South American Publishing

SAP Netnography. 2026;4:270
doi: 10.62486/net2026270

SYSTEMATIC REVIEW

AI in Qualitative Digital Research: A Systematic Mapping Review of Methodological Rigor, Ethics, and Human-AI Collaboration with Implications for Netnography and Digital Ethnography

IA en la investigación digital cualitativa: una revisión sistemática de mapeo sobre el rigor metodológico, la ética y la colaboración humano-IA, con implicaciones para la netnografía y la etnografía digital

Maicol Stiven Vanegas-Nieto[1,2]
Sofia Belen Castro Brenes[3]

[1] Instituto Internacional de Estudios Políticos Avanzados Ignacio Manuel Altamirano de la Universidad Autónoma de Guerrero, México.
[2] Universidad del Valle, Colombia.
[3] Universidad Autónoma del Estado de Morelos, Nicaragua.

Submitted: 15-04-2026 | Revised: 08-06-2026 | Accepted: 10-08-2026 | Published: 12-08-2026

Corresponding Author: Maicol Stiven Vanegas-Nieto

ABSTRACT

Artificial intelligence (AI), including large language models, machine learning, automated coding, sentiment analysis, topic modelling, and pattern detection, is increasingly incorporated into qualitative digital research. This systematic mapping review examines how these tools are integrated into netnography and digital ethnography, which tasks are delegated to AI, and which remain under human interpretive control. A total of 793 records were identified, 29 duplicates were removed or truncated, and 764 records were screened. After screening, 242 reports were sought, 96 were assessed for eligibility, and 57 were included in the synthesis. Evidence published between 2020 and 2026 shows a recurring division of labour: AI is mainly used for transcription, initial coding, code suggestion, topic modelling, summarization, pattern detection, and preliminary classification, while researchers retain responsibility for contextual interpretation, theoretical integration, reflexivity, ethical judgement, and final meaning-making. Validation practices include Cohen's kappa, F1 scores, cosine similarity, human review, and triangulation. Reported failures include hallucination, stochastic instability, context-window truncation, error propagation, category bias, and plausible but ungrounded outputs. Although privacy, consent, anonymization, and AI-use disclosure are frequently reported, standardized ethical frameworks, audit trails, and model-version reporting remain underdeveloped. The review proposes testable propositions concerning task allocation, contextual grounding, human interpretive authority, validation, ethical governance, reflexivity, transparency, and auditability. Limitations include single-reviewer screening, database coverage, inaccessible full texts, heterogeneous study designs, absence of formal quality appraisal, and a Scopus-indexed source share below the protocol target. These findings support a human-led approach in which computational assistance complements, rather than replaces, culturally grounded interpretation and methodological accountability.

Keywords: AI-augmented netnography; digital ethnography; large language models; qualitative coding; human-AI collaboration; reflexivity; research ethics; auditability; thick description; methodological rigor; systematic mapping review

RESUMEN

La inteligencia artificial (IA), incluidos los modelos de lenguaje de gran escala, el aprendizaje automático, la codificación automatizada, el análisis de sentimientos, el modelado de temas y la detección de patrones, se incorpora a la investigación digital cualitativa. Esta revisión sistemática de mapeo examina cómo estas herramientas se integran en la netnografía y la etnografía digital, qué tareas se delegan a la IA y cuáles permanecen bajo control interpretativo humano. Se identificaron 793 registros, se eliminaron o depuraron 29 duplicados y se examinaron 764. Tras el cribado, se buscaron 242 informes, 96 fueron evaluados para determinar su elegibilidad y 57 se incluyeron en la síntesis. La evidencia publicada entre 2020 y 2026 muestra una división del trabajo: la IA se utiliza para transcripción, codificación inicial, sugerencia de códigos, modelado de temas, síntesis, detección de patrones y clasificación preliminar, mientras los investigadores conservan la responsabilidad sobre la interpretación contextual, integración teórica, reflexividad, juicio ético y construcción final de significado. La validación incluye kappa de Cohen, puntuaciones F1, similitud del coseno, revisión humana y triangulación. Los fallos reportados incluyen alucinaciones, inestabilidad estocástica, truncamiento, propagación de errores, sesgos y resultados plausibles sin fundamento. Aunque se reportan privacidad, consentimiento, anonimización y declaración del uso de IA, siguen poco desarrollados los marcos éticos, registros de auditoría y reporte de versiones. Se proponen proposiciones comprobables sobre asignación de tareas, anclaje contextual, autoridad humana, validación, gobernanza ética, reflexividad, transparencia y auditabilidad. Las limitaciones incluyen cribado por un único revisor, cobertura, textos inaccesibles, heterogeneidad y ausencia de evaluación.

Palabras clave: netnografía aumentada mediante IA; etnografía digital; modelos de lenguaje de gran escala; codificación cualitativa; colaboración humano-IA; reflexividad; ética de la investigación; auditabilidad; descripción densa; rigor metodológico; revisión sistemática de mapeo.

© 2026; The Authors. This Open access article is distributed under CC BY 4.0 licence.

INTRODUCTION

Netnography and digital ethnography are interpretive research traditions concerned with online communities, digitally mediated sociality, cultural meaning, and the situated practices through which people produce and contest social worlds. They depend on sustained engagement, contextual understanding, and the researcher's reflexive positioning within the field. Over the past several years, artificial intelligence has entered this methodological space not simply as a topic to be studied but as a set of tools that can be used for data collection, transcription, coding, thematic analysis, pattern detection, summarization, and even the generation of synthetic research material(1,2).

Large language models have made some of these capabilities newly accessible to qualitative researchers who do not possess advanced computational skills, while machine learning and natural language processing have extended the scale and speed with which digital traces can be organized(3,4).

This methodological shift raises foundational questions. If AI systems can produce codes, themes, summaries, or conversational interviews, what exactly is being delegated, and with what consequences for interpretive validity? How do researchers preserve thick description, cultural context, and emic meaning when computational systems operate through patterns derived from large, often decontextualized corpora? What forms of validation, transparency, reflexivity, bias control, privacy protection, and auditability are being reported? And what reproducible models of human-AI collaboration can be identified across the emerging literature?

The existing evidence is fragmented. Systematic reviews and mapping studies have begun to survey AI use in qualitative research, but they often cover broad disciplinary fields rather than netnography and digital ethnography specifically(5-7). Studies of AI-assisted thematic analysis frequently use interview or survey data rather than online community data(8,9).

Conversely, computational ethnographies of online communities sometimes treat AI as a technical pipeline rather than as a methodological collaborator with ethical and reflexive consequences(4,10). There is therefore a need to synthesize evidence at the intersection of AI and qualitative digital research, while keeping the interpretive and methodological focus central and being explicit that much of the included evidence concerns qualitative analysis more broadly rather than netnography proper.

This review addresses that gap. Its research question is: how is artificial intelligence, including large language models, machine learning, automated coding, sentiment analysis, and pattern detection, integrated into qualitative digital research with implications for netnography and digital ethnography; what tasks are delegated to AI and what remains under human interpretive control; how are cultural context, thick description, validation, transparency, reflexivity, bias control, privacy protection, and auditability reported; and what reproducible models of human-AI collaboration and methodological failure modes can be identified?

The review is organized around four domains: AI task delegation; human interpretive control; ethical governance; and transparency, validation, and auditability. It concludes by proposing a set of conditional propositions for rigorous AI-augmented netnography that is derived from the included evidence and is explicitly advanced as a synthesis of heterogeneous, unappraised reports rather than as a validated instrument.

Background and Conceptual Framework

Netnography and digital ethnography are not simply qualitative methods applied online. They involve the adaptation of ethnographic sensibilities to digitally mediated environments, including attention to platform affordances, community norms, insider language, temporality, visibility, and the ethical ambiguities of public or semi-public data(10,4).

