relevant qualitative systematic reviews were identified, and the following information was recorded: year of publication, topic, number of included studies, and standards of reporting used. A systematic review by Hogg et al. [4] was selected, which synthesizes qualitative research on diverse stakeholder perspectives around the implementation of AI in clinical settings, as the source of our study sample for assessing quality of reporting. This review was chosen because it comprised a large number of open-access studies ( $n = 68$ ) that varied in setting, participants, and methodology. We restricted our sample to open-access studies to avoid potential privacy and ethical concerns around uploading restricted files (i.e., studies behind paywalls) to genAI platforms.

We selected SRQR for our pilot study due to its rigour and wide application in qualitative research [16,24]. The SRQR checklist consists of 21 essential items across all aspects of primary qualitative research, from the qualitative approach used to contributions to the field, to ensure complete and transparent reporting [16]. In addition to assisting authors in drafting transparent manuscripts, these standards can also help peer reviewers in evaluating manuscripts for publication, and readers in assessing the reporting gaps within primary qualitative research [16].

### 2.2. Manual Data Collection

A Google Sheets (Google LLC, California, United States of America) spreadsheet was created to capture both manual and AI extractions of the SRQR checklist items with the following headings: unique article ID, link to the open-access article, reporting checklist used in the original article, 21 items for manual extraction, and 21 items for genAI extraction.

Prior to full manual assessments, an initial phase involved two reviewers (SJ, AT) independently assessing a subset of five studies. A strong overall human–human inter-rater reliability was observed (Cohen’s kappa  $> 0.80$ ) [25]. Subsequently, manual human checklist assessments were performed for the remaining included studies using the SRQR checklist by a reviewer (SJ) and followed by verification by a second reviewer (AT). Any disagreements were resolved through discussion between reviewers. This consolidated file served as a reference standard for comparison with genAI extractions of the SRQR checklists across all studies.

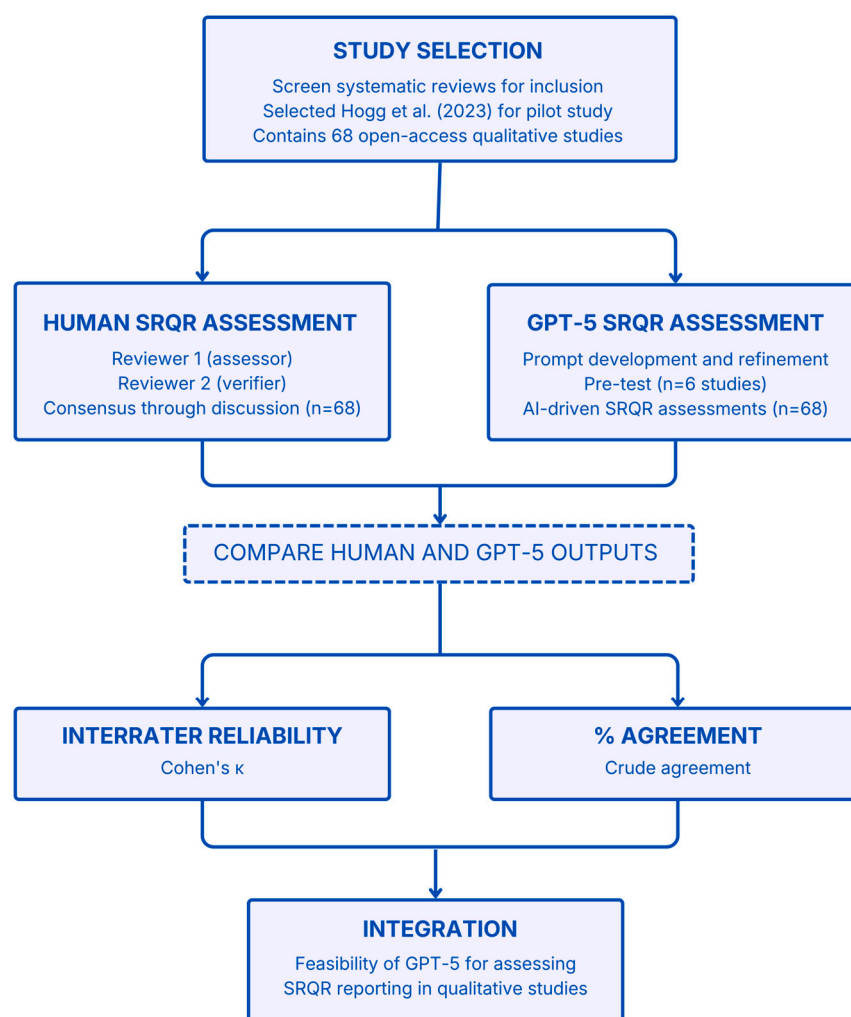


**Figure 1.** Methodological framework of the pilot study. A preliminary literature scan identified an eligible systematic review of qualitative studies on artificial intelligence in healthcare [4]. Sixty-eight open-access primary studies were assessed independently using the SRQR checklist by human reviewers and by GPT-5 using a standardized prompt. Human consensus ratings served as the reference standard. Agreement between human and GPT-5 assessments was evaluated using Cohen's kappa and crude agreement.

