**Edcomtech:** Jurnal Kajian Teknologi Pendidikan
Vol 11, No 1, April 2026, pp. 108-127
https://doi.org/10.17977/um039v11i12026p108-127
pISSN: 2548-9879 (print) eISSN: 2599-2139 (online)
http://journal-fip.um.ac.id/index.php/edcomtech

# The Role of Generative AI in Advancing Educational Technology Research: A Systematic Review of Qualitative Data Analysis

**Dedi Aco[1*], Ming-Chou Liu[2], Harmita Sari[3]**
[1 2]*National Dong Hwa University, Shoufeng Township, Hualien County, Taiwan*
[3]*Universitas Muhammadiyah Palopo, Palopo, Sulawesi Selatan, Indonesia*
**Email corresponding author: 811288117@gms.ndhu.edu.tw*

**Article Info**

***Article history:***
Received 16-04-2026
Revised 25-04-2026
Accepted 27-04-2026
Published 30-04-2026

**How to cite:** Aco, D., Liu, M., & Sari, H. (2026). From The Role of Generative AI in Advancing Educational Technology Research: A Systematic Review of Qualitative Data Analysis. *Edcomtech: Jurnal Kajian Teknologi Pendidikan, 11*(1), 108–127. https://doi.org/10.17977/um039v11i12026p108-127



***Abstrak***

*Large language models (LLMs) semakin banyak digunakan dalam analisis data kualitatif; namun, pertanyaan terkait reliabilitasnya dibandingkan dengan peneliti manusia masih menjadi perhatian. Dengan mengikuti pedoman PRISMA 2020, tinjauan sistematis ini mensintesis bukti empiris mengenai penggunaan kecerdasan buatan generatif untuk pengodean data wawancara dan diskusi kelompok terarah. Dari 1.085 artikel yang diperoleh dari enam basis data akademik pada periode 2020–2026, sebanyak 30 studi memenuhi kriteria inklusi. Hasil penelitian menunjukkan bahwa LLMs, yang didominasi oleh model GPT-4, mampu mencapai tingkat kesepakatan tematik yang moderat hingga tinggi dengan peneliti manusia, dengan nilai Cohen's kappa berkisar antara 0,40 hingga 0,91 (median 0,72) serta tingkat akurasi antara 77% hingga 96%. Reliabilitas meningkat secara signifikan melalui penggunaan prompting yang dioptimalkan dan metode multi-run ensemble. Meskipun LLMs menunjukkan efisiensi yang sangat tinggi dengan penghematan waktu sebesar 80% hingga 95%, model ini masih mengalami keterbatasan dalam memahami nuansa budaya, kedalaman interpretasi, serta konteks pengodean yang kompleks. Oleh karena itu, bukti saat ini mendukung pemanfaatan LLMs sebagai alat bantu, bukan pengganti peneliti manusia. Pendekatan hibrida manusia-AI yang menggabungkan kecepatan komputasi dengan ketelitian interpretatif manusia menjadi strategi paling menjanjikan untuk analisis kualitatif yang rigor. Bagi peneliti pendidikan, temuan ini menunjukkan potensi LLMs dalam mengembangkan analitik pembelajaran kualitatif melalui pemrosesan data mahasiswa dalam skala besar secara cepat. Dengan demikian, penerapan pendekatan hibrida ini memungkinkan diperolehnya pemahaman yang lebih mendalam terhadap lingkungan pembelajaran berbasis teknologi tanpa mengorbankan nuansa pedagogis.*

***Kata Kunci:*** *Large Language Models; Analisis Data Kualitatif; Pengodean Tematik; Reliabilitas Antar Penilai; Kolaborasi Manusia-AI.*

**Abstract**

Large language models (LLMs) are increasingly used for qualitative data analysis; however, questions remain regarding their reliability compared to human coders. Following PRISMA 2020 guidelines, this

systematic review synthesizes empirical evidence on the use of generative artificial intelligence for coding interview and focus group data. Of the 1,085 records retrieved from six academic databases between 2020 and 2026, 30 studies met the inclusion criteria. The findings indicate that LLMs, predominantly GPT-4, achieve moderate to substantial thematic agreement with human coders, with Cohen's kappa values ranging from 0.40 to 0.91 (median 0.72) and accuracy rates between 77% and 96%. Reliability significantly improves with optimized prompting strategies and multi-run ensemble methods. Although LLMs demonstrate exceptional efficiency, reducing analysis time by 80% to 95%, they still face limitations in capturing cultural nuance, interpretive depth, and context-dependent coding. Therefore, current evidence supports the use of LLMs as an augmentation tool rather than a replacement for human researchers. Hybrid human-AI workflows, combining computational efficiency with human interpretive rigor, represent the most promising approach for robust qualitative analysis. For educational researchers, these findings highlight the potential of LLMs to advance qualitative learning analytics by enabling rapid processing of large-scale student data. Ultimately, this hybrid approach allows for deeper insights into technology-enhanced learning environments without sacrificing pedagogical nuance.

**Keywords:** *Large Language Models; Qualitative Data Analysis; Thematic Coding; Inter-rater Reliability; Human-AI Collaboration.*

## INTRODUCTION

Qualitative research methods, particularly the thematic analysis of interviews and focus groups, have long been foundational to understanding complex human experiences, behaviors, and social phenomena across disciplines (Klieger et al., 2024; Li et al., 2024). Traditional qualitative coding requires substantial human expertise, time investment, and often multiple coders to establish reliability (Prescott et al., 2024). Within educational technology research, qualitative inquiry is indispensable for evaluating learning-related applications, deeply understanding student experiences, and capturing complex classroom discourse. Scholars frequently rely on rigorous thematic analysis to investigate pedagogical phenomena, such as the integration of AI tools in language learning, bilingual instruction dynamics, and the overall efficacy of technology-enhanced learning environments. However, the labor-intensive nature of analyzing vast amounts of educational data ranging from student reflections to interview transcripts frequently limits the scale and scope of these studies. Finding methodological innovations to efficiently and reliably process this qualitative data without losing pedagogical nuance is therefore a critical priority for educational technology scholars. The emergence of large language models (LLMs), such as GPT-4, Claude, and Gemini, has introduced new possibilities for automating or augmenting qualitative data analysis (Mellon et al., 2024; Qiao et al., 2024).

Generative AI models demonstrate remarkable natural language understanding capabilities, raising questions about their potential to perform qualitative coding tasks that are traditionally reserved for trained human researchers (Parkington et al., 2025). Proponents argue that large language models (LLMs) could democratize qualitative research by reducing the time and resource barriers while maintaining analytical rigor (Borse et al., 2025; Sakaguchi et al., 2025). However, critics have expressed concerns about the interpretive depth, cultural

sensitivity, and contextual understanding required for high-quality qualitative analysis (Pattyn, 2024; Zhang et al., 2024).

The reliability of LLM-generated codes compared to that of human coding remains a critical empirical question. Inter-rater reliability metrics, such as Cohen's kappa (κ), have been the gold standard for assessing coding consistency in qualitative research (Raza et al., 2025). As researchers increasingly experiment with LLMs for coding interviews, focus groups, and open-ended survey responses, systematic evidence is needed to evaluate whether these tools can achieve acceptable reliability thresholds (Long et al., 2024; Nyaaba et al., 2025).

Several factors may influence LLM coding reliability, including model architecture and version, prompting strategies, the complexity and cultural specificity of themes, and the type of qualitative data analyzed (Theelen et al., 2024; Wachinger et al., 2025). Understanding these factors is essential for developing best practices and identifying appropriate use cases for LLM-assisted qualitative analysis (Yue et al., 2024).

This systematic review aims to examine the current state of empirical evidence on the use of generative AI, particularly large language models (LLMs), for qualitative coding of interviews and focus groups. It seeks to analyze the reliability metrics reported when comparing LLM-generated codes with human coding, including the range of agreement observed across studies. In addition, the review explores the types of LLM models, prompting approaches, and methodological configurations that have been evaluated in prior research. It also investigates the reported strengths and limitations of LLM-based qualitative coding in comparison to traditional human coding practices. Furthermore, this study aims to identify the key factors that influence the reliability and quality of LLM-generated qualitative codes, as well as to examine the specific implications of applying these tools within the field of educational technology research.