AI enters this space in several ways. It can be used as a computational scaffold for reading large corpora, as a coding assistant, as a pattern detector, as a translation or transcription tool, as a simulator of participants or personas, and as an analytic interlocutor that produces interpretations for human review(3,11,12).

The first conceptual domain is AI task delegation. The literature distinguishes between tasks that are primarily mechanical or classificatory and tasks that require conceptual synthesis. AI has been used for transcription, language translation, initial open coding, code suggestion, deductive classification, topic modelling, sentiment analysis, clustering, summarization, and the generation of interview follow-up questions(7,12,13).

In several studies, AI is framed as a suggestion provider rather than an autonomous coder(14,15). In others, multi-agent systems are designed to perform most steps of thematic analysis

with minimal human intervention(16,17).

The second domain is human interpretive control. Qualitative research depends on contextual judgement, theoretical sensitivity, and the ability to recognize irony, silence, ambiguity, and culturally specific meaning (18,19). Several studies report that AI can produce plausible codes or themes but struggles with higher-level interpretation, contextual nuance, and the integration of findings into a theoretical narrative(20,21). This has led to repeated calls for human-in-the-loop or human-in-the-lead models in which AI supports discrete tasks while interpretive authority remains with the researcher(22,23).

The third domain is ethical governance. AI-augmented research raises questions about informed consent, data privacy, anonymization, platform terms of service, cross-border data transfer, proprietary model retention, bias, and the risk of harm to already marginalized groups(2,24,25).

Digital ethnographic data are often identifiable, relational, and collected in contexts where consent is difficult to obtain in the conventional sense. The introduction of third-party AI services can complicate those problems further, especially when data are sent to external APIs(26,27).

The fourth domain is transparency, validation, and auditability. Because AI outputs are probabilistic and model versions change, reproducibility is challenging(1). Researchers have responded with a variety of validation strategies, including human review, inter-coder agreement, triangulation with a second model, comparison against a gold standard, and the collection of audit trails(28,23,29). Yet there is no settled standard for reporting AI use in qualitative research, and reviewers may lack the expertise to assess it(5,22).

These four domains are not independent. Decisions about task delegation shape what kinds of validation are possible; validation practices shape what can be claimed about interpretive control; ethical governance depends on transparency about data flows and model choices; and auditability depends on the granularity with which human and machine contributions are recorded. The synthesis below therefore treats them as cross-cutting concerns rather than as separate checklists.

METHODS

Review type and protocol

This is a systematic mapping review. The protocol specified a systematic design with a defined research question, eligibility criteria, search concepts, search queries, date range, and planned quality appraisal. The workflow metadata contain reproducible search and screening information, including databases, queries, record counts, and full-text retrieval outcomes.

The protocol followed PRISMA-style reporting guidance, but the completed review does not meet PRISMA requirements for dual independent screening, reviewer-level agreement statistics, or completed quality appraisal. Screening and selection were conducted by a single reviewer in the completed workflow, which is a deviation from the protocol's stated plan for dual independent screening. The review is therefore reported as a systematic mapping review rather than as a PRISMA-compliant systematic review.

The protocol stated that dual independent screening and data extraction would be used, with disagreements resolved by discussion or a third reviewer, and that quality appraisal would use CASP, MMAT, or JBI tools depending on study design.

The workflow metadata report the aggregate screening and retrieval counts but do not include reviewer-level agreement statistics or completed appraisal scores. The evidence library contains an internal confidence value for each record, but this is not a substitute for formal quality appraisal and is not used as evidence quality in this review.

Search strategy

Searches were run on 15 March 2026 in Scopus, Web of Science, SciELO, OpenAlex and Semantic Scholar. Ten Boolean queries were attempted in each database, and all twenty queries completed without recorded errors.

The queries combined terms for AI-augmented netnography, computational and digital ethnography, large language models in qualitative research, automated coding, human-AI collaboration, ethics and privacy, reflexivity and transparency, sentiment analysis and pattern detection, and specific intersections of AI with netnography, digital ethnography, online ethnography, qualitative coding, ethics, reflexivity, and rigor.

The ten queries were: "AI augmented netnography" OR "artificial intelligence netnography"

OR “LLM netnography” OR “large language models netnography”; “AI digital ethnography” OR “computational ethnography” OR “automated digital ethnography” OR “machine learning digital ethnography”; “large language models qualitative research” OR “LLM qualitative coding” OR “AI-assisted qualitative data analysis” OR “ChatGPT qualitative research”; “automated coding qualitative research” OR “machine learning qualitative coding” OR “AI thematic analysis” OR “natural language processing qualitative analysis”; “human-AI collaboration qualitative research” OR “human in the loop ethnography” OR “augmented qualitative research” OR “AI-assisted ethnography”; “AI research ethics digital ethnography” OR “privacy netnography” OR “algorithmic bias qualitative research” OR “ethical AI ethnography”; “reflexivity AI qualitative research” OR “transparency AI qualitative analysis” OR “auditability AI ethnography” OR “thick description AI”; “sentiment analysis netnography” OR “pattern detection digital ethnography” OR “social media ethnography AI” OR “online ethnography artificial intelligence”; “AI qualitative analysis” AND (“netnography” OR “digital ethnography” OR “online ethnography”); and “large language model” AND (“ethnography” OR “netnography” OR “qualitative coding”) AND (“ethics” OR “reflexivity” OR “rigor”). The date range was 2020–2026.

Eligibility criteria

Inclusion criteria were peer-reviewed journal articles, book chapters, conference papers, or systematic reviews; publication from 2020 to 2026, with seminal older studies included only through citation chasing or hand searching; a focus on netnography, digital ethnography, or qualitative online research with explicit ethnographic or interpretive elements; explicit discussion of AI, machine learning, large language models, automated coding, sentiment analysis, pattern detection, or related computational tools in qualitative digital research; attention to at least one of methodological rigor, task delegation, human-AI collaboration, ethics, reflexivity, bias, privacy, transparency, auditability, thick description, or cultural context; English language; and sufficient methodological detail for quality appraisal and data extraction.

Exclusion criteria were purely technical AI or machine learning papers without qualitative, ethnographic, or netnographic application; social media analytics, sentiment analysis, or computational social science studies without an interpretive or netnographic framing; studies that treat AI only as a substantive topic rather than as a methodological tool or collaborator; opinion pieces, editorials, or commentaries lacking methodological, ethical, or reflexive analysis relevant to AI-augmented netnography; purely quantitative content analysis without a human interpretive or ethnographic component; non-English publications; duplicates, retracted articles, and sources from predatory venues.

Screening and selection

The workflow recorded 793 identified records. After removal or truncation of 29 duplicates, 764 records were screened. Title and abstract screening excluded 520 records, and two records remained uncertain after a second pass. Reports were sought for 242 records, but 146 could not be retrieved, leaving 96 reports assessed for eligibility. No reports were recorded as unreadable by automation. Of the 96 reports assessed, 20 were excluded at full text, 19 remained uncertain at full text and were kept out of the automatic synthesis, and 57 were included.

Data extraction and synthesis

Data extraction covered study design, objective, population or corpus, sample size, setting, methods, AI tools and interventions, outcomes, key findings, effect estimates where reported, limitations, relevance to the review, evidence basis, and confidence. The synthesis was thematic and organized around the four conceptual domains described above. Because the included studies are methodologically heterogeneous, the synthesis does not pool effect sizes. Quantitative metrics are reported where they were supplied in the evidence records, but they are treated as indicators of reported performance rather than as directly comparable estimates.

Quality appraisal

The protocol planned CASP, MMAT, or JBI appraisal depending on study design. The workflow metadata do not contain completed appraisal scores. The evidence library does contain confidence values ranging from approximately 0.78 to 1.0, but these are internal evidence-record confidence ratings rather than formal quality appraisal outcomes. They are not used as quality appraisal in this review. No synthesis claim should be read as implying appraised evidence quality. All synthesis claims are based on unappraised reports. All records used for substantive claims in this review are

marked as full-text evidence in the supplied library. The synthesis distinguishes between full-text evidence and abstract-only evidence where relevant, and it avoids detailed claims from records that are not full text.