### 2.3. Prompt Development and AI Extraction

A pretest was conducted on six studies from the original systematic review to assess GPT-5's ability to assess the quality of reporting using the SRQR checklist and produce consistent outputs. The pretest revealed minor inconsistencies in outputs (i.e., formatting issues) and guided iterative refinements of the prompt provided to GPT-5 in September 2025 to ensure consistent outputs across studies.

The pilot was conducted using the ChatGPT Plus interface with access to the GPT-5 foundation model (model snapshot gpt-5-2025-08-07, accessed during September 2025). Prompts were input using default configuration settings, relying on the model's standard automated routing and default reasoning configuration. Each SRQR assessment was conducted in a new conversation, whereby the open-access article in PDF format was uploaded directly to the interface along with the prespecified prompt. These articles were only assessed once. No external applications or additional tools were used.

All items within the SRQR checklist were included within our prompt, in addition to attaching a PDF copy of each study. New chats were created for each new assessment to ensure consistent context and prevent prior conversation history from influencing subsequent

outputs or introducing hallucinated information from other studies. An amended version of the finalized prompt is shown below, and the full prompt that contains all checklist items to guide GPT-5 is provided in Supplementary File S1. The following prompt, in addition to a description of each item directly from the SRQR checklist, was applied to each study in GPT-5:

Using the attached paper, apply the Standards for Reporting Qualitative Research reporting guideline as a checklist written below and list which of the components it meets. For each number in the checklist, provide a 1 sentence justification and location where it is found in a table format and use “Yes”, “No”, and “Unclear” as your final decision. Please ensure that the output is presented in 4 columns: (i) item No., (ii) item name, (iii) Met?, (iv) justification & page number.

The following information was recorded for each GPT-5 extraction: item number, item name, presence of the item, a brief justification, and the page number where the item is found in the study.

#### 2.4. Data Analysis

Descriptive characteristics of the 68 articles included in this study were taken from the original review (e.g., publication year, setting, method), and the extent to which the articles followed reporting guidelines. Human and GPT-5 assessments were descriptively reported. While human assessments involved a binary response (yes vs. no), GPT-5 was instructed to generate a third “unclear” response to facilitate further descriptive examination of where the model struggled. Afterwards, “Unclear” outputs were ultimately collapsed to “No” when estimating interrater reliability to match human assessments. We also performed a sensitivity analysis to examine whether the exclusion of “Unclear” responses impacted coefficients in any meaningful way. Agreement between human and GPT-5 assessments was assessed using the ‘irr’ package in RStudio (version 2025.05.0 running R 4.4.0). Some items were reported across most or all studies, and thus, the kappa could not be computed. Therefore, we also calculated a crude agreement, defined as the percentage of studies for which the human and GPT-5 ratings matched exactly. Furthermore, potentially statistically significant differences in SRQR reporting between studies that followed a checklist (vs. did not follow) were assessed using the non-parametric Mann–Whitney U test because the assumptions of normality and homogeneity of variance were violated. All analyses were conducted using RStudio.

All data supporting this study are openly available at <https://doi.org/10.17605/OSF.IO/P6XW2>, including: the consolidated human SRQR reference-standard ratings; results of the six-study prompt pretest; the full item-level GPT-5 extractions (rating, justification, and page number for each of the 21 SRQR items across all 68 studies); and descriptive characteristics of the included studies.

### 3. Results

#### 3.1. Summary of Included Studies

This study included 68 peer-reviewed, open-access qualitative studies, sourced from Hogg et al.’s systematic review [4]. Detailed information on the descriptive characteristics of included studies can be accessed in the open database: <https://doi.org/10.17605/OSF.IO/P6XW2>. Most studies were published between 2019 and 2020 (47%), and conducted in the United States, United Kingdom, or Canada (53%) (Table 1). Studies captured the perspectives of various stakeholders, including healthcare professionals, developers, patients, caregivers, and the public, with the median number of participants being 23 (range: 3–16,519). Interviews were the most commonly used data collection method (50%). Within

the included studies, the vast majority (85%) did not use a reporting checklist. Of those that did, COREQ was the most commonly used (10%).

**Table 1.** Descriptive characteristics of 68 qualitative studies (n = 68 studies) (adapted from [4]).

| Study Characteristics       | n  | %    |
|-----------------------------|----|------|
| Publication Year            |    |      |
| 2014                        | 6  | 8.8  |
| 2015                        | 5  | 7.4  |
| 2016                        | 4  | 5.9  |
| 2017                        | 3  | 4.4  |
| 2018                        | 9  | 13.2 |
| 2019                        | 14 | 20.6 |
| 2020                        | 18 | 26.5 |
| 2021                        | 9  | 13.3 |
| Country                     |    |      |
| United States               | 15 | 22.1 |
| United Kingdom              | 14 | 20.6 |
| International <sup>a</sup>  | 6  | 8.8  |
| Canada                      | 7  | 10.3 |
| Australia                   | 5  | 7.4  |
| Germany                     | 3  | 4.41 |
| Netherlands                 | 3  | 4.41 |
| Other <sup>b</sup>          | 15 | 22.1 |
| Data Collection Method      |    |      |
| Mixed or Multi <sup>c</sup> | 15 | 22.1 |
| Interviews                  | 34 | 50.0 |
| Focus Groups                | 8  | 11.8 |
| Survey                      | 9  | 13.2 |
| Other <sup>d</sup>          | 2  | 2.9  |
| Reporting Checklists Used   |    |      |
| COREQ                       | 7  | 10.3 |
| RATS                        | 2  | 2.9  |
| RAMESES II                  | 1  | 1.5  |
| None                        | 58 | 85.3 |