## METHOD

### Protocol and Regitration

This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (2020) statement (Kondo et al., 2024). The review protocol was developed a priori and specified the eligibility criteria, search strategy, screening procedures, and data extraction methods. Although not formally registered in PROSPERO, this decision was made due to two primary factors. First, PROSPERO's strict inclusion criteria prioritize systematic reviews related to direct human health or social care outcomes, often rejecting purely methodological evaluations of technology or software tools. Second, the rapidly evolving landscape of large language model research necessitates a swift review cycle that is incompatible with standard registry wait times. To ensure rigorous transparency and reproducibility despite the lack of formal registration, a detailed internal protocol was developed a priori, specifying exact search strings, inclusion criteria, and extraction matrices, which was strictly adhered to throughout the study.

### Eligibility Criteria

Studies were included in this review if they met several specific criteria. The target population consisted of qualitative data derived from interviews (structured, semi-structured, or unstructured), focus groups, or open-ended survey responses that required qualitative coding. The required intervention was the application of generative AI large language models, such as GPT-3, GPT-3.5, GPT-4, Claude, Gemini, LLaMA, PaLM, or similar transformer-based models, for qualitative coding or thematic analysis. Furthermore, studies had to feature a

direct comparator consisting of human coding performed by trained researchers and report quantitative reliability or validity outcomes, including Cohen's kappa, Krippendorff's alpha, percent agreement, accuracy, precision, recall, F1-scores, or semantic similarity measures. Eligible study designs encompassed empirical research (quantitative, qualitative, or mixed methods), including comparative studies, methodological validation studies, and experimental designs. Finally, publication characteristics were restricted to English-language peer-reviewed journal articles, conference proceedings, and preprints containing empirical data published between January 2020 and March 2026. Conversely, studies were excluded if they (1) utilized traditional NLP or non-generative machine learning approaches; (2) did not report quantitative reliability metrics; (3) lacked a direct comparison to human coders; (4) were opinion pieces, editorials, or reviews devoid of original empirical data; (5) focused exclusively on data collection, literature review assistance, or non-coding tasks; or (6) analyzed non-conversational text without an interview or focus group context.

**Information Sources**

Given the rapid pace of publication in generative AI research, traditional academic databases often suffer from significant indexing delays. Therefore, we deliberately incorporated AI-powered academic search engines, specifically SciSpace Deep Search and SciSpace Full Text Search, alongside Google Scholar and arXiv. SciSpace was selected because it utilizes semantic search algorithms that are exceptionally effective at capturing cutting-edge empirical studies, recent preprints, and interdisciplinary conference proceedings that might be missed by traditional keyword-based queries. This approach ensured a comprehensive capture of the most current LLM methodological evaluations across diverse academic domains.

**Search Strategy**

A comprehensive search strategy was developed by combining three core concepts using Boolean operators to capture all relevant literature. The first concept focused on generative AI and large language models, utilizing terms such as "large language model*," LLM, GPT, GPT-3, GPT-4, ChatGPT, "generative AI," "generative pre-trained transformer," Claude, Gemini, or "artificial intelligence." *The second concept* targeted qualitative methodologies, incorporating keywords like qualitative, interview*, "focus group*," "thematic analysis," "qualitative data analysis," QDA, coding, "open-ended," or "grounded theory." *The final concept* captured reliability and comparative evaluation metrics, employing terms including reliability, "inter-coder," intercoder, "inter-rater," interrater, agreement, validity, accuracy, "human vs," comparison, "Cohen's kappa," or Krippendorff. To accommodate variations in search engine syntax, this full search string was carefully adapted for each specific database. Furthermore, to ensure literature saturation, the reference lists of all included studies and relevant prior reviews were hand-searched to identify any additional eligible studies (Kondo et al., 2024).

**Selection Process**

The study selection process involved a rigorous multi-stage approach, beginning with the importation of 1,085 retrieved records into a reference management system. In the initial deduplication phase, AI-powered tools identified and merged duplicate records based on DOI, title, and author matching, yielding 242 unique records. Subsequently, two independent human reviewers conducted title and abstract screenings against the established eligibility

criteria. While this stage utilized AI-assisted screening tools to initially rank records by semantic relevance, human oversight remained absolute. The human reviewers independently screened all records ranked above the relevance threshold, alongside a 20% random sample of the lower-ranked records, to validate the tool's accuracy and prevent the exclusion of relevant literature. Any conflicts between the two independent reviewers were resolved through direct discussion until 100% consensus was achieved. Following this, full-text PDFs were obtained for 64 potentially eligible studies, which were independently assessed by two reviewers against all inclusion criteria. Disagreements during this full-text assessment were resolved through consensus discussions, and explicit reasons for exclusion were thoroughly documented. Ultimately, 30 studies successfully met all criteria and were included in the qualitative synthesis. The complete selection process is visually illustrated in the PRISMA flow diagram (Figure 1).

**Data Collection Process**

Data extraction was independently performed by two reviewers using a standardized form developed and pilot-tested in five studies. Discrepancies were resolved through discussion. For studies with insufficient detail, the authors were not contacted because of time constraints, and the available data were extracted with limitations noted.

**Data Items**

A comprehensive set of data items was extracted from each included study to facilitate a thorough systematic synthesis. First, general study characteristics were recorded, including the authors, publication year, publication venue, country of origin, research discipline, and any declared funding sources. To understand the methodological context, sample characteristics were detailed, capturing the type of qualitative data (such as interviews, focus groups, or open-ended surveys), sample sizes pertaining to both transcripts and participants, population characteristics, and the overarching research context. Furthermore, specific large language model (LLM) characteristics were meticulously documented, encompassing the exact models utilized (e.g., GPT-4, Claude 3.5), version details, parameter settings such as temperature, any fine-tuning or optimization methods, and the prompting approaches applied, whether zero-shot, few-shot, chain-of-thought, or iterative. The data extraction process also gathered vital details regarding the qualitative coding approach, distinguishing between inductive and deductive methods, and noting the specific coding frameworks used, the total number of codes or themes generated, and the human coders' characteristics, such as their quantity and level of expertise. To evaluate analytical performance, various reliability metrics were compiled, including Cohen’s kappa (κ), Krippendorff’s alpha, percent agreement, accuracy, precision, recall, F1-scores, semantic similarity measures, and the original authors' interpretations of these agreement levels. Finally, the extraction captured the key findings of each study, specifically the primary results comparing LLM and human coding performance, identified strengths and limitations, time and cost efficiency comparisons, and future recommendations alongside crucial risk of bias indicators, which assessed methodological transparency, the thoroughness of prompt reporting, sample size adequacy, and any potential conflicts of interest.

**Study Risk of Bias Assessment**

Due to the methodological diversity among the included studies and the lack of a validated risk-of-bias tool for large language model (LLM) evaluation research, a custom

quality assessment framework was developed. This framework was based on principles of methodological rigor for both qualitative inquiry and AI research. Studies were evaluated across seven criteria: methodological transparency, examining coding procedures and human-to-LLM implementation clarity; level of prompt reporting, from full to partial disclosure; and sample adequacy in size and diversity relative to the research question. The assessment also scrutinized appropriate metrics and their interpretation, the quality of the human comparator regarding coder expertise, and overall reproducibility based on model versions, parameters, and data characteristics. Finally, conflicts of interest were assessed through funding declarations and potential bias statements. Each criterion was rated low, moderate, or high risk, and an overall quality category was determined: high quality (0–1 high-risk criteria), moderate quality (2–3 high-risk criteria), or low quality (4 or more high-risk criteria). This risk-of-bias assessment was conducted independently by two reviewers, with discrepancies resolved through consensus discussion.