RESULTS

Descriptive Overview of Included Studies

The 57 included records span 2020–2026. One study was published in 2020(4), one in 2021(31), ten in 2023, twelve in 2024, seventeen in 2025, and sixteen in 2026. The distribution confirms the protocol's expectation that generative AI would substantially change the literature from 2023 onward. The included studies cover methodological framework development, comparative coding experiments, system design and user evaluation, interview studies with qualitative researchers, conceptual analyses, computational ethnographies, and systematic or quasi-systematic reviews.

Several records are directly concerned with thematic analysis or inductive coding. These include the LLM-in-the-loop framework for thematic analysis(8), the ChatGPT prompt-design study(9), the exploration of GPT-3.5 on inductive thematic analysis(32), the comparison of human experts and LLMs in open coding(33), CollabCoder(15), the Guided AI Thematic Analysis framework(11), the augmented qualitative researcher model(20), and the multi-agent AutoTheme framework (16). Others focus on qualitative data analysis workflows, validation, or tool design(23,34,35).

A second cluster concerns computational ethnography, digital ethnography, and online community analysis. These include the COMETH project on generating ethnographic models from online data(4), the computational ethnography of data science on Twitter(10), the Reddit mental health analysis using LLMs and topic modelling(28), the AI-driven documentation of East Sumba ikat traditions(36), the digital ethnography of an organizational LLM deployment(37), and the LLM-assisted reflexive thematic analysis of AI chatbot risk discourse on Reddit(38). The chatbot risk discourse study is included because it integrates LLM-assisted reflexive thematic analysis into a qualitative digital ethnography, not because of its substantive topic of AI chatbot risk. These studies are especially relevant to netnography because they work with online communities and digital traces, although not all use the term netnography.

A third cluster examines AI in qualitative research practice more broadly, including researcher perceptions and ethical concerns(2), low-income African contexts(24), engineering education quality frameworks(18), higher education faculty perceptions(35), and sociological and marketing research(19). Several reviews map AI use across qualitative analysis and identify reporting and validation gaps(5-7,21). A smaller set of studies is explicitly about prompt engineering, LLM agents, or multi-agent consensus in coding (17,39-41).

The evidence is heterogeneous in discipline, data type, and AI tool. Coding studies use GPT-3.5, GPT-4, GPT-4o, Claude, Gemini, Llama, Mistral, Falcon, BERT, BERTopic, and other open-source models. Digital ethnographies use NLP, computer vision, topic modelling, sentiment analysis, knowledge graphs, and LLM-assisted thematic analysis. This diversity is a strength for mapping the field but a limitation for cumulative validation. A substantial share of the included records are thematic-analysis, qualitative-coding, or software-engineering studies without an explicit netnographic or digital-ethnographic framing. The review therefore synthesizes evidence on AI in qualitative digital research broadly, with implications for netnography and digital ethnography, rather than claiming to be a review of AI-augmented netnography proper.

AI Tasks Delegated in Qualitative Digital Research

Across the included evidence, AI is delegated a recognizable set of tasks. The most common are transcription and translation, initial or open coding, code suggestion, deductive classification, topic modelling, sentiment or emotion labelling, pattern detection, summarization, and first-pass thematic grouping. AI is also used for data collection through chatbot interviewing, for generating personas or synthetic respondents, and for prompt refinement.

Transcription and translation are among the most widely adopted uses. A quasi-systematic review of AI-supported interviews reports that transcription tools such as Whisper and ChatGPT reduce time and cost with near-human accuracy in structured contexts, though manual correction is still required for jargon, informal speech, and non-standard accents(7). The same review reports word error rates of 2.5% to 3.36% in adult transcription studies and sub-1% in one ChatGPT refinement study, but notes that such metrics do not capture semantic accuracy or interpretive

suitability(7). Qualitative researchers interviewed about LLM use were comfortable with task-specific tools such as grammar check, speech-to-text transcription, and translation, and used them extensively, while remaining wary of LLMs for complex interpretive work(2).

Coding is the most frequently studied delegated task. The LLM-in-the-loop framework delegates initial code generation, code refinement, and theme identification to GPT-3.5-turbo while retaining human dialogue and final decisions; it reports almost perfect agreement between human and machine coders on two datasets, with Cohen's kappa of 0.87 and 0.81, and high cosine similarity to the original authors' codes(8). The human-expert versus LLM comparison for inductive open coding found that fine-tuning with as few as 100 examples could achieve sufficient performance, that Falcon and Mistral performed best when fine-tuned, and that performance plateaued after approximately 100 training examples(33). Qualitative code suggestion has been framed as a human-centric alternative to automatic coding, with information-retrieval and zero-shot prompting approaches achieving the highest ranking scores while novel-code detection remained difficult(14).

Deductive coding studies show a similar pattern. A comparison of hierarchical and direct prompting for coding communication data found that direct prompting achieved the highest agreement with human coding, with Cohen's kappa of 0.591, comparable to human-human agreement of 0.582; two-step hierarchical prompting was vulnerable to error propagation from main-category misclassification to subcategory assignment(41). A controlled study of psychological safety coding in software engineering communities found that all three tested models achieved fair to moderate agreement with a human gold standard, with kappa values between 0.33 and 0.44, and that multi-shot prompting significantly improved agreement for one model only(40). Another comparison found that LLM coding aligned closely with human coding for systematic review data, with final Cohen's kappa of 0.899 for study aims and 0.823 for discussions(42).

Thematic analysis is frequently delegated at the level of code and theme generation. ChatGPT prompt design improved perceived usefulness, transparency, and trust among qualitative researchers, but participants still emphasized human oversight and double-checking(9). GPT-3.5-Turbo was able to infer most main themes in one inductive thematic analysis, though it failed to infer some themes valued by human analysts and produced hallucinations in later phases(32). GPT-4 generated codes, subcodes, clusters, and themes in a Guided AI Thematic Analysis workflow, but initial themes often resembled clusters and lacked participant-connected contextual meaning (11). A multi-agent framework produced more detailed design-feature themes than LDA topic modelling, with 14 themes versus 6, but required substantial computational resources and did not examine error propagation across agents (16).

Pattern detection and topic modelling are commonly delegated to AI in digital ethnographic work. BERTopic was ranked first by 8 of 12 qualitative researchers and produced higher topic coherence and diversity than LDA and NMF, though its large number of topics could be overwhelming and it lacked hierarchical visualization in the tested toolkit(13). Computational ethnography has used topic modelling, network community detection, and weighted log odds ratios to study data science on Twitter, with the researcher retaining control over boundary specification and interpretation(10). The COMETH project used automated metaphor extraction, clustering, and sentiment analysis to model community worldviews from online language, with human validation studies used to check whether metaphor-based models distinguished communities(4). The Reddit mental health study combined GPT-3.5-turbo coding with BERTopic clustering and human validation in a human-in-the-loop pipeline(28).

AI is also being delegated tasks at the front end of research. Chatbot-based qualitative data collection used LLM modules for dynamic probing and member checking, but the study found that LLM chatbots achieved high-quality conversations on established communication metrics while rarely capturing participants' specific motives or personalized examples(12). Synthetic participants and personas have been generated for interviews(43)and for writing up thematic results(44). Prompt engineering itself has become a delegated or semi-delegated task, with frameworks proposing structured prompts, role-play, few-shot examples, chain-of-thought, retrieval-augmented generation, and hierarchical prompt systems(11,45,19).