<sup>a</sup> This category comprised studies conducted across multiple countries. <sup>b</sup> The following countries were represented: Austria (n = 1), Brazil (n = 2), China (n = 1), Colombia (n = 1), France (n = 2), Ghana (n = 2), Ireland (n = 1), Italy (n = 1), Kenya (n = 1), Sweden (n = 1), and Switzerland (n = 2). <sup>c</sup> This category comprised studies that employed multiple qualitative data collection methods (e.g., interviews and focus groups). <sup>d</sup> This category comprised two studies that employed alternative qualitative data collection methods: one used observation, and the other used the think-aloud protocol.

### 3.2. GPT-5 and Human Agreement on SRQR Items

GPT-5 and human responses to SRQR checklist items across 68 qualitative studies were compared. For the “Yes” category, items 2–4 (i.e., abstract, problem, and purpose) had perfect alignment, and items 10, 12, 18, 20, and 21 (i.e., links to literature, implications, participants, data collection, and funding) demonstrated high alignment (Table 2). In contrast, the lowest alignment was observed for items 1, 5, 7, 8, and 14 (i.e., title, design/paradigm, data analysis, sampling, and context). Overall, responses reported as “No” showed poor alignment in comparison, except for items 2–4, as both GPT-5 and human reviewers provided identical outputs. Cohen’s kappa coefficients were computed to assess the level of agreement between GPT-5 and human responses. Agreement levels ranged from none to near-perfect agreement (0.00–0.85). Kappa coefficients could not be calculated for items 2–4 due to homogeneous between-reviewer responses. Furthermore, GPT-5 showed no variability across some checklist items (i.e., 7, 8, 10, 12, 16, 18), resulting in coefficient values of 0; crude percent agreement was reported alongside these values. Items falling below

a moderate level of agreement ( $\kappa < 0.41$ ), reflecting checklist items with the greatest disagreement between GPT-5 and human reviewers, are bolded in Table 2.

**Table 2.** GPT-5 and human responses to SRQR checklist items across 68 qualitative studies and the level of human–GPT-5 agreement between assessments.

| Item No.                | Items Reported Across Studies (n, %) |            | Level of Agreement       |           |
|-------------------------|--------------------------------------|------------|--------------------------|-----------|
|                         | GPT-5                                | Human      | Kappa <sup>a,b</sup>     | Crude (%) |
| 1. Title                | 61 (89.7%)                           | 37 (54.4%) | <b>0.18</b>              | 61.8      |
| 2. Abstract             | 68 (100%)                            | 68 (100%)  | -                        | 100       |
| 3. Problem              | 68 (100%)                            | 68 (100%)  | -                        | 100       |
| 4. Purpose              | 68 (100%)                            | 68 (100%)  | -                        | 100       |
| 5. Design/paradigm      | 39 (57.4%)                           | 21 (30.9%) | <b>0.28</b>              | 61.8      |
| 6. Reflexivity          | 10 (14.7%)                           | 8 (11.8%)  | <b>0.23</b>              | 82.4      |
| 7. Context              | 68 (100%)                            | 59 (86.8%) | <b>0.00</b>              | 86.8      |
| 8. Sampling             | 68 (100%)                            | 52 (76.5%) | <b>0.00</b>              | 76.5      |
| 9. Ethics               | 63 (92.6%)                           | 59 (86.8%) | 0.53                     | 91.2      |
| 10. Data collection     | 68 (100%)                            | 65 (95.6%) | <b>0.00</b> <sup>c</sup> | 95.6      |
| 11. Instruments         | 65 (95.6%)                           | 60 (88.2%) | <b>0.13</b>              | 86.8      |
| 12. Participants        | 68 (100%)                            | 65 (95.5%) | <b>0.00</b> <sup>c</sup> | 95.6      |
| 13. Data processing     | 50 (73.5%)                           | 54 (79.4%) | 0.43                     | 79.4      |
| 14. Data analysis       | 65 (95.6%)                           | 48 (70.6%) | <b>0.20</b>              | 75.0      |
| 15. Trustworthiness     | 31 (45.6%)                           | 40 (58.8%) | <b>0.22</b>              | 60.3      |
| 16. Key findings        | 68 (100%)                            | 62 (91.2%) | <b>0.00</b>              | 91.2      |
| 17. Empirical data      | 67 (98.5%)                           | 60 (88.2%) | <b>0.20</b>              | 89.7      |
| 18. Links to literature | 68 (100%)                            | 67 (98.5%) | <b>0.00</b> <sup>c</sup> | 98.5      |
| 19. Limitations         | 66 (97.1%)                           | 60 (88.2%) | <b>0.37</b>              | 91.2      |
| 20. Implications        | 64 (94.1%)                           | 65 (95.6%) | 0.85                     | 98.5      |
| 21. Funding             | 63 (92.6%)                           | 63 (92.6%) | 0.78                     | 97.1      |

<sup>a</sup> Some coefficients could not be calculated due to homogeneous responses across studies. <sup>b</sup> Bolded values indicate anything below a moderate level of agreement ( $<0.41$ ). <sup>c</sup> Low variability between responses, interpret with caution.