**Effect Measures**

The primary effect measures for this study evaluated alignment between large language model (LLM) outputs and human coding using several statistical lenses. Central to this evaluation was Cohen's kappa, measuring agreement between the LLM and human reviewers while correcting for chance. These values followed (Landis & Koch, 1977) framework: below 0.00 as poor, 0.00–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1.00 almost perfect. Besides kappa, percent agreement and accuracy metrics determined the proportion of codes where LLM and human coders agreed. For instances where human coding was the gold standard, the model's performance was analyzed using precision, recall, and F1-score. To capture thematic overlap beyond exact matches, semantic similarity was assessed through cosine similarity and other measures. Complementing these indicators, the study included secondary measures for practical feasibility, including time efficiency, cost-benefit comparisons, and qualitative assessments of code quality.

**Synthesis Methods**

Given the substantial heterogeneity in study designs, large language model (LLM) architectures, qualitative data types, and specific coding tasks, a meta-analysis was deemed inappropriate for this study. Instead, a narrative synthesis was conducted and organized into five distinct areas: a descriptive summary of study and sample characteristics; a categorization of LLM models and configurations, including prompting and optimization strategies; a synthesis of reliability metrics, such as Cohen’s kappa and accuracy; a comparative analysis of factors influencing reliability levels; and a thematic synthesis of the reported strengths and limitations of LLM coding. Detailed summary tables were constructed to facilitate cross-study comparisons. Subgroup analyses were performed to explore potential variations in reliability based on model type, prompting approach, data category, or research domain, whereas sensitivity analyses were employed to examine the influence of study quality on the reported reliability metrics.

**Study Selection**

Database searches retrieved 1,085 records. After AI-powered deduplication and merging, 242 unique records remained. Title and abstract screening excluded 178 records that did not meet the eligibility criteria (non-empirical studies, no LLM use, no qualitative coding focus, or no reliability metrics). Full-text assessments were performed on 64 articles, of which

34 were excluded for the following reasons: no quantitative reliability metrics were reported (n=18), no direct human comparison was made (n=9), non-generative AI methods were used (n=4), and duplicate publications were identified (n=3). Thirty studies met all inclusion criteria and were included in the qualitative synthesis. A PRISMA flow diagram is presented in Figure 1.

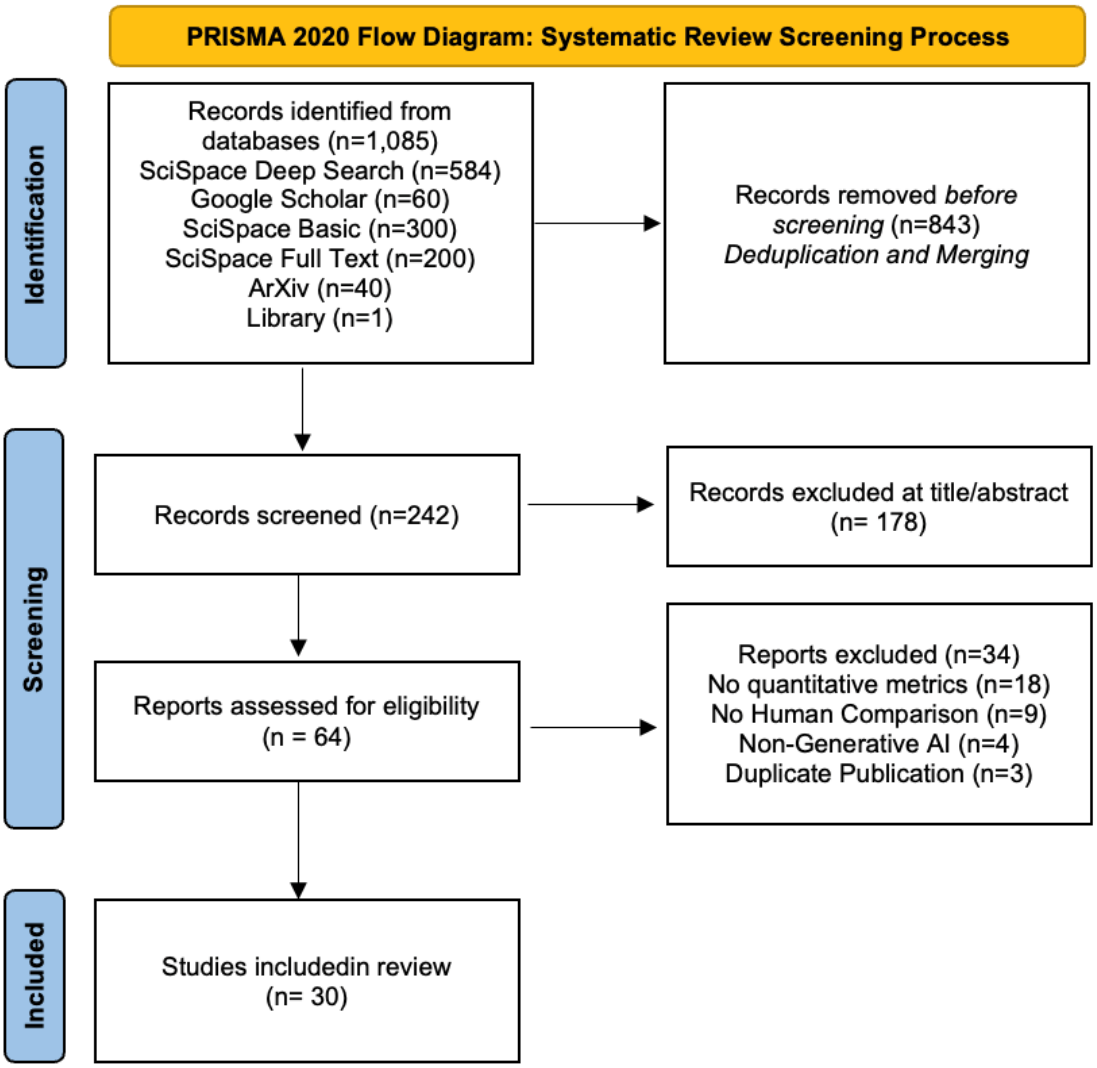


**Figure 1. PRISMA 2020 Flow Diagram of Study Selection Process**

**Study Characteristics**

The 30 included studies were published between 2023 and 2025, reflecting the recent emergence of generative AI for qualitative analysis. The majority were published in 2024 (n=16) and 2025 (n=11), with three studies from 2023. Publication venues included peer-reviewed journals (n=12), conference proceedings (n=6), and preprints (n=12).

Studies spanned diverse research disciplines, including healthcare and medical research (n=9), education (n=5), social sciences (n=6), software engineering and human–computer interaction (n=4), policy and public health (n=3), and mixed domains (n=3). Geographic representation included studies from North America (n=14), Europe (n=7), Asia (n=5), and multinational collaborations (n=4).

Sample sizes varied considerably. The number of transcripts or text units analyzed ranged from 1 to 237 (median=28). Participant numbers, when reported, ranged from 1 to

122 (median=20). Types of qualitative data included semi-structured interviews (n=18), focus groups (n=4), open-ended survey responses (n=5), and mixed or other conversational data (n=3).

Table 1 presents the detailed characteristics of all included studies, organized by publication year and first author.