The overall picture is not that AI replaces the researcher uniformly. Instead, delegation clusters around tasks that can be framed as classification, retrieval, summarization, or suggestion. Tasks requiring theoretical synthesis, contextual judgement, ethical reasoning, and the construction of a coherent interpretive narrative are repeatedly retained by humans, even in studies that automate substantial portions of coding(16,46).

Human Interpretive Control, Cultural Context, and Thick Description

The included evidence converges on the conclusion that human interpretive control remains central, but the reasons differ. Some studies emphasize semantic and pragmatic limits: LLMs can assist with coding but cannot grasp the semantic and pragmatic aspects of data, and their outputs are surface-level and lack granularity(18). Others emphasize contextual immersion: AI can identify patterns but lacks the lived experience, positionality, and accountability that ethnographic interpretation requires(29,47). Still others emphasize theoretical integration: AI may produce plausible themes, but it struggles to connect them to a theoretical framework or to explain why they matter(21,48).

The concept of thick description is directly challenged by AI's tendency to smooth, summarize, or decontextualize. In the chatbot interviewing study, LLM-based chatbots performed well on relevance, specificity, clarity, and informativeness but poorly on richness metrics such as cognitive empathy, palpability, follow-up, and self-awareness; the authors conclude that LLM chatbots function as adaptive surveyors rather than standalone interviewers capable of eliciting rich qualitative data(12).

A comparative study of GPT-3.5-generated and human interview responses found that key themes were strikingly similar, but also found hyper-accuracy distortion and second-order inference bias, and concluded that high algorithmic fidelity does not equate to safe, ethical, or inclusive usage (43). The same study argues that human expert validation remains indispensable even when LLMs achieve apparent fidelity(43).

Cultural context is a recurring boundary. The deep-learning scaffolding framework argues that what remains primarily under human interpretive control is contextual and incommensurable, such as elements depending on local cultural factors(3). The ethnographic prompt-engineering framework in social work attempts to embed cultural themes into prompt components, but it also notes that the ethnographic data analysis was conducted manually without AI and that the automated prompt refinement process was more challenging than expected(45).

The AI-driven documentation of East Sumba ikat traditions used NLP, computer vision, ontology modelling, and knowledge graphs, but it also introduced sacredness flags and access restrictions based on selective semantic disclosure by Indigenous actors, showing that cultural context can be encoded only through explicit human governance decisions(36).

A study of generative AI in anthropology teaching found that AI-generated ethnographic personas often flattened cultural complexities and reinforced stereotypes, requiring critical data literacy and process reflection(49).

Another anthropological reflection argues that embodied, lived experience remains difficult if not impossible for chatbots to replicate, anchoring ethnographic authenticity in human presence(47).

The literature also documents specific interpretive failures. An LLM-assisted Template Analysis workflow found that GPT-4 could generate codes, subcodes, clusters, and themes, but initial themes often resembled clusters and lacked participant-connected contextual meaning; the context window could not process the full corpus, leading to data reduction and potential hallucination(11).

A study of LLMs coding interviews with firearm violence survivors found that guardrails led to substantial narrative erasure: on average 44% of each prompt request refused to generate output, with up to 65% of some subjects' experiences ignored, primarily because of graphic violence, explicit language, sexual activity, and race discussions that were focal study topics(50).

The same study found that the machine pipeline initially generated around 3,000 unique codes, later reduced to fewer than 100 via BERTopic, and that formal machine codes appeared to have lower validity and higher likelihood of hallucination than initial codes(50).

Video analysis extends these concerns to multimodal data. A study of a multimodal LLM applied to educational video identified description hallucinations in 13 of 36 clips, interpretation hallucinations in 11, and instruction hallucinations in 6; examples included misidentifying AR headsets as VR headsets, counting errors, gender misidentification, misidentifying a facilitator's role, and fabricating an interaction(51).

The authors conclude that multimodal LLMs may be more effective as initial coding assistants with human oversight than as fully automated analysis tools(51).

Even when AI is used successfully, human interpretive control is described as necessary for final meaning-making. The researcher using GPT-4 as a research assistant retained the role of reflexive instrument and intellectual leader, with final interpretive decisions remaining human(11).

The augmented qualitative researcher model found that higher-level analytical queries were less reliable and required substantial empirical verification, and that human researchers remained

central to valid interpretive analysis(20). The Q3 framework analysis concludes that researcher expertise and contextual knowledge are necessary to supplement LLM limitations, especially for sensitive or complex topics(18).

The antipatterns catalog argues for a human-in-the-lead rather than human-in-the-loop approach to preserve constructivist and reflexive foundations(22). A methodological article on the researcher-AI tandem reports that AI coding fails to capture silence, irony, and quoted opinions, requiring manual coding, and that AI cannot capture silence as diagnostic data(19).

Thick description therefore survives in AI-augmented qualitative digital research not by being automated but by being protected. The evidence suggests that researchers preserve it through explicit task boundaries, manual review of AI outputs, contextual annotation, theoretical framing, and reflexive journaling. Where those protections are absent, the result is often a thinner, more generic, or biased account.

Validation, Transparency, Reflexivity, and Auditability

Validation in the included studies is plural and often improvised. The most common quantitative strategies are inter-coder agreement metrics such as Cohen's kappa, Krippendorff's alpha, F1 scores, precision, recall, cosine similarity, and topic coherence. These are used to compare AI outputs with human coding, with a gold standard, or with a second model. The metrics reported in this section are not commensurable. Cohen's kappa, Krippendorff's alpha, F1, precision, recall, cosine similarity, and topic coherence index different constructs, use different baselines, and are not pooled or meta-analyzed here. Reporting them in one narrative is intended to map the range of validation practices, not to imply cumulative evidence. The LLM-in-the-loop study reports Cohen's kappa of 0.87 and 0.81 between human and machine coders, alongside cosine similarity above 0.88 to the original authors' codes(8).

The inductive open-coding comparison reports low human inter-coder reliability, with Krippendorff's alpha of 0.2, and argues that low reliability suggests the task is highly subjective(33). The communication-data coding study reports human-human agreement of 0.582 and direct-prompting agreement of 0.591, and notes that no shared benchmarks exist for cross-study comparison(41). The psychological safety coding study reports kappa values of 0.33–0.44 and argues for multi-run evaluation and category-level bias monitoring(40).

The systematic review coding study reports final Cohen's kappa of 0.899 and 0.823 and proposes human arbitration to resolve discrepancies(42). The hybrid human-machine learning study reports F1 of 0.63 for machine-centered BERT, alpha of 0.80 for human-only coding, and F1 of 0.88 for the hybrid strategy(31).

These metrics are informative but not sufficient. Several studies note that inter-rater reliability may be ambiguous for human-AI coding, that conventional metrics may not capture interpretive quality, and that high agreement can coexist with shared bias(5,17,46). The multi-agent consensus study found that neither temperature nor persona pairing led to robust improvements in coding accuracy, and that single agents matched or outperformed multi-agent consensus in most conditions; only 5 of 432 condition-level comparisons significantly favored the multi-agent system after correction(39).

The comparison of a human and a multi-agent thematic analysis found that AI-generated codes were rated appropriate in 91% of data points versus 81% for human codes, but experts still preferred human code sets in direct comparisons, and attribution of whether codes were AI- or human-generated was only 44% correct(46). The AI-coder discussion study warns that AI "agreement" may project an illusion of rigor while bypassing the epistemic labor of reflexive engagement (17).

Human validation remains the most common safeguard. The Reddit mental health pipeline used two independent random samples of 250 code-text pairs, human reading, and manual thematic analysis, and still found that human validation was required for context-dependent nuances(28).

The RAG-based writing-analysis study found that human participation remained essential for checking edge cases and errors, and noted that human-AI interaction was unilateral, with humans selecting valuable AI insights without dialogue about reasoning(52). The GPT-4 coding transparency study used process mining and idea-thread visualizations to trace code changes to data, researcher discussions, and AI output, and reported that AI codes were not directly adopted but stimulated discussion and refinement(53).