Refer to the following open database: <https://doi.org/10.17605/OSF.IO/P6XW2> for the full item-level dataset, including individual GPT-5 and human ratings, justifications, and page-number citations for each study.

### 3.3. Sensitivity Analysis

Although human assessments did not involve an “Unclear” criterion (i.e., the item was either reported or not), GPT-5 was instructed in prompts to designate Yes, No, or Unclear for each checklist item across studies. Thus, descriptive statistics of GPT-5 initiated SRQR assessments resulting in “Unclear” were carried out. Overall, GPT-5 reported at least one unclear decision in 11 checklist items (52%). Of these, the median number of unclear decisions was 3 (IQR: 2–18), with the model being uncertain on the following items: research design/paradigm (41%), researcher reflexivity (41%), data processing (27%), and trustworthiness (46%). A sensitivity analysis was performed to assess interrater reliability coefficients excluding all “Unclear” GPT-5 responses, which revealed that most items had a drastic reduction in coefficients. The only items that saw a positive change in the level of agreement were reflexivity (+0.10) and trustworthiness (+0.22). Full counts and proportions of “Unclear” responses, along with the interrater reliability scores can be accessed in the following open database: <https://doi.org/10.17605/OSF.IO/P6XW2>.

### 3.4. Impact of Using a Reporting Guideline on SRQR Adherence

Only a small proportion of studies in our analysis used a reporting checklist ( $n = 10$ , 14.7%). Of these studies, most used COREQ ( $n = 7$ ), followed by RATS ( $n = 2$ ), and RAMESES II ( $n = 1$ ). There was a statistically significant difference in the number of SRQR items reported between studies that followed reporting checklists ( $M = 19.3$ ,  $SD = 0.8$ ) compared to those that did not ( $M = 16.5$ ,  $SD = 2.8$ ,  $p < 0.001$ ). Figure 2 depicts the distribution of adherence to SRQR based on within-study use of a reporting guideline. For detailed information on the reporting checklists used, along with additional descriptive characteristics for all 68 studies, refer to the following open database: <https://doi.org/10.17605/OSF.IO/P6XW2>.

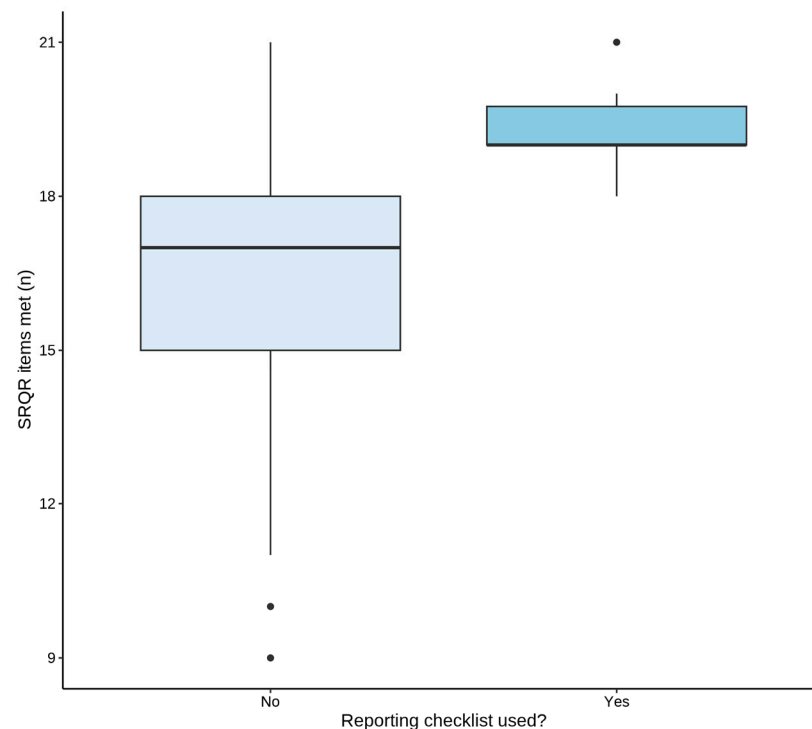


**Figure 2.** Boxplots indicating the distribution of adherence to the SRQR checklist items across the 68 qualitative studies based on whether the studies had followed a reporting guideline (medium blue plot) or not (pale blue plot). Bullets outside the box lengths represent outliers.

## 4. Discussion

### 4.1. Statement of Principal Findings

This study examined whether an AI chatbot, specifically GPT-5, could reliably assess the reporting quality of qualitative studies. Our results showed that alignment and agreement between GPT-5 and human responses to the SRQR checklist varied widely across 21 items. The complexity and explicitness of each respective checklist item may explain this variability between GPT-5 and human reviewers. For instance, perfect alignment and agreement generally occurred for straightforward and easily identifiable items, such as abstract, problem, and purpose. In contrast, poor alignment and agreement were observed in more nuanced and ambiguous items that are open to interpretation, including study design/paradigm, reflexivity, data analysis and sampling. Interestingly, most included studies in our study sample did not use a reporting checklist (85%). A statistically significant difference in SRQR adherence between studies that followed reporting checklists and those that did not was observed. The following sections will cover the use of reporting checklists in qualitative research, the strengths of GPT-5 in assessing quality of reporting,

areas where GPT-5 underperformed, and implications for the use of genAI in assessing the quality of reporting of qualitative research.