**Table 1. Characteristics of Included Studies (n=30)**

| Study | Year | Domain | Data Type | Sample Size |
|---|---|---|---|---|
| Li et al. | 2024 | Healthcare | Semi-structured interviews | 20 patients |
| Borse et al. | 2025 | Healthcare | Interviews | Not reported |
| Pattyn | 2024 | Social sciences | Open-ended surveys | 122 responses |
| Lockwood et al. | 2025 | Education | Interviews | Not reported |
| Shah et al. | 2025 | Software engineering | Requirements interviews | Not reported |
| Yi et al. | 2025 | Mixed | Interviews | Not reported |
| Parkington et al. | 2025 | Mental health | Digital health data | Not reported |
| Prescott et al. | 2023 | Public health | SMS text messages | 40 messages |
| Theelen et al. | 2024 | Education | Student interviews | Multiple datasets |
| Simon et al. | 2025 | Mixed | Interviews | Not reported |
| Liu & Sun | 2023 | Policy | Stakeholder interviews | Not reported |
| Raza et al. | 2025 | Healthcare | Parent interviews | Not reported |
| Zhang et al. | 2024 | Mixed | Interviews | Multiple datasets |
| Kondo et al. | 2024 | Medical Education | Student interviews | Not reported |
| Wachinger et al. | 2024 | Healthcare | Patient interviews | Not reported |
| Qiao et al. | 2024 | Maternal health | In-depth interviews | Not reported |
| Yue et al. | 2024 | Mixed | Interviews | Not reported |
| Sakaguchi et al. | 2025 | Healthcare (Japan) | Clinical interviews | Not reported |
| Mellon et al. | 2024 | Social sciences | Open-ended surveys | Large-scale |
| Nyaaba et al. | 2025 | Mixed | Interviews | Not reported |
| Long et al. | 2024 | Education | Classroom dialogue | Not reported |
| Kim et al. | 2025 | Mixed | Complex qualitative data | Not reported |
| Jain et al. | 2025 | Psychotherapy | Therapist interview | 1 transcript |
| Klieger et al. | 2024 | Software engineering | Team messages | 237 messages |

**LLM Models and Configurations**

GPT-4 and its variants were the most frequently evaluated models, being used in 18 of the 30 studies (60%). ChatGPT-3.5 was used in eight studies (27%), Claude models (versions 1.3, 3.5 Sonnet) in five studies (17%), Gemini/PaLM in three studies (10%), and Google Bard in two studies (7%). Several studies have compared multiple models (Jain et al., 2025; Li et al., 2024; Prescott et al., 2024).

Prompting approaches varied widely. Zero-shot prompting (providing task instructions without examples) was used in 12 studies, few-shot prompting (including example codes) in eight studies, iterative or chain-of-thought prompting in seven studies, and multi-run ensemble approaches in four studies (Jain et al., 2025; Li et al., 2024). Temperature settings, when reported, ranged from 0.0 (deterministic) to 1.0 (creative), with most studies using default settings or not reporting parameters.

Only three studies have reported fine-tuning or supervised optimization of LLMs for specific coding tasks (Shah et al., 2025; Yi et al., 2025). Most studies used off-the-shelf models

via APIs or web interfaces. Prompt transparency varied: 12 studies provided full prompts in the appendices, 10 provided partial prompt descriptions, and 8 did not report prompt details.

Table 2 summarizes the LLM models and configurations used across the included studies.

**Table 2. LLM Models and Configuration in Included Studies**

| Study | Year | LLM Model | Prompting Approach | Key Configuration Details |
|---|---|---|---|---|
| Li et al. | 2024 | GPT-4 | Iterative, two phase | Naive and comparison phases |
| Borse et al. | 2025 | GPT-4o, GPT-4.5 | Optimized prompts | Hyperparameter tuning |
| Pattyn | 2024 | ChatGPT | Deductive coding | Improved inter-coder reliability |
| Lockwood et al. | 2025 | ChatGPT-4 | Not reported | Compared to novice coders |
| Parkington et al. | 2025 | GPT-4o | Out-of-the-box | Knowledge-based variants tested |
| Prescott et al. | 2024 | ChatGPT, Bard | Inductive/deductive | Both models compared |
| Simon et al. | 2025 | Claude 3.5 Sonnet | Multi-agent system | Consistency benchmarking |
| Liu & Sun | 2023 | GPT-4 | Not reported | Theme-level coding |
| Zhang et al. | 2024 | QualiGPT tool | Multiple scenarios | Substantial IRR reported |
| Sakaguchi et al. | 2025 | ChatGPT-4 | Japanese context | Cultural interpretation focus |
| Mellon et al. | 2024 | Claude 1.3 | Large-scale categorization | Best accuracy: 93.9% |
| Long et al. | 2024 | GPT-4 (customized) | Expert-guided | High consistency achieved |
| Jain et al. | 2025 | Gemini 2.5, GPT-4o, Claude 3.5 | Multi-run ensemble | Dual reliability metrics |
| Klieger et al. | 2024 | GPT-4 Turbo, GPT-3.5 | Iterative classification | Bales' IPA framework |

**LLM Models and Configurations**

The risk of bias assessment revealed moderate overall quality across the included studies. Methodological transparency was high in 18 studies (60%), moderate in 10 (33%), and low in 2 (7%). Prompt reporting was a common weakness; only 12 studies (40%) provided full reports, limiting reproducibility. Sample adequacy was generally appropriate, although 8 studies (27%) used very small samples (n<10 transcripts), raising concerns about generalizability.

The appropriate use of reliability metrics was high in 24 studies (80%), with 6 studies (20%) showing methodological concerns, such as unclear metric calculation or interpretation. Comparator quality (description of human coding) was adequate in 22 studies (73%) but limited in 8 studies (27%) that did not fully describe human coder training or expertise.

Reproducibility was moderate overall; 14 studies (47%) provided sufficient detail for replication, 12 (40%) provided partial detail, and 4 (13%) lacked key information, such as model versions or parameters. Conflict-of-interest declarations were present in 20 studies (67%), with 10 studies (33%) not reporting funding or potential biases.

Overall study quality was rated as high in 11 studies (37%), moderate in 16 studies (53%), and low in 3 studies (10%). Sensitivity analyses (see Section 3.6) examined whether study quality influenced reported reliability metrics.

## RESULTS

### Reliability Metrics: Cohen's Kappa

Across the 22 evaluated studies, Cohen's kappa values comparing large language models (LLMs) and human coders ranged from 0.40 to 0.91, with a median of 0.72, indicating substantial overall agreement. Specifically, the distribution of these agreement levels included moderate agreement (κ = 0.41–0.60) in 5 studies (23%), a predominance of substantial agreement (κ = 0.61–0.80) in 11 studies (50%), and almost perfect agreement (κ = 0.81–1.00) in 6 studies (27%). Several specific findings highlight performance variations that are highly dependent on the chosen model and methodological approach. For instance, Li et al. (2024) reported moderate agreement (κ = 0.401) when using GPT-4 to code patient interviews, noting that the model identified fewer subthemes than human researchers. In contrast, Klieger et al. (2024) observed stronger performance with substantial agreement (κ = 0.729) using GPT-4 Turbo for team communication messages, and Parkington et al. (2025) achieved strong agreement (κ = 0.84) using GPT-4o for mental health research excerpts. The highest levels of reliability were reported by (Jain et al., 2025), who achieved almost perfect agreement through a multi-run ensemble approach utilizing Gemini 2.5 Pro (κ = 0.907), GPT-4o (κ = 0.853), and Claude 3.5 (κ = 0.842). Furthermore, beyond model selection, prompting strategies proved to be a critical factor; Borse et al. (2025) demonstrated that prompt optimization significantly improved agreement from moderate to substantial (κ > 0.60) for models such as ChatGPT-4o and GPT-4.5-preview.

### Reliability Metrics: Percent Agreement and Accuracy

Eighteen studies evaluated large language model (LLM) performance using percent agreement or accuracy metrics, reporting values that ranged broadly from 36% to 96%, with a median of 81%. Several key findings illustrate the varying capabilities and limitations of these models depending on the task and context. For instance, Mellon et al. (2024) demonstrated that Claude 1.3 achieved 93.9% accuracy on large-scale open-text categorization, closely approaching the human performance baseline of 94.7%. Similarly, Liu & Sun (2023) found that GPT-4 aligned with human coders at 77.89% for specific themes and reached 96.02% for broader theme categories in policy interviews. Furthermore, Qiao et al. (2024) reported that ChatGPT maintained over 80% overall coding accuracy in maternal health interviews while simultaneously achieving an 81% reduction in coding time. However, performance is not universally high across all studies and coding approaches. Prescott et al. (2024) observed notably lower agreement rates, with ChatGPT showing only 47% and 37% agreement for inductive and deductive coding, respectively, while Bard achieved just 37% and 36% in the same categories. Additionally, Sakaguchi et al. (2025) highlighted significant contextual limitations, reporting that while agreement exceeded 80% for descriptive themes, it dropped to approximately 30% for culturally nuanced themes in Japanese clinical contexts.