The Co-Refine system constrains LLM consistency scores within ±0.15 of deterministic embedding metrics, uses an immutable edit history and append-only consistency records, and reports a mean SUS of 77.77 (23). These are among the clearest examples of auditability in the included evidence.

Transparency reporting varies. Some studies report model versions, dates, prompts, temperature, and random seeds(11,41,18). The Q3 framework analysis argues that transparency in reporting LLM use, including prompts, model version, and dates, is needed for process reliability(18).

The systematic mapping study of LLMs in software engineering qualitative analysis recommends disclosing the LLM tool and version, experimenting with prompt strategies, establishing data protection protocols, ensuring human participation, and discussing validity threats contextually(6). A review of AI in qualitative research identifies the lack of standardized quality criteria and reporting norms as a central gap(5). The antipatterns catalog recommends that reviewers assess human engagement, traceability, transparency, and grounding in data(22).

Reflexivity is addressed in several ways. The AI-assisted dissertation case study reports that AI collaboration augmented the researcher's analytical skills and fostered reflexivity, and that using AI encouraged authentic engagement with bias and positionality(54). The anthropological teaching study used logs, annotated outputs, and reflective journals to support transparency and critical awareness, and describes a movement from "knowledge telling" to "knowledge transforming"(49). The researcher-AI tandem article argues that AI integration intensifies researcher reflexivity rather than replacing it, making biases more visible, and proposes a two-level validation procedure and bias audit(19). The computational ethnography of Twitter uses asymmetric comparison of differently bounded datasets to reveal locally meaningful scenes that larger datasets miss, an explicitly reflexive strategy(10). The organizational LLM ethnography frames accountability as a social, technical, institutional, economic, and legal achievement rather than a technical property(37).

Auditability is less developed. The clearest examples are the provenance-oriented tools and procedures described in the business anthropology essay, which argues for provenance standards and explicit task boundaries between humans and machines(29), and the Co-Refine system's append-only consistency records and deterministic grounding(23).

The East Sumba cultural documentation project used metadata schemas, RDF triples, and knowledge graphs with sacredness flags and access restrictions, which provides a model for auditable cultural governance(36). The GPT-4 coding transparency study used process mining and visualizations to document which parts of coding involved AI(53). Beyond these, auditability is often limited to reporting prompts and model versions, without a full trace of how AI outputs were transformed into final interpretations.

Ethical Risks, Privacy, Consent, Bias, and Algorithmic Accountability

Ethical concerns appear in nearly every included study, but they are rarely operationalized in a standardized way. The most frequently mentioned issues are data privacy, informed consent, anonymization, proprietary data retention, bias, transparency, and the risk of over-reliance.

Qualitative researchers interviewed about LLM use called attention to an urgent lack of norms and tooling to guide ethical LLM use, and worried that available tools do not present researchers with the right options to confidently preserve privacy(2). They recommended updating consent forms, avoiding proprietary models that train on input for participant data, validating LLM performance task-specifically, and considering cultural and political bias(2).

The augmented qualitative researcher article lists data privacy, anonymization, informed consent, GDPR compliance, proprietary retention of user data, and environmental costs as ethical concerns(20). The GPT-4 Template Analysis study notes data privacy concerns for third-party platforms and the risk of "dumping" data into generative AI, which can oversimplify nuanced data and erode reflexivity(11). The Q3 framework analysis warns that using commercial LLMs to handle participant data raises ethical issues including consent, data management plan violations, and breach of trust(18).

Privacy and anonymization practices vary. The Reddit mental health study used named entity anonymization with GLiNER before LLM coding(28). The firearm violence survivor study manually de-identified interviews before machine coding, but still found that guardrails erased narratives related to violence, race, and sexuality(50).

The low-income African context review notes that ethical concerns, data privacy, infrastructure limitations, and the black-box nature of AI models challenge transparency and trust, and recommends capacity building, ethical frameworks, infrastructure investment, open-source solutions, and community engagement (24). The organizational ethnography of an internal LLM deployment found that data privacy and confidentiality concerns prompted formal governance interventions, including an LLM policy and restrictions on use(37).

Bias is reported in multiple forms. The silicon-participant study found hyper-accuracy distortion and second-order inference bias, such as female silicon participants referring to husbands

while no male silicon participants referred to wives or partners(43).

The psychological safety coding study found systematic over-prediction of "Sharing Negative Feedback" by up to 5.25 times and under-prediction of "Expressing Concerns" across all models and prompt configurations, with bias patterns persisting under both zero-shot and multi-shot prompting(40). The value-alignment study found systematic overemphasis on the Security value across models, suggesting possible model-induced value biases, and found that uncertainty structures diverged from expert patterns(55). The video-analysis study found gender misidentification and role misidentification(51). The antipatterns catalog warns that AI tendencies such as hallucination, generalization, overinterpretation, and sycophancy can reinforce bias and reduce depth and reflexivity(22).

Algorithmic accountability is discussed in several registers. The organizational ethnography argues that accountability requires social, technical, institutional, economic, organizational, and legal components, not just technical compliance, and that localised, bounded deployments enable stakeholder preferences to be elicited and oversight to be maintained(37).

The predictive-care protocol for computational ethnography in emergency psychiatry explicitly evaluates machine learning models for intersectional bias, false positive and false negative parity, and incomplete EHR data, noting that structured risk scales can have high false positive rates and that EHR data may be missing on over 50% of some sociodemographic variables(25). The Q3 framework analysis argues that LLM training-data biases can perpetuate dominant narratives and misrepresent marginalized experiences, threatening theoretical validation(18). The systematic review of AI in qualitative research identifies algorithmic bias, interpretive validity, ethics, and data privacy as central concerns, and notes that institutional review boards often lack expertise to assess AI-related risks and consent language (5).

Disclosure is another recurring theme. The higher education faculty study found that disclosure of AI use was a common concern among all participants, and that one participant suggested students should manually code and then use AI to validate and compare differences(35).

The researcher-AI tandem article proposes algorithm transparency, human-involved validation, bias audit, and result interpretability as criteria for AI-supported validity(19). The SUPERVISE framework chapter argues for responsible documentation and equitable supervisory practice, and notes that technical parameters such as tokenisation, context windows, temperature, and platform guardrails function as methodological variables(21). The quasi-systematic review of AI-supported interviews notes that ethical concerns including bias, data privacy, and compliance with GDPR and HIPAA are particularly salient in sensitive contexts(7).

What is striking is the gap between awareness and infrastructure. Many studies mention ethics, but few provide a complete account of consent, data flow, model retention, anonymization, bias auditing, and disclosure. The evidence suggests that ethical practice in AI-augmented qualitative digital research is currently more a matter of individual researcher vigilance than of shared standards.

Models and Failure Modes of Human-AI Collaboration

The included studies contain several reproducible models of human-AI collaboration. The LLM-in-the-loop framework structures thematic analysis as a dialogue between a human coder and a machine coder across four steps: data familiarization and exemplar generation, initial code generation, iterative code refinement and codebook development, and theme identification and evaluation(8). CollabCoder integrates LLMs into collaborative qualitative analysis as a suggestion provider in open coding, a mediator and facilitator in discussions, and a support for codebook development, with user autonomy and coding independence maintained(15).

The Guided AI Thematic Analysis framework operationalizes an adapted Template Analysis with GPT-4 and the ACTOR prompting framework, while the researcher remains the reflexive instrument and intellectual leader(11). The augmented qualitative researcher model integrates LLMs into interpretive text analysis through iterative prompt development, holdout coding, robustness checks, and human-led verification(20). The research-as-assemblage framework positions hermeneutic and algorithmic tasks as interacting epistemically in a non-hierarchical, rhizomatic way(3). The Reddit mental health study presents a human-in-the-loop pipeline combining LLM coding, BERTopic clustering, and human validation(28).