#### 4.2. Interpretation Within the Context of the Wider Literature

Although reporting checklists are widely available for qualitative studies, which emphasize their utility in supporting transparent and rigorous reporting, we found that most included studies (85%) did not use a reporting checklist. This aligns with the broader literature highlighting inconsistencies in the application of reporting checklists in qualitative research. For instance, a 2021 review that applied COREQ guidelines to qualitative studies in dentistry found that only 8% of studies had good reporting (i.e.,  $\geq 25$  items reported out of 32 criteria) [13]. Similarly, results from a study investigating the quality of reporting of 197 qualitative studies in nursing social science revealed that 95% of the studies were either moderately (i.e., 17–24 items reported) or poorly reported (i.e.,  $\leq 16$  items reported) [14]. Collectively, these findings underscore the need for more consistent use of reporting guidelines to increase transparency, completeness, and the overall quality of qualitative research [13,14].

Overall, LLMs tend to excel on tasks that are clear and simple, where relevant information is explicitly presented in the article's text [10,11,26,27]. This may be explained by the fact that LLMs are trained on large volumes of data, which allows them to accurately recognize patterns and locate predictable, structured pieces of information commonly found in research articles, such as the study purpose or objective [2,10,28]. This is reflected in our findings, as there was perfect alignment and agreement between GPT-5 and human reviewers for checklist items 2–4 (i.e., abstract, problem, purpose), which were among the most clearly defined and explicitly presented checklist items. These findings are consistent with evidence from studies evaluating LLMs for data extraction [10,11,26,27]. For instance, results from Reason et al. [27] showed that ChatGPT extracted PICO (i.e., Population, Intervention, Comparator, Outcome) research questions from almost 700,000 randomized controlled trial abstracts with 98% accuracy. Similarly, findings from an exploratory study on automated data extraction of COVID-19 modelling research demonstrated that ChatGPT extracted clearly stated data, such as study settings, with 93% accuracy [10]. These results suggest that GPT-5's performance is not solely determined by the clarity and simplicity of the prompt itself, but is also influenced by the explicitness and straightforwardness of information presented in an article's text [10,11]. Therefore, both our findings and the broader literature highlight the importance of working within the known limits of LLMs and applying them to general tasks that are clear, simple, and explicit to achieve high accuracy.

Our findings indicated that GPT-5 performed poorly on checklist items that contained multiple components and required deduction. First, this was evidenced by the high "Unclear" responses observed for some checklist items generated by GPT-5. The literature posits that LLMs do not currently have the capacity to handle complex and subjective tasks that require contextual reasoning [10,11,26,27,29,30]. LLMs provide outputs based on predictions and patterns drawn from their training data, rather than connecting and interpreting contextually relevant pieces of information presented throughout an article [29,30]. Notably, results from a recent study conducted by Zhang et al. [30] that investigates the relationship between working memory and contextual reasoning demonstrate that, despite increases in model size, a major limitation of modern LLMs remains their capacity to hold and process information effectively, as they lack true comprehension of prompts and struggle to process information meaningfully. Similarly, a study assessing the agreement between genAI and human data extraction methods concluded that LLMs are highly capable of completing straightforward extraction tasks; however, due to a lack of contextual

reasoning and underlying comprehension, complex tasks require human verification and oversight [29].

Our results are consistent with these findings, as poor alignment and agreement were observed between GPT-5 and human reviewers for title, design/paradigm, sampling, and data analysis, which are some of the more nuanced and subjective checklist items that are most open to interpretation and require working knowledge of qualitative methods to judge. This challenge is compounded by the nature of the SRQR tool itself, which requires not only identification of reporting elements but also an understanding of the methodological approach and which of the subcomponents within each item are relevant to that method [16]. This reasoning—recognizing when and how specific SRQR subcomponents of items apply, depending on the overall study design and methodology—was beyond the current capabilities of GPT-5. Because these models do not reveal how decisions are made or how information is weighted, their reasoning processes remain opaque, reinforcing what has been referred to as the “black box” of LLM decision-making [31]. Thus, while we can acknowledge the potential for complex checklist items to be modified to improve LLM interpretation and performance, at present, human qualitative methods expertise continues to be required to reliably assess such items.