### Reliability Metrics: Semantic Similarity and Other Measures

In addition to traditional agreement percentages, several studies have employed alternative reliability metrics to evaluate large language model (LLM) performance. Five studies utilized semantic similarity measures, such as cosine similarity and thematic overlap

scores, as either complementary or primary evaluation metrics. For instance, Jain et al. (2025) reported highly accurate cosine semantic similarity scores ranging from 92% to 95% for top-performing LLMs in multi-run evaluations. Similarly, Raza et al. (2025) applied thematic similarity metrics to demonstrate that LLM-enhanced pipelines consistently outperformed baseline methods in aligning with human-generated themes. Additionally, three studies evaluated model performance using precision, recall, and F1-scores by treating human coding as the definitive gold standard. Long et al. (2024) observed high coding consistency when applying a customized GPT-4 model to classroom dialogue analysis, although specific F1 values were not detailed in the available abstracts.

**Table 3. Summary Realiability Metrics Reported in Included Studies**

| Study | Year | Cohen's Kappa | Accuracy/Agreement/Interpretation |
|---|---|---|---|
| Li et al. | 2024 | 0.401 | Not reported |
| Jain et al. | 2025 | 0.842-0.907 | 92-95% (semantic) |
| Klieger et al. | 2024 | 0.729 | 78.1% |
| Parkington et al. | 2025 | 0.84 | Not reported |
| Borse et al. | 2025 | >0.6 | Not reported |
| Prescott et al. | 2023 | Not reported | 36-47% |
| Mellon et al. | 2024 | Not reported | 93.9% |
| Liu et al. | 2023 | Not reported | 77.89-96.02% |
| Qiao et al. | 2024 | Not reported | >80% |
| Sakaguchi et al. | 2025 | Not reported | 30-80% (themedependent) |
| Pattyn | 2024 | Improved vs baseline | Not reported |
| Long et al. | 2024 | Not reported | High consistency |

**Factors Influencing Reliability**

A synthesis of current literature reveals critical factors influencing the reliability of large language models (LLMs) in qualitative coding tasks. Foremost is model sophistication. Empirical evaluations show advanced architectures like GPT-4 and Claude 3.5 outperform earlier versions like GPT-3.5 and Claude 1.3 (Klieger et al., 2024; Li et al., 2024). Methodological approaches matter; multi-run ensemble strategies improve coding reliability over single-run outputs (Jain et al., 2025; Li et al., 2024). Beyond architecture, prompting strategy is crucial. Optimized prompts with clear instructions and context-specific examples enhance inter-coder agreement (Borse et al., 2025; Theelen et al., 2024). Few-shot prompting generally outperforms zero-shot approaches (Theelen et al., 2024). While CoT prompting enhances interpretability, it does not universally improve coding reliability (Yue et al., 2024).

The nature of qualitative data and the analytical framework dictate LLM performance. LLMs perform better with descriptive themes than interpretive, culturally nuanced, or emotionally complex constructs (Sakaguchi et al., 2025; Wachinger et al., 2025). Deductive coding with predefined frameworks yields higher reliability than inductive approaches (Pattyn, 2024; Shah et al., 2025). LLMs show higher agreement with shorter, structured texts like SMS messages compared to long, complex interview transcripts (Prescott et al., 2024; Qiao et al., 2024). Linguistic and cultural contexts also limit performance, with lower reliability for non-English or culturally specific content (Sakaguchi et al., 2025). LLM coding reliability is linked to the human comparator used. Studies show lower agreement with expert coders than with novices (Kondo et al., 2024; Parkington et al., 2025). Using multiple human coders for a consensus reference standard provides more robust reliability estimates for AI performance (Raza et al., 2025; Simon et al., 2025).

**Reporting Biases**

The assessment of reporting biases was limited by few studies and outcome heterogeneity. Nonetheless, publication bias is a concern; the field's rapid growth and 40% preprints suggest a publication lag for negative findings. Studies showing higher LLM reliability are more likely published quickly. Selective outcome reporting is evident, with some studies focusing on best-performing models without detailing negative results (Prescott et al., 2024; Zhang et al., 2024). Prompt optimization bias is related, as studies optimizing prompts until a reliability threshold may overestimate AI performance (Borse et al., 2025; Theelen et al., 2024). Small sample bias is present; studies with small samples ($n < 10$) often report higher reliability, likely due to model overfitting or insufficient theme diversity (Li et al., 2024; Yue et al., 2024). A standard funnel plot analysis to quantify these biases was not feasible due to the lack of common metrics and effect size measures.

**Certainty of Evidence**

The certainty of evidence was assessed using adapted GRADE criteria for diagnostic and methodological studies. The certainty was rated as moderate for the primary finding that large language models (LLMs) can achieve moderate to substantial agreement with human coding. Factors increasing this certainty include the robust consistency of findings across studies and contexts, the use of established reliability metrics like Cohen's kappa, and successful replication across different LLM architectures and domains. Conversely, critical factors decrease the overall certainty. Chiefly, there is high heterogeneity in study designs, LLM configurations, and underlying data types. Additionally, there is a moderate risk of bias due to limited prompt reporting and small sample sizes, compounded by a possible publication bias favoring positive findings. The literature also lacks long-term or large-scale validation studies, and the rapid pace of technological change limits the long-term generalizability of these findings. Thus, while current evidence shows LLMs have significant promise for qualitative coding, it remains insufficient to recommend replacing human coders, especially for complex, highly interpretive analytical tasks.

**DISCUSSION**

**Summary of Primary Findings**

This systematic review synthesized evidence from 30 empirical studies comparing generative AI (LLM) coding with human coding of interviews and focus groups. The evidence shows current LLMs can achieve moderate to substantial agreement with human coders in diverse qualitative contexts, with Cohen's kappa ($\kappa$) values from 0.40 to 0.91 (median 0.72) and accuracy rates from 77% to 96%. Several key findings emerged. First, LLMs consistently show reliable performance on many standard coding tasks. Most studies (77%) reported substantial or almost perfect agreement ($\kappa > 0.61$) between LLMs and human coders (Borse et al., 2025; Klieger et al., 2024; Li et al., 2024; Parkington et al., 2025). This suggests that for selected tasks, LLMs can produce codes comparable to those of human researchers, particularly with advanced models like GPT-4, Claude 3.5, and Gemini 2.5 using optimized prompting strategies and multi-run ensemble approaches (Jain et al., 2025; Li et al., 2024).

However, performance varies with task complexity and theme type. LLMs perform well on descriptive themes with clear boundaries but struggle with interpretive, culturally nuanced, or emotionally complex constructs (Sakaguchi et al., 2025; Wachinger et al., 2025). For instance, Sakaguchi et al. (2025) observed that while agreement exceeded 80% for

descriptive themes, it dropped to about 30% for culturally specific themes in Japanese clinical contexts. This pattern across studies underscores a limitation in LLMs' capacity for deep cultural and contextual interpretations (Pattyn, 2024; Zhang et al., 2024).

Furthermore, methodological factors substantially influence overall coding reliability. Prompt engineering has emerged as a critical variable; studies employing optimized, iterative prompting protocols achieved significantly higher agreement than those relying on simple zero-shot approaches (Borse et al., 2025; Theelen et al., 2024). Additionally, multi-run ensemble methods, in which multiple LLM outputs are systematically synthesized, have consistently outperformed traditional single-run approaches (Li et al., 2024). The specific model version also dictates performance, with GPT-4 substantially outperforming its predecessor, GPT-3.5, and the latest iteration of models (e.g., GPT-4o, Claude 3.5 Sonnet, Gemini 2.5 Pro) demonstrating the highest baseline reliability (Klieger et al., 2024; Li et al., 2024).