The HHMLA model uses BERT embeddings and logistic regression to prioritize recall, followed by human review to filter false positives, and reports higher reliability than either human-only or machine-centered coding(31). The RAG-based writing-analysis workflow uses retrieval to select relevant responses for expert rating and thematic analysis, with humans checking edge cases(52).

The multi-agent AutoTheme framework uses independent coder agents, a code reconciler, theme generators, and a theme reconciler, but still requires human involvement for comprehensive narrative construction(16). The SUPERVISE framework proposes hybrid human-AI workflows with responsible documentation and equitable supervision(21). The researcher-AI tandem proposes four procedures: analysis architecture design, iterative feedback, two-level validation, and a protocol for detecting systemic biases(19). The business anthropology essay describes provenance-tracked tools that codify task boundaries between humans and machines(29). The antipatterns catalog argues for human-in-the-lead rather than human-in-the-loop(22).

These models share several features. They allocate classification, retrieval, suggestion, and pattern detection to AI. They reserve contextual interpretation, theoretical integration, ethical judgement, and final meaning-making for humans. They build in iterative review rather than one-shot automation. They increasingly attempt to document prompts, model versions, and decision traces. The strongest models also include explicit mechanisms for contesting AI output, such as human review of machine-positive cases, second-coder validation, or constrained LLM scoring.

The failure modes are equally consistent. Hallucination is reported in multiple studies, including fabricated codes, invented data points, unsupported justifications, and misidentified roles or objects(32,11,51,39). Stochastic instability appears as variability across runs, sensitivity to prompt wording, temperature effects, and inconsistent adherence to requested formats(26,41,40).

Context-window limits lead to data reduction, forgetting, and potential hallucination(11). Error propagation occurs when hierarchical decisions constrain later classifications(41). Category bias and minority-category underperformance are documented in coding studies(40). Refusal-driven narrative erasure occurs when guardrails block discussion of violence, race, sexuality, or explicit language that is central to the research question(50).

Over-reliance and cognitive surrender are reported or warned against, including the risk that AI outputs are treated as insights rather than as suggestions(22,35). Low inter-rater reliability between LLM and human assessment is documented in chatbot-based data collection(12). AI agreement can create an illusion of rigor while bypassing reflexive engagement (17). Contextual nuance and cultural meaning are frequently lost (49,47). Model deprecation and version change threaten replication(28,20). Proprietary data processing raises privacy and reproducibility concerns(27,26).

The evidence therefore supports a cautious conclusion: AI can augment qualitative digital research in specific, bounded ways, but it introduces predictable methodological and ethical failure modes that require explicit design responses. The most defensible models are those that combine task-level delegation with strong human interpretive authority, contextual grounding, validation against human judgement, and auditable records of what the AI did and how its output was used. These models and failure modes are synthesized further in the propositions below.

DISCUSSION

The synthesis reveals several reporting gaps. First, model and prompt reporting is inconsistent. Some studies report model versions, dates, temperatures, and prompts(11,41,18), while others report only the model family or provide limited prompt detail. Because LLM outputs are stochastic and model versions change, this inconsistency limits reproducibility(1,20). Second, validation standards are unsettled. Conventional inter-rater reliability is widely used but may be ambiguous for human-AI coding, and it may not capture interpretive quality or shared bias(5,46). Third, ethical reporting is often generic. Privacy, consent, and bias are frequently mentioned but rarely documented through concrete data-flow descriptions, consent language, anonymization procedures, or bias audits(2,24). Fourth, auditability is underdeveloped. Few studies provide a full trace from raw data through AI output to final interpretation(23,53,29). Fifth, the division of interpretive labour is often described in general terms rather than mapped task by task(15,19). Sixth, the literature says relatively little about how AI affects the researcher-participant relationship, the co-construction of knowledge, or the researcher's own reflexive practice over time(5,21). Seventh, there is limited attention to cultural context, language diversity, and social positioning as factors that shape AI performance(5,36,24).

These gaps have methodological implications for qualitative digital research, including netnography and digital ethnography. Netnographic data are often relational, longitudinal, and context-dependent. A coding pipeline that works on short survey responses may fail on threaded conversations, ironic exchanges, multimodal posts, or community-specific language.

The evidence suggests that AI is better at identifying the presence of constructs than at assessing their quality, relevance, or meaning (48). It also suggests that AI is better at supporting

open coding than at axial or selective coding, and better at pattern detection than at theoretical integration(21).

This means that AI-augmented netnography should not be designed as a pipeline that runs from data to findings. It should be designed as a set of bounded collaborations in which AI outputs become objects of human interpretation rather than substitutes for it.

The evidence also implies that AI changes the researcher's work rather than simply reducing it. Several studies report that AI shifts effort from mechanical coding to prompt design, output review, error correction, bias monitoring, and documentation(26,19,31). The researcher-AI tandem article argues that AI adoption transforms qualitative sociologist competencies: mechanical coding skills diminish while critical thinking, conceptual creativity, and prompt engineering skills strengthen(19). The higher education faculty study raises similar concerns about students not learning to verify AI outputs(35). The antipatterns catalog warns of community-level skill erosion(22). These are not merely efficiency questions; they concern the reproduction of interpretive expertise.

Finally, the review highlights a tension between scale and depth. AI enables researchers to work with larger corpora, more languages, and more heterogeneous data, which can democratize access and reduce costs(24,3). But scale can also encourage thinner analysis, overgeneralization, and the substitution of pattern detection for cultural understanding(49,22). The most credible studies in this review do not treat scale as an end in itself. They use AI to surface patterns that humans then interrogate, contextualize, and validate.

Conditional Propositions for Rigorous AI-Augmented Netnography

The following propositions are derived from a heterogeneous and unappraised set of studies. They are conditional and testable rather than prescriptive, and they do not constitute a validated framework or reporting standard. They are organized around eight components and phrased as propositions for future testing.

Proposition 1: Task allocation mapping. If researchers map each task in the netnographic workflow before analysis and decide whether it will be performed by humans, by AI, or through a defined collaboration, then accountability for final decisions may be clearer. AI may be most defensible for transcription, translation, initial or open coding, code suggestion, topic modelling, sentiment or emotion labelling, pattern detection, summarization, and first-pass classification, while contextual interpretation, theoretical integration, ethical judgement, and final meaning-making are likely to remain human-led(7,14,15,22). The mapping could specify what the AI receives, what it returns, how outputs are reviewed, and who is accountable(19,29).

Proposition 2: Contextual grounding and thick description. If AI prompts and workflows are grounded in the specific cultural, linguistic, and platform context of the field, then outputs may be more likely to preserve thick description. This could include using community-specific examples, annotating local meanings, and treating cultural context as a constraint on automated classification rather than as noise(3,45,36). Researchers might explicitly assess whether AI outputs preserve or flatten thick description, and could document cases where AI misses irony, silence, ambiguity, or quoted speech(19,49).

Proposition 3: Human interpretive authority. If the workflow adopts a human-in-the-lead model rather than a fully automated model, then interpretive authority may be better preserved(22). Human researchers would retain authority over codebook finalization, theme construction, theoretical interpretation, and the narrative account. AI may generate suggestions, but those suggestions are treated as inputs to human judgement, not as findings(20,11,46). Where AI is used to apply a validated coding scheme, both the scheme and the model's accuracy are checked separately(56).

Proposition 4: Validation and contestation. If validation combines quantitative agreement metrics with human interpretive review and explicit contestation, then the limits of AI output may be better understood. Useful practices could include human review of AI-positive cases, second-coder validation, comparison against a gold standard, multi-run stability testing, triangulation with a second model or method, and attention to minority-category performance(28,31,40,23). Validation might also assess whether AI outputs introduce systematic bias, not only whether they agree with human coders(40,55). Disagreement may be analytically useful rather than a failure, following evidence that AI can disrupt consensus and introduce alternative perspectives for human consideration(39,17).