#### *4.3. Strengths and Limitations*

Some limitations were identified. First and most importantly, this pilot study only assessed the feasibility of using GPT-5 to assess the quality of reporting in studies from one systematic review that synthesized global stakeholder perspectives on the implementation of AI in clinical settings. A subset of these studies was used to develop the prompt, leading to the possibility that the prompt may have been optimized for these studies in particular. Additionally, many studies included in this pilot had little variability in reporting, leading to homogeneous responses and skewed kappa coefficients towards zero. Therefore, our results may not be generalizable to other topics beyond clinical AI implementation in the healthcare context, and we advise readers to exercise caution in the interpretation of findings. We also note that reliance on a single, rapidly evolving proprietary tool poses a broader methodological limitation beyond generalizability alone: model behaviour may change over time without transparency or version control, which limits both reproducibility and consistency. To help mitigate this, we report the specific model version and assessment date used in this study (GPT-5, September 2025). Next, there are potential concerns with within-study reliability since it was observed that GPT-5 would occasionally differ in its responses to more nuanced checklist items such as trustworthiness, qualitative paradigm, and reflexivity. Thirdly, studies that utilize multiple methods (i.e., two or more qualitative methods) were challenging for GPT-5, as it likely used any information presented within the article to base its decision on a checklist item instead of carefully considering each method individually to develop meaningful and context-sensitive responses. Although this study did not involve a deeper analysis of GPT-5’s performance on multimethod and mixed method designs, further examination is needed. Notably, items such as ethical approval, author contributions, and funding were occasionally only provided on the journal website, causing GPT-5 to miss sufficiently reported items not found in the PDFs, further emphasizing the need for human oversight. Moreover, this pilot did not include a sensitivity analysis of prompt variations; future work should systematically evaluate how changes in prompt wording affect model outputs. Additionally, this study only assessed GPT-5’s ability to assess the quality of reporting in qualitative research, and thus, our findings may not be applicable to other AI chatbots. Lastly, we used ChatGPT’s standard web interface rather than API-based access with controlled parameters (e.g., temperature set to zero), which limits technical reproducibility across runs. This choice was intentional,

as it reflects how our intended audience, public health researchers without programming backgrounds, would realistically use this tool in practice; however, we acknowledge this trade-off between real-world applicability and strict reproducibility.

Since our study is the first to explore the feasibility of using GPT-5 to assess the quality of reporting in qualitative research, additional research should be conducted using other AI chatbots (e.g., Google Gemini) and other reporting checklists (e.g., COREQ) to improve understanding of the current capabilities of AI chatbots in automating research processes within qualitative research. To date, some reporting checklists, such as the Consolidated Standards of Reporting Trials [32], now have AI extensions to support the transparent and ethical use of AI in research (i.e., CONSORT-AI) [33]. Future research could explore the development of extensions for SRQR or COREQ checklists to mirror this in qualitative research [33].

#### 4.4. Implications for Policy, Practice and Research

There is an opportunity to explore the development of LLM-friendly checklist adaptations. Most reporting checklists are not optimized for genAI interpretation and application. Our pilot findings provide evidence of these challenges, particularly regarding contextual reasoning and multi-component checklist items. As seen across broader literature and best practice guidelines, prompt engineering influences genAI's performance [34,35]. Rather than recommending immediate checklist redesign, future work could explore the feasibility of developing LLM-friendly adaptations or extensions to existing tools. For instance, a dedicated section containing simplified, structured prompts or definitions may improve AI performance. Finally, we echo best practices from Cochrane, the Campbell Collaboration, JBI, and the Collaboration for Environmental Evidence on the responsible and ethical use of AI in evidence synthesis research [36], which emphasize human oversight, transparent reporting of AI use, and following evolving ethical standards.

## 5. Conclusions

To the best of our knowledge, this study is the first of its kind to assess the quality of reporting in health-related qualitative studies with generative AI. The pilot revealed that GPT-5 performed well on simple tasks where relevant information was explicitly presented and performed inconsistently on checklist items that often required forming logical deductions based on contextual elements. There was opacity in GPT-5's internal decision-making processes, raising potential reliability and transparency concerns. Future research should be conducted using other AI chatbots, with larger sample sizes, different areas of research, and reporting checklists (i.e., COREQ) to deepen our understanding of the applications and limitations of LLMs in performing quality of reporting assessments of qualitative research. Preliminary evidence suggests that the development of LLM-friendly reporting checklists or an AI extension to SRQR or COREQ could strengthen compatibility and improve the reliability of automated quality of reporting assessments.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/knowledge6030024/s1>, File S1: ChatGPT Prompt.

**Author Contributions:** Conceptualization, A.T. and M.M.; Methodology, S.J., A.T., I.Y. and M.M.; Validation, S.J., A.T., I.Y. and M.M.; Formal analysis, S.J. and A.T.; Investigation, S.J., A.T. and M.M.; Resources, M.M.; Writing—original draft, S.J., A.T. and M.M.; Writing—review and editing, A.T., I.Y. and M.M.; Supervision, A.T. and M.M.; Project administration, A.T. and M.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data supporting the findings of this study, including the consolidated human and GPT-5 SRQR assessments, pretest results, and descriptive characteristics of all 68 included studies, are openly available in the Open Science Framework at <https://doi.org/10.17605/OSF.IO/P6XW2>.