Despite interpretive limitations, LLMs offer undeniable and substantial efficiency gains. Multiple studies have reported dramatic time savings ranging from 80% to 95% when compared with traditional human coding (Long et al., 2024; Prescott et al., 2024; Qiao et al., 2024). Notably, Qiao et al. (2024) documented an 81% reduction in total coding time while maintaining an accuracy rate above 80%. This profound efficiency advantage suggests that LLMs could effectively democratize qualitative research by drastically reducing resource barriers, provided that rigorous quality assurance protocols remain securely in place (Borse et al., 2025; Sakaguchi et al., 2025).

Ultimately, hybrid human-AI approaches currently show the greatest methodological promise. Several studies have concluded that combining LLM-generated initial codes with rigorous human review and refinement successfully optimizes both efficiency and analytical quality (Nyaaba et al., 2025; Raza et al., 2025; Simon et al., 2025). This collaborative paradigm strategically leverages the speed and consistency of the LLM while preserving the interpretive depth and nuanced contextual understanding of human researchers (Wachinger et al., 2025; Yue et al., 2024).

**Comparison with Existing Literature**

The findings of this systematic review closely align with the broader literature concerning the integration of artificial intelligence in qualitative research, which has historically documented both the distinct opportunities and profound challenges inherent in automating interpretive tasks. Notably, previous reviews evaluating traditional natural language processing (NLP) methods for text analysis have found significantly lower reliability metrics than those reported here for modern large language models (LLMs). This stark contrast suggests that generative AI represents a highly meaningful and substantial methodological advance within the field (Liu & Sun, 2023; Mellon et al., 2024).

When evaluating practical applications, the reliability range observed in this synthesis ($\kappa$ = 0.40–0.91) is highly comparable to standard inter-rater reliability benchmarks reported for human coders in qualitative research, where kappa values between 0.60 and 0.80 are common and generally considered acceptable (Raza et al., 2025; Simon et al., 2025). This equivalency strongly suggests that LLMs have reached a critical threshold of practical utility for many standard coding applications. However, it is imperative to acknowledge that they do not consistently match the nuanced, highly contextualized performance of expert human coders.

Furthermore, our specific finding that LLMs actively struggle with deep cultural nuance closely echoes ongoing concerns raised in critical AI literature. These critiques frequently highlight the inherent limitations of models primarily trained on English-language, Western-centric data corpora, which naturally restrict their cross-cultural applicability (Kim et al., 2025; Sakaguchi et al., 2025). The documented lower performance on highly interpretive themes also strongly aligns with foundational theoretical arguments positing that authentic qualitative analysis requires a level of embodied, lived human understanding that current AI architectures cannot replicate (Noble, 2018; Wachinger et al., 2025).

Finally, the substantial efficiency gains documented in our review are entirely consistent with broader cross-disciplinary trends regarding the AI augmentation of knowledge work. In these paradigms, AI tools are successfully deployed to accelerate routine descriptive tasks, thereby liberating human researchers to focus their cognitive resources on complex analytical judgment (Creswell & Poth, 2016; Long et al., 2024). Nevertheless, this review simultaneously highlights the significant risks associated with over-reliance on AI, particularly the potential degradation of interpretive depth and the vital loss of essential researcher reflexivity (Pattyn, 2024; Wachinger et al., 2025).

**Strengths and Limitations of the Review**

Strengths This systematic review has several notable methodological strengths. Primarily, the study strictly adhered to PRISMA 2020 guidelines, ensuring a transparent and comprehensive framework (Kondo et al., 2024). The team conducted rigorous searches across multiple databases, retrieving over 1,000 records and utilizing AI-assisted screening efficiently. A critical strength is its focus on studies providing quantitative reliability metrics and direct human comparisons, yielding robust evidence addressing the research question. By extracting detailed data on specific LLM configurations, the review enabled nuanced analysis of variables influencing coding reliability. Finally, the systematic risk of bias assessment allows for a contextualized interpretation of findings in light of the methodological quality of the primary literature.

**Limitations**

Despite these strengths, several important limitations must be acknowledged. First, the rapid pace of LLM development means these findings may quickly become outdated as new architectures emerge. The studies in this synthesis mainly evaluated GPT-4 and its predecessors; thus, newer iterations like GPT-4.5, Claude 3.7, or Gemini 2.0 may perform differently. Methodologically, the heterogeneity across study designs, LLM configurations, and qualitative data types precluded a quantitative meta-analysis, limiting the ability to calculate precise overall effect sizes. Furthermore, many studies showed a moderate risk of bias, especially concerning incomplete prompt reporting and small sample sizes, which may skew reliability estimates. This risk is compounded by a likely publication bias favoring positive findings on AI performance.

Culturally and linguistically, most synthesized research was conducted in English-speaking contexts using English data, making generalizability to other languages and cultures uncertain (Sakaguchi et al., 2025). The review did not contact original authors for missing data, potentially limiting data completeness. It focused on interviews and focus groups, so conclusions may not generalize to other qualitative formats, like ethnographic notes or visual data. The review did not systematically assess the quality of human coding used as the

reference standard, which varied significantly across studies, influencing reported reliability estimates.

**Transferability to Educational Contexts**

While only five of the synthesized studies focused strictly on educational settings (Kondo et al., 2024; Lockwood et al., 2024; Long et al., 2024; Theelen et al., 2024), the findings regarding LLM reliability in healthcare and social sciences are highly transferable to education. Qualitative data in education such as student interviews, teacher focus groups, and classroom observations share the same thematic complexity and conversational structure as the clinical and social data analyzed in the majority of the reviewed studies (Borse et al., 2025; Li et al., 2024; Mellon et al., 2024; Pattyn, 2024). The substantial agreement (median $\kappa = 0.72$) found across disciplines suggests that LLMs like GPT-4 are robust enough to assist educational researchers in coding large-scale feedback or identifying descriptive patterns in learning experiences (Theelen et al., 2024; Zhang et al., 2024). Although only a minority of the synthesized studies explicitly targeted educational contexts (Kondo et al., 2024; Lockwood et al., 2024; Long et al., 2024; Theelen et al., 2024), the insights gained from research on LLM reliability in healthcare and social sciences provide valuable implications for education. The qualitative data types examined across these disciplines such as interviews, focus groups, and observational records share core characteristics in thematic richness and conversational dynamics. This similarity indicates that analytical approaches validated in clinical and social research can be effectively adapted to educational research settings, where understanding nuanced learner and teacher experiences is critical. Consequently, the demonstrated robustness of LLMs like GPT-4 in managing complex qualitative datasets supports their potential as scalable tools for educational researchers tasked with coding extensive qualitative feedback or discerning recurring themes in learning environments.

Moreover, the median interrater agreement ($\kappa = 0.72$) reported across multiple disciplines underscores the consistency and reliability with which LLMs can interpret and categorize qualitative data. This level of agreement suggests that LLM-assisted coding can reduce human workload while maintaining analytical rigor, enabling more efficient processing of large-scale educational data such as student reflections, teacher narratives, and classroom interactions. As educational research increasingly incorporates qualitative methodologies to capture the depth of learning experiences, integrating LLMs could enhance both the speed and quality of data analysis. This cross-disciplinary transferability highlights the strategic value of leveraging advancements in AI-driven qualitative analysis beyond their initial healthcare and social science applications to enrich educational inquiry.

**Limitations in Educational Settings**

Despite the promising potential for Large Language Models (LLMs) to enhance research methodologies, their application within educational settings presents distinct challenges that limit their effectiveness. Educational research relies heavily on intricate pedagogical theories and frameworks that demand nuanced interpretive skills, which current LLMs have yet to fully master. This limitation is particularly evident in tasks requiring context-dependent coding, where subtle shifts in cognition or instructional intent may be overlooked by AI systems. Such oversights risk reducing the complexity of educational phenomena to overly simplistic codes, thereby undermining the depth and accuracy essential for meaningful qualitative analysis (Pattyn, 2024; Lockwood et al., 2024; Zhang et al., 2024).