Proposition 5: Ethical governance and privacy. If ethical governance is specified before data collection and updated as the workflow evolves, then privacy and consent risks may be reduced. It could address informed consent, anonymization, data minimization, third-party API retention, cross-border transfer, model training policies, and platform terms of service(2,24,20). Where pos-

sible, researchers might prefer tools that do not train on participant data, use local or self-hosted models for sensitive material, and document data flows(27,57). Bias auditing could be planned, not ad hoc, and could examine category-level performance, refusal patterns, and representation of marginalized experiences(40,50,25).

Proposition 6: Reflexivity and positionality. If AI-augmented qualitative digital research incorporates reflexive practices that examine how AI use shapes the researcher's assumptions, choices, and relationship to the field, then interpretive awareness may be strengthened. This could include reflexive journaling, prompt logs, annotation of AI outputs, and periodic review of how AI suggestions influence interpretation(54,49,19). Reflexivity might also address the researcher's position relative to the community and the ethical implications of using computational tools to interpret others' meaning-making (37,36).

Proposition 7: Transparency and reporting. If reports disclose the AI tools used, model versions and dates, prompt strategies, temperature and seed settings where applicable, data segmentation or truncation decisions, validation procedures, and the division of interpretive labour, then reproducibility may improve(18,41,6,21). Where proprietary models are used, reports could describe what data were sent, what retention policies applied, and what limitations follow(27,26). Transparency might extend to negative results, failed prompts, hallucinations, and cases where AI output was rejected(22,39).

Proposition 8: Auditability and provenance. If an audit trail records how data were collected, what was sent to AI, what AI returned, how outputs were reviewed, what was changed, and who made final decisions, then accountability may be strengthened(23,53,29). Auditability may be especially important in qualitative digital research because field boundaries, community norms, and data provenance are often contested. Provenance standards could also address cultural governance, including what cultural knowledge is shareable, what is sacred, and what access restrictions apply(36).

Taken together, these propositions imply a workflow that is slower and more documented than a fully automated pipeline but faster and more scalable than purely manual analysis. They are intentionally conservative about automation and treat AI as a collaborator in specific tasks, not as an author of ethnographic interpretation. Their value depends on future empirical testing rather than on the unappraised evidence from which they are derived.

CONCLUSION

AI is being integrated into qualitative digital research, including netnography and digital ethnography, in ways that are both promising and uneven. The evidence shows that AI can support transcription, translation, initial coding, code suggestion, topic modelling, sentiment analysis, pattern detection, summarization, and first-pass thematic grouping, often with substantial gains in speed and scale. It also shows that human interpretive control remains indispensable for contextual understanding, theoretical integration, ethical judgement, reflexivity, and final meaning-making.

The most credible models of collaboration are neither fully automated nor purely manual. They allocate bounded tasks to AI, subject AI outputs to human review, validate against human judgement, and document the process.

The review also identifies persistent failure modes, including hallucination, stochastic instability, context-window truncation, error propagation, category bias, refusal-driven narrative erasure, low reliability for minority categories, cultural flattening, and over-reliance on plausible but ungrounded output.

Ethical reporting is widespread but under-specified, and auditability remains the weakest link in most studies. The conditional propositions indicate a possible way forward by treating AI as a bounded collaborator, prioritizing contextual grounding and human interpretive authority, and requiring validation, ethical governance, reflexivity, transparency, and auditability.

Their value will depend on whether future research tests them against the messy realities of online communities, and whether the field develops shared norms for reporting AI use that are proportionate to the interpretive and ethical stakes of qualitative digital research. Because this review rests on heterogeneous and unappraised reports, its conclusions should be read as a mapping of the field and a set of testable propositions rather than as settled guidance.

REFERENCES

1. Thomas Davidson. Start Generating: Harnessing Generative Artificial Intelligence for Sociological Research. 2023. doi:10.31235/osf.io/u9nft.
2. Hope Schroeder, Marianne Aubin Le Quéré, Casey Randazzo, David Mimno, Sarita Schoenebeck. Large Language Models in Qualitative Research: Uses, Tensions, and Intentions. arXiv (Cornell University). 2024. doi:10.48550/arxiv.2410.07362.
3. André Luís A. da Fonseca, Paula Chimenti, Maribel Carvalho Suarez. Using Deep Learning Language Models as Scaffolding Tools in Interpretive Research. Revista de Administração Contemporânea. 2023. doi:10.1590/1982-7849rac2023230021.en.
4. Tomek Strzalkowski, Anna Newheiser, Nathan Kemper, Ning Sa, Bharvee Acharya, Gregorios Katsios. Generating Ethnographic Models from Communities' Online Data. 2020. doi:10.18653/v1/2020.figlang-1.23.
5. Dimple Ravindra Patil, Nitin Liladhar Rane, Obizue Mirian Ndidi, Jayesh Rane. Qualitative research using artificial intelligence: Methods, techniques, challenges, and future directions. International Journal of Applied Resilience and Sustainability. 2026. doi:10.70593/deepsci.0202015.
6. Matheus de Morais Leça, Lucas Valença, Reydne Santos, Ronnie de Souza Santos. Applications and Implications of Large Language Models in Qualitative Analysis: A New Frontier for Empirical Software Engineering. arXiv (Cornell University). 2024. doi:10.48550/arxiv.2412.06564.
7. Tomasz Kupiec. AI-Supported Individual Interviews: Opportunities and Threats for Evaluation. Polski Przegląd Ewaluacyjny. 2025. doi:10.65468/487cvb.
8. Shih-Chieh Dai, Aiping Xiong, Lun-Wei Ku. LLM-in-the-loop: Leveraging Large Language Model for Thematic Analysis. 2023. doi:10.18653/v1/2023.findings-emnlp.669.
9. He Zhang, Chuhao Wu, Jingyi Xie, Yao Lyu, Jie Cai, John M. Carroll. Redefining Qualitative Analysis in the AI Era: Utilizing ChatGPT for Efficient Thematic Analysis. arXiv (Cornell University). 2023. doi:10.48550/arxiv.2309.10771.
10. Philipp Brandt. Data science's cultural construction: qualitative ideas for quantitative work. Frontiers in Big Data. 2024. doi:10.3389/fdata.2024.1287442.
11. Kien Nguyen-Trung. ChatGPT in Thematic Analysis: Can AI become a research assistant in qualitative research?. 2024. doi:10.31219/osf.io/vefwc.
12. Cuevas, Alejandro, Jennifer V. Scurrell, Eva Maxfield Brown, Jason Entenmann, Madeleine I. G. Daepp. Collecting Qualitative Data at Scale with Large Language Models: A Case Study. arXiv (Cornell University). 2023. doi:10.48550/arxiv.2309.10187.
13. Amandeep Kaur, James R. Wallace. Moving Beyond LDA: A Comparison of Unsupervised Topic Modelling Techniques for Qualitative Data Analysis of Online Communities. arXiv (Cornell University). 2024. doi:10.48550/arxiv.2412.14486.
14. Cesare Piano, Samira Abbasgholizadeh Rahimi, Jackie Chi Kit Cheung. Qualitative Code Suggestion: A Human-Centric Approach to Qualitative Coding. 2023. doi:10.18653/v1/2023.findings-emnlp.993.
15. Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, et al. CollabCoder: A Lower-barrier, Rigorous Workflow for Inductive Collaborative Qualitative Analysis with Large Language Models. arXiv (Cornell University). 2023. doi:10.48550/arxiv.2304.07366.
16. Abdullah Wahbeh, Omar El-Gayar, Mohammad Al-Ramahi, Tareq Nasralah, Ahmed Elnoshokaty. AutoTheme: A Multi-Agent Framework for Inductive Thematic Analysis with LLMs. Proceedings of the .. Annual Hawaii International Conference on System Sciences/Proceedings of the Annual Hawaii International Conference on System Sciences. 2026. doi:10.24251/hicss.2026.214.
17. Mitchell, John. How AI Coders Discuss, Disagree, and Reach Consensus: Challenges and Opportunities for LLM-Based Qualitative Coding. arXiv (Cornell University). 2026. doi:10.48550/arxiv.2609.11109.
18. David Reeping, Cynthia Hampton, Desen Özkan. Interrogating the Use of Large Language Models in Qualitative Research Using the Qualifying Qualitative Research Quality Framework. Studies in Engineering Education. 2025. doi:10.21061/see.174.
19. V. O. Mikryukov. Features and Emerging Challenges of the "Researcher + AI" Tandem in Qualitative Sociological and Marketing Research. Inter. 2026. doi:10.19181/inter.2026.18.2.7.
20. Elida Izani Ibrahim, Andrea Voyer. The Augmented Qualitative Researcher: Using Generative AI in Qualitative Text Analysis. 2024. doi:10.31235/osf.io/gkc8w.
21. Anouck Butraud-Assathian, Cécile Méadel, Jaércio DA SILVA. Artificial Intelligence and Social Research: Methods, contexts, imaginaries. OAR@UM (University of Malta). 2025. doi:10.69146/55440895.
22. Rashina Hoda, Carolyn Seaman, Victória Gomes, Rodrigo Spinola. Antipatterns in AI-assisted Qualitative Data Analysis: A Catalog of Temptations and Pitfalls for Software Engineering Researchers. arXiv (Cornell University). 2026. doi:10.48550/arxiv.2608.27927.
23. Athikash Jeyaganthan, Kai Xu, Franziska Becker, Steffen Koch. Co-Refine: AI-Powered Tool Supporting Qualitative Analysis. arXiv (Cornell University). 2026.
24. K. Isangula. Navigating Barriers: Challenges and Strategies for Adopting Artificial Intelligence in Qualitative Research in Low-Income African Contexts. Tanzania Journal of Health Research. 2025. doi:10.4314/thrb.v26i3.14.
25. Laura Sikstrom, Marta M. Maslej, Zoe Findlay, Gillian Strudwick, Katrina Hui, Juveria Zaheer, et al. Predictive care: a protocol for a computational ethnographic approach to building fair models of inpatient violence in emergency psychiatry. BMJ Open. 2023. doi:10.1136/bmjopen-2022-069255.
26. A. Cevik, F. Abu-Zidan. Utilizing AI-Powered Thematic Analysis: Methodology, Implementation, and Lessons Learned. Cureus. 2025. doi:10.7759/cureus.85338.