**Acknowledgments:** ChatGPT-5 was used as part of the research to evaluate the application and limitations in assessing the quality of reporting in qualitative research.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Xu, Y.; Liu, X.; Cao, X.; Huang, C.; Liu, E.; Qian, S.; Liu, X.; Wu, Y.; Dong, F.; Qiu, C.-W.; et al. Artificial Intelligence: A Powerful Paradigm for Scientific Research. *Innovation* **2021**, *2*, 100179. [[CrossRef](#)] [[PubMed](#)]
2. Fui-Hoon Nah, F.; Zheng, R.; Cai, J.; Siau, K.; Chen, L. Generative AI and ChatGPT: Applications, Challenges, and AI-Human Collaboration. *J. Inf. Technol. Case Appl. Res.* **2023**, *25*, 277–304. [[CrossRef](#)]
3. Feuerriegel, S.; Hartmann, J.; Janiesch, C.; Zschech, P. Generative AI. *Bus. Inf. Syst. Eng.* **2024**, *66*, 111–126. [[CrossRef](#)]
4. Hogg, H.D.J.; Al-Zubaidy, M.; Talks, J.; Denniston, A.K.; Kelly, C.J.; Malawana, J.; Papoutsis, C.; Teare, M.D.; Keane, P.A.; Beyer, F.R.; et al. Stakeholder Perspectives of Clinical Artificial Intelligence Implementation: Systematic Review of Qualitative Evidence. *J. Med. Internet Res.* **2023**, *25*, 39742. [[CrossRef](#)] [[PubMed](#)]
5. Nica, E.; Popescu, G.; Grupac, M.; Hurjui, M.C.; Parvu, I.A. Revolutionizing Public Sector Operations with Generative AI: Enhancing Efficiency, Optimizing Resources, and Empowering Data-Driven Decision-making for Modern Governance. *Manag. Res. Pract.* **2025**, *17*, 35.
6. Yang, L.; Wang, J. Factors Influencing Initial Public Acceptance of Integrating the ChatGPT-Type Model with Government Services. *Kybernetes* **2024**, *53*, 4948–4975. [[CrossRef](#)]
7. Hu, B.; Tomini, E.; Corrin, T.; Pussegoda, K.; Sandner, E.; Henriques, A.; Simniceanu, A.; Fontana, L.; Wagner, A.; Brazeau, S.; et al. Enhancing Evidence Synthesis Efficiency: Leveraging Large Language Models and Agentic Workflows for Optimized Literature Screening. *Cochrane Evid. Synth. Methods* **2025**, *3*, 70042. [[CrossRef](#)] [[PubMed](#)]
8. Morgan, D.L. Exploring the Use of Artificial Intelligence for Qualitative Data Analysis: The Case of ChatGPT. *Int. J. Qual. Methods* **2023**, *22*, 16094069231211248. [[CrossRef](#)]
9. Sakaguchi, K.; Sakama, R.; Watari, T. Evaluating ChatGPT in Qualitative Thematic Analysis with Human Researchers in the Japanese Clinical Context and Its Cultural Interpretation Challenges: Comparative Qualitative Study. *J. Med. Internet Res.* **2025**, *27*, 71521. [[CrossRef](#)] [[PubMed](#)]
10. Mahmoudi, H.; Chang, D.; Lee, H.; Ghaffarzadegan, N.; Jalali, M.S. Critical Assessment of Large Language Models' (ChatGPT) Performance in Data Extraction for Systematic Reviews: Exploratory Study. *JMIR AI* **2025**, *4*, 68097. [[CrossRef](#)] [[PubMed](#)]
11. Motzfeldt Jensen, M.; Brix Danielsen, M.; Riis, J.; Assifuah Kristjansen, K.; Andersen, S.; Okubo, Y.; Jørgensen, M.G. ChatGPT-4o Can Serve as the Second Rater for Data Extraction in Systematic Reviews. *PLoS ONE* **2025**, *20*, 0313401. [[CrossRef](#)] [[PubMed](#)]
12. Bijker, R.; Merkouris, S.S.; Dowling, N.A.; Rodda, S.N. ChatGPT for Automated Qualitative Research: Content Analysis. *J. Med. Internet Res.* **2024**, *26*, e59050. [[CrossRef](#)] [[PubMed](#)]
13. Al-Moghrabi, D.; Tschlaki, A.; Alkadi, S.; Fleming, P.S. How Well Are Dental Qualitative Studies Involving Interviews and Focus Groups Reported? *J. Dent.* **2019**, *84*, 44–48. [[CrossRef](#)] [[PubMed](#)]
14. Walsh, S.; Jones, M.; Bressington, D.; McKenna, L.; Brown, E.; Terhaag, S.; Shrestha, M.; Al-Ghareeb, A.; Gray, R. Adherence to COREQ Reporting Guidelines for Qualitative Research: A Scientometric Study in Nursing Social Science. *Int. J. Qual. Methods* **2020**, *19*, 1609406920982145. [[CrossRef](#)]
15. The EQUATOR Network. What Is a Reporting Guideline? Available online: <https://www.equator-network.org/about-us/what-is-a-reporting-guideline/> (accessed on 20 February 2026).
16. O'Brien, B.C.; Harris, I.B.; Beckman, T.J.; Reed, D.A.; Cook, D.A. Standards for Reporting Qualitative Research: A Synthesis of Recommendations. *Acad. Med.* **2014**, *89*, 1245–1251. [[CrossRef](#)] [[PubMed](#)]
17. Tong, A.; Sainsbury, P.; Craig, J. Consolidated Criteria for Reporting Qualitative Research (COREQ): A 32-Item Checklist for Interviews and Focus Groups. *Int. J. Qual. Health Care* **2007**, *19*, 349–357. [[CrossRef](#)] [[PubMed](#)]
18. Creswell, J.W.; Creswell, J.D. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 6th ed.; SAGE Publications: Thousand Oaks, CA, USA, 2022.
19. Lim, W.M. What Is Qualitative Research? An Overview and Guidelines. *Australas. Mark. J.* **2025**, *33*, 199–229. [[CrossRef](#)]
20. Stickley, T.; O'Caithain, A.; Homer, C. The Value of Qualitative Methods to Public Health Research, Policy and Practice. *Perspect. Public Health* **2022**, *142*, 237–240. [[CrossRef](#)] [[PubMed](#)]