Moreover, classroom dialogues, a fundamental data source in educational technology research, are deeply embedded within specific cultural and social contexts unique to each learning environment. LLMs often struggle to grasp these localized cultural nuances and the emotional subtleties that shape interactions in educational settings. Consequently, AI-generated coding may inadvertently flatten the rich social dynamics and relational intricacies that influence learning processes (Sakaguchi et al., 2025; Kim et al., 2025). In addition to these interpretive challenges, ethical and data privacy concerns are heightened in education due to the sensitive nature of student data and institutional regulations. The reliance on cloud-based LLMs introduces risks of data exposure and conflicts with privacy policies, necessitating a cautious and human-centered hybrid approach. This approach integrates AI assistance with rigorous human oversight to ensure ethical compliance and preserve the integrity of qualitative inquiry (Kondo et al., 2024; Pattyn, 2024; Theelen et al., 2024).

**Implications for Practice**

Based on synthesized evidence, several recommendations emerge for integrating large language models (LLMs) into qualitative coding. LLMs should be strategically used for coding descriptive themes in structured data, like patient symptoms or service experiences (Long et al., 2024; Mellon et al., 2024; Qiao et al., 2024). They are valuable for large-scale tasks where efficiency and defined thematic boundaries are crucial (Liu & Sun, 2023; Mellon et al., 2024). Human oversight is essential, especially for interpretive or complex analyses (Pattyn, 2024; Sakaguchi et al., 2025; Wachinger et al., 2025). Hybrid workflows are recommended, with LLMs generating initial codes for human analysts to review and refine (Nyaaba et al., 2025; Raza et al., 2025; Simon et al., 2025). Methodological rigor is crucial, starting with engineering investments. Coding reliability depends on prompt quality; thus, prompts should be developed and tested with clear instructions and examples, considering advanced approaches (Borse et al., 2025; Theelen et al., 2024). Ensemble methods are recommended, executing LLMs multiple times with variations to improve reliability and reduce variation (Jain, et al., 2025; Li et al., 2024). Researchers must validate LLM performance within their context before scaling. As AI reliability varies with data type and complexity, piloting LLMs on representative data, calculating inter-rater reliability, and ensuring agreement with scholarly standards are imperative before full deployment (Borse et al., 2025; Raza et al., 2025).

Deploying this technology requires considering its ethical implications and a commitment to transparent reporting. LLMs raise urgent questions about data privacy, especially with sensitive interview transcripts, and concerns about algorithmic bias and changing qualitative research expertise (Kim et al., 2025; Pattyn, 2024). Researchers must secure ethical approvals, ensure data security, and reflect on how AI tools shape research practice and knowledge production (Noble, 2018; Wachinger et al., 2025). To ensure accountability, reproducibility, and support knowledge building, studies must adhere to transparent methodological reporting standards. Researchers must document the LLM model and version, prompts and parameters used, computational runs executed, and reliability metrics achieved, aligning with AI reporting guidelines in qualitative research (Kondo et al., 2024).

**Implications for Future Research**

Building on the current evidence, critical priorities for future research emerge. Foremost is the need for large-scale, longitudinal validation studies. As most included literature relies on small qualitative samples with a median of 28 transcripts, expansive research with diverse

data and broader participant populations is necessary to establish true generalizability and boundary conditions of LLM reliability (Liu & Sun, 2023; Mellon et al., 2024). Coupled with this is a mandate for cross-cultural and multilingual research. With only one included study on non-English data (Sakaguchi et al., 2025), future investigations must evaluate LLM performance across languages, cultural frameworks, and global contexts, especially concerning underrepresented populations (Kim et al., 2025). Furthermore, the field requires comparative studies. Systematic evaluations contrasting multiple LLM architectures using identical datasets and methodologies are needed to clarify which models are best for different qualitative tasks (Li et al., 2024; Prescott et al., 2024).

Future research should explore AI-assisted analysis mechanisms and human-AI collaboration optimization. Investigating LLM success or failure in coding tasks by examining attention patterns, reasoning, and failure modes will inform better prompt designs and usage parameters (Yi et al., 2025; Yue et al., 2024). This understanding is crucial for optimizing hybrid workflows. Although collaborative human-AI approaches show promise, the best workflow configuration is unclear. Future studies should compare division-of-labor strategies, like LLM versus human first-pass coding, or parallel versus sequential coding frameworks (Nyaaba et al., 2025; Raza et al., 2025; Simon et al., 2025).

Epistemological and ethical impacts of this technology require rigorous scholarly attention. Future research should examine how LLM integration affects the depth, creativity, and sophistication of qualitative analysis, and its impact on researcher skill development and qualitative inquiry epistemology (Noble, 2018; Wachinger et al., 2025). Developing standardized evaluation frameworks is urgent. Consensus on reliability metrics, open-source datasets, and evaluation protocols will align qualitative AI with computational testing standards (Kondo et al., 2024; Sankaranarayanan et al., 2025). Methodological maturation must include investigations into ethical implications, addressing data privacy, algorithmic bias, research equity, and the evolving role of qualitative researchers in an AI-augmented landscape (Kim et al., 2025; Pattyn, 2024).

## CONCLUSION

This systematic review synthesizes empirical evidence on the reliability of large language models (LLMs) for qualitative coding of interviews and focus groups. Based on 30 studies from 2023 to 2025, evidence shows LLMs achieve moderate to substantial reliability, with Cohen's kappa values of 0.40 to 0.91 and accuracy rates of 77% to 96%. AI-human agreement is comparable to standard inter-rater reliability among human coders. However, performance varies significantly with task characteristics. LLMs excel at analyzing descriptive themes in structured data but struggle with interpretive, culturally nuanced, or emotionally complex constructs. Reliability is maximized using advanced architectures like GPT-4, Claude 3.5, and Gemini 2.5, with optimized prompting and multi-run ensemble approaches.

Beyond baseline reliability, LLM integration offers substantial practical benefits, reducing coding time by 80% to 95%. These savings can democratize qualitative research by lowering resource barriers but must be balanced against quality considerations. Human oversight remains essential, as current evidence does not support completely replacing human coders. Hybrid human-AI workflows that leverage computational efficiency while preserving human interpretive depth are the most methodologically sound.

Methodological rigor is critical for implementing these hybrid systems. Researchers must pilot LLM coding in specific contexts and report methodologies transparently for reproducibility. The evidence base, though rapidly growing, has structural limitations,

including design heterogeneity, moderate risk of bias, potential publication bias, and a lack of non-English research contexts. The overall certainty of current evidence is rated as moderate, highlighting the need for rigorous methodological investigation.

**Future Directions**

As LLM technology evolves, several research priorities have emerged. First, standardized evaluation frameworks and benchmark datasets are needed for rigorous assessments of qualitative coding. Second, researchers must conduct cross-cultural validation studies across diverse populations and languages to ensure global applicability. Third, investigation is required to optimize hybrid human-AI workflows, identifying the division of labor between computational speed and human interpretive depth. Fourth, the academic community must address ethical implications, including data privacy, algorithmic bias, and impacts on research integrity. Additionally, longitudinal studies are essential to track how coding reliability shifts with new model architectures. Finally, future explorations should investigate LLM integration in higher-level qualitative tasks, such as analytical memo writing and complex theory development.

**Final Remarks**

Integrating generative AI into qualitative research is a methodological crossroads, offering extraordinary opportunities and challenges. Large language models (LLMs) can accelerate qualitative analysis, reduce resource barriers, and enable large-scale studies previously infeasible due to time and cost constraints. However, these tools raise questions about interpretive inquiry, human judgment, and the epistemological foundations of qualitative research.

Empirical evidence in this review suggests LLMs have reached practical utility for specific qualitative coding applications, especially when used to augment, not replace, human expertise. As technology advances, the qualitative research community must engage critically with these systems. Establishing standardized practices, ethical guidelines, and quality criteria is imperative to preserve the rigor and interpretive depth of high-quality qualitative inquiry.

Researchers, methodologists, journal editors, and funding agencies share the responsibility of integrating AI into social sciences responsibly. By combining technological innovation with methodological rigor and ethical reflection, the field can harness LLMs' benefits while safeguarding the core values of qualitative research. This review provides an evidence-based foundation for these conversations. As new studies emerge and LLM capabilities evolve, regular updates to this synthesis will guide the field's trajectory through this changing methodological landscape.

**REFERENCES**

Borse, N. S., Subramaniam, R. C., & Rebello, N. S. (2025). *Investigation of the Inter-Rater Reliability between Large Language Models and Human Raters in Qualitative Analysis* (Version 2). arXiv.https:/doi.org/10.48550/ARXIV.2508.14764

Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.

Jain, N., Suh, H., Adeyinka, S., Roseman, L., & Allsop, A. (2025). *Multi-LLM Thematic Analysis with Dual Reliability Metrics: Combining Cohen's Kappa and Semantic Similarity for Qualitative Research Validation* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2512.20352

Kim, C., Ke, F., Zhang, N., & Barrett, A. (2025). *LLM-supported Thematic Analysis: Evaluating GATOS Workflow on Complex Qualitative Data*. https://doi.org/10.5281/ZENODO.15870242

Klieger, B., Charitsis, C., Suzara, M., Wang, S., Haber, N., & Mitchell, J. C. (2024). *ChatCollab: Exploring Collaboration Between Humans and AI Agents in Software Teams* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2412.01992

Kondo, T., Miyachi, J., Jönsson, A., & Nishigori, H. (2024). *A mixed-methods study comparing human-led and ChatGPT-driven qualitative analysis in medical education research* (No. 4). Nagoya University Graduate School of Medicine, School of Medicine. https://doi.org/10.18999/nagjms.86.4.620

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Li, K. D., Fernandez, A. M., Schwartz, R., Rios, N., Carlisle, M. N., Amend, G. M., Patel, H. V., & Breyer, B. N. (2024). Comparing GPT-4 and Human Researchers in Health Care Data Analysis: Qualitative Description Study. *Journal of Medical Internet Research*, *26*, e56500. https://doi.org/10.2196/56500

Liu, A., & Sun, M. (2023). *From Voices to Validity: Leveraging Large Language Models (LLMs) for Textual Analysis of Policy Stakeholder Interviews* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2312.01202

Lockwood, A., Newman, D., Mossing, K., Glubzinski, A., & Cohen, E. (2025). *Human vs. Machine: A Comparative Analysis of Qualitative Coding by Humans and ChatGPT-4*. PsyArXiv. https://doi.org/10.31234/osf.io/8g36r

Long, Y., Luo, H., & Zhang, Y. (2024). Evaluating large language models in analysing classroom dialogue. *Npj Science of Learning*, *9*(1), 60. https://doi.org/10.1038/s41539-024-00273-3

Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., Miori, M., & Schmedeman, P. (2024). Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale. *Research & Politics*, *11*(1), 20531680241231468. https://doi.org/10.1177/20531680241231468

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. In *Algorithms of oppression*. New York university press.

Nyaaba, M., SungEun, M., Apam, M. A., Acheampong, K. O., & Dwamena, E. (2025). *Optimizing Generative AI's Accuracy and Transparency in Inductive Thematic Analysis: A Human-AI Comparison* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2503.16485

Parkington, K., Teferra, B. G., Rouleau-Tang, M., Perivolaris, A., Rueda, A., Dubrowski, A., Kapralos, B., Samavi, R., Greenshaw, A., Zhang, Y., Cao, B., Wu, Y., Rambhatla, S., Krishnan, S., & Bhat, V. (2025). *Human vs. LLM-Based Thematic Analysis for Digital Mental Health Research: Proof-of-Concept Comparative Study* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2507.08002

Pattyn, F. (2024). The Value of Generative AI for Qualitative Research: A Pilot Study. *Journal of Data Science and Intelligent Systems*. https://doi.org/10.47852/bonviewJDSIS4202964

Prescott, M. R., Yeager, S., Ham, L., Rivera Saldana, C. D., Serrano, V., Narez, J., Paltin, D., Delgado, J., Moore, D. J., & Montoya, J. (2024). Comparing the Efficacy and Efficiency of Human and Generative AI: Qualitative Thematic Analyses. *JMIR AI*, *3*, e54482. https://doi.org/10.2196/54482

Qiao, S., Fang, X., Wang, J., Zhang, R., Li, X., & Kang, Y. (2024). *Generative AI for Thematic Analysis in a Maternal Health Study: Coding Semi-structured Interviews using Large Language Models (LLMs)*. Public and Global Health. https://doi.org/10.1101/2024.09.16.24313707

Raza, M. Z., Xu, J., Lim, T., Boddy, L., Mery, C. M., Well, A., & Ding, Y. (2025). *LLM-TA: An LLM-Enhanced Thematic Analysis Pipeline for Transcripts from Parents of Children with Congenital Heart Disease* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2502.01620

Sakaguchi, K., Sakama, R., & Watari, T. (2025). *Evaluating ChatGPT in Qualitative Thematic Analysis With Human Researchers in the Japanese Clinical Context and Its Cultural Interpretation Challenges: Comparative Qualitative Study (Preprint)*. Journal of Medical Internet Research. https://doi.org/10.2196/preprints.71521

Sankaranarayanan, S., Borchers, C., Simon, S., Tajik, E., Ataş, A. H., Celik, B., Balzan, F., & Shahrokhian, B. (2025). *Automating Thematic Analysis with Multi-Agent LLM Systems*. EdArXiv. https://doi.org/10.35542/osf.io/kq8zh_v1

Shah, S. T. U., Hussein, M., Barcomb, A., & Moshirpour, M. (2025). *From Inductive to Deductive: LLMs-Based Qualitative Data Analysis in Requirements Engineering* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2504.19384

Simon, S., Sankaranarayanan, S., Tajik, E., Borchers, C., Shahrokhian, B., Balzan, F., Strauß, S., Viswanathan, S. A., Ataş, A. H., Čarapina, M., Liang, L., & Celik, B. (2025). *Comparing a Human's and a Multi-Agent System's Thematic Analysis: Assessing Qualitative Coding Consistency*. EdArXiv. https://doi.org/10.35542/osf.io/ez8wc_v1

Theelen, H., Vreuls, J., & Rutten, J. (2024). Doing Research with Help from ChatGPT: Promising Examples for Coding and Inter-Rater Reliability. *International Journal of Technology in Education*, *7*(1), 1–18. https://doi.org/10.46328/ijte.537

Wachinger, J., Bärnighausen, K., Schäfer, L. N., Scott, K., & McMahon, S. A. (2025). Prompts, Pearls, Imperfections: Comparing ChatGPT and a Human Researcher in Qualitative Data Analysis. *Qualitative Health Research*, *35*(9), 951–966. https://doi.org/10.1177/10497323241244669

Yi, S., Nguyen, J., Xu, H., Lim, T., Skrovan, J., Beri, M., Modi, H., Well, A., Leqi, L., Markey, M., & Ding, Y. (2025). *SFT-TA: Supervised Fine-Tuned Agents in Multi-Agent LLMs for Automated Inductive Thematic Analysis* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2509.17167

Yue, Y., Liu, D., Lv, Y., Hao, J., & Cui, P. (2024). *A Practical Guide and Assessment on Using ChatGPT to Conduct Grounded Theory: Tutorial (Preprint)*. Journal of Medical Internet Research. https://doi.org/10.2196/preprints.70122

Zhang, H., Wu, C., Xie, J., Rubino, F., Graver, S., Kim, C., Carroll, J. M., & Cai, J. (2024). *When Qualitative Research Meets Large Language Model: Exploring the Potential of QualiGPT as a Tool for Qualitative Coding* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2407.14925