21. Nguyen-Trung, K. ChatGPT in Thematic Analysis: Can AI Become a Research Assistant in Qualitative Research? *Qual. Quant.* **2025**, *59*, 4945–4978. [CrossRef]
22. Gottfried, G.; Bishop, W.; Anderson, M.; Faverio, M.; Park, E.; McClain, C. Americans and AI 2026: Chatbots, Smart Devices and Views on Impact. Available online: <https://www.pewresearch.org/internet/2026/06/17/americans-and-ai-2026-chatbots-smart-devices-and-views-on-impact/> (accessed on 25 February 2026).
23. Singh, H.; Singh, A. ChatGPT: Systematic Review, Applications, and Agenda for Multidisciplinary Research. *J. Chin. Econ. Bus. Stud.* **2023**, *21*, 193–212. [CrossRef]
24. Aref, H.A.T.; Witry, M.; Olufemi-Yusuf, D.; Guirguis, L.M. Ensuring Quality Qualitative Research Reporting in Community Pharmacy: A Systematic Literature Review. *Int. J. Pharm. Pract.* **2021**, *29*, 416–427. [CrossRef] [PubMed]
25. McHugh, M.L. Interrater Reliability: The Kappa Statistic. *Biochem. Med.* **2012**, *22*, 276–282. [CrossRef]
26. Mitchell, E.; Are, E.B.; Colijn, C.; Earn, D.J.D. Using Artificial Intelligence Tools to Automate Data Extraction for Living Evidence Syntheses. *PLoS ONE* **2025**, *20*, 0320151. [CrossRef] [PubMed]
27. Reason, T.; Langham, J.; Gimblett, A. Automated Mass Extraction of Over 680,000 PICOs from Clinical Study Abstracts Using Generative AI: A Proof-of-Concept Study. *Pharm. Med.* **2024**, *38*, 365–372. [CrossRef] [PubMed]
28. Liu, H.; Chen, Z.; Ding, Q.; Zhang, J.; Liu, J.; Hu, J.; Lu, W. Research on Paragraph-Level Functional Structure Recognition in Scientific Literature: A Data Augmentation Method Based on LLMs and Lexical Function. *Scientometrics* **2025**, *130*, 5473–5502. [CrossRef]
29. Daraqel, B.; Owayda, A.; Khan, H.; Koletsi, D.; Mheissen, S. Artificial Intelligence as a Tool for Data Extraction Is Not Fully Reliable Compared to Manual Data Extraction. *J. Dent.* **2025**, *160*, 105846. [CrossRef] [PubMed]
30. Zhang, C.; Jian, Y.; Ouyang, Z.; Vosoughi, S. Working Memory Identifies Reasoning Limits in Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 12–16 November 2024*; Al-Onaizan, Y., Bansal, M., Chen, Y.-N., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024.
31. Yang, W.; Wei, Y.; Wei, H.; Chen, Y.; Huang, G.; Li, X.; Li, R.; Yao, N.; Wang, X.; Gu, X.; et al. Survey on Explainable AI: From Approaches, Limitations and Applications Aspects. *Hum.-Centric Intell. Syst.* **2023**, *3*, 161–188. [CrossRef]
32. Hopewell, S.; Chan, A.-W.; Collins, G.S.; Hróbjartsson, A.; Moher, D.; Schulz, K.F.; Tunn, R.; Aggarwal, R.; Berkwits, M.; Berlin, J.A.; et al. CONSORT 2025 Statement: Updated Guideline for Reporting Randomised Trials. *BMJ* **2025**, *389*, e081123. [CrossRef] [PubMed]
33. Liu, X.; Cruz Rivera, S.; Moher, D.; Calvert, M.J.; Denniston, A.K. Reporting Guidelines for Clinical Trial Reports for Interventions Involving Artificial Intelligence: The CONSORT-AI Extension. *Lancet Digit. Health* **2020**, *2*, 537–548. [CrossRef] [PubMed]
34. Government of Canada Guide on the Use of Generative Artificial Intelligence. Available online: <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/guide-use-generative-ai.html> (accessed on 11 November 2024).
35. Meskó, B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. *J. Med. Internet Res.* **2023**, *25*, 50638. [CrossRef] [PubMed]
36. Flemyng, E.; Noel-Storr, A.; Macura, B.; Gartlehner, G.; Thomas, J.; Meerpohl, J.J.; Jordan, Z.; Minx, J.; Eisele-Metzger, A.; Hamel, C. Position Statement on Artificial Intelligence (AI) Use in Evidence Synthesis across Cochrane, the Campbell Collaboration, JBI and the Collaboration for Environmental Evidence 2025. *Campbell Syst. Rev.* **2025**, *21*, e70074. [CrossRef] [PubMed]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.