27. Sreyoshi Bhaduri, Satya Kapoor, Alex Gil, Anshul Mittal, Rutu Mulkar. Reconciling Methodological Paradigms: Employing Large Language Models as Novice Qualitative Research Assistants in Talent Management Research. arXiv (Cornell University). 2024. doi:10.48550/arxiv.2408.11043.
28. Drin Ferizaj, Christopher Lalk, Nils Lahmann, Sandra Strube-Lahmann, Julian Rubel. Identifying Yalom's group therapeutic factors in anonymous mental health discussions on Reddit: a mixed-methods analysis using large language models, topic modeling and human supervision. Frontiers in Psychiatry. 2025. doi:10.3389/fpsyt.2025.1503427.
29. Adam Gamwell, Phil Surles. Reclaiming Relevance: A New Agenda for Business Anthropology in the Age of AI. Journal of Business Anthropology. 2026. doi:10.22439/jba.v15i1.7815.
30. Álvaro Sebastián Cumbe Saquisilí, Juan Bautista Solís-Muñoz, Diego Sebastián Flores Cantos. Etnografía digital innovada por IA. 2026. doi:10.58995/lb.redlic.81.374.
31. Zhuofan Li, Daniel Dohan, Corey M. Abramson. Qualitative Coding in the Computational Era: A Hybrid Approach to Improve Reliability and Reduce Effort for Coding Ethnographic Interviews. 2021. doi:10.31235/osf.io/gpr4n.
32. Stefano De Paoli. Can Large Language Models emulate an inductive Thematic Analysis of semi-structured interviews? An exploration and provocation on the limits of the approach and the model. arXiv (Cornell University). 2023. doi:10.48550/arxiv.2305.13014.
33. Angelina Parfenova, Andreas Marfurt, Jürgen Pfeffer, Alexander Denzler. Text Annotation via Inductive Coding: Comparing Human Experts to LLMs in Qualitative Data Analysis. 2025. doi:10.18653/v1/2025.findings-naacl.361.
34. Yilmaz Saglam. Beyond Automation: Prompt Design and Trustworthiness in AI-Assisted Inductive Coding. International Journal on Social and Education Sciences. 2026. doi:10.46328/ijonses.8906.
35. Abigail M. Nubla-Kung, Pressley R. Rankin, Mary Dereshiwsky. Higher Education Faculty Members' Perceptions of an AI-driven Qualitative Data Analysis Tool for Their Research: An Exploratory Study. Journal of Online Graduate Education. 2026. doi:10.65201/001c.158996.
36. Laely Indah Lestari, Evi Novianti, Dadang Sugiana, Ute Lies Siti Khadijah. From Weaving Patterns to Data Patterns: AI-Driven Cultural Documentation of East Sumba's Ikat Traditions. Data & Metadata. 2025. doi:10.56294/dm20251129.
37. Kelsie Nabben. AI as a constituted system: accountability lessons from an LLM experiment. Data & Policy. 2024. doi:10.1017/dap.2024.58.
38. Jianfeng Zhu, Karin G. Coifman, Ruoming Jin. Understanding Risk and Dependency in AI Chatbot Use from User Discourse. arXiv (Cornell University). 2026.
39. Conrad Borchers, Bahar shahrokhian, Francesco Balzan, Elham Tajik, Sreecharan Sankaranarayanan, Sebastian Simon. Temperature and Persona Shape LLM Agent Consensus With Minimal Accuracy Gains in Qualitative Coding. arXiv (Cornell University). 2025. doi:10.48550/arxiv.2507.11198.
40. Moaath Alshaikh, Tasneem Alshaher, Ricardo Vieira, Beatriz Santana, Clelio Xavier, José Amâncio, et al. Prompt Engineering Strategies for LLM-based Qualitative Coding of Psychological Safety in Software Engineering Communities: A Controlled Empirical Study. arXiv (Cornell University). 2026.
41. Wenju Cui, Jiangang Hao, Yang Jiang, Patrick C. Kyllonen, Emily Kerzabi. Automated coding of communication data using large language models: a comparison of hierarchical and direct prompting strategies. Frontiers in Education. 2026. doi:10.3389/feduc.2026.1764154.
42. Kim Uittenhove, Paolo Martinelli, Angélique Roquet. Large Language Models in Psychology: Application in the Context of a Systematic Literature Review. 2024. doi:10.31234/osf.io/nq4d2.
43. Aliya Amirova, Theodora Fteropoulli, Nafiso Ahmed, Martín Cowie, Joel Z. Leibo. Framework-based qualitative analysis of free responses of Large Language Models: Algorithmic fidelity. PLoS ONE. 2024. doi:10.1371/journal.pone.0300024.
44. Stefano De Paoli. Writing user personas with Large Language Models: Testing phase 6 of a Thematic Analysis of semi-structured interviews. arXiv (Cornell University). 2023. doi:10.48550/arxiv.2305.18099.
45. Miroslavas Seniutis, Valentas Gružauskas, Artūras Sas, Valentinas Navickas, Mantas Švažas. Designing a framework for ethnography-driven prompt engineering in social work. Human Technology. 2025. doi:10.14254/1795-6889.2025.21-1.5.
46. Sebastian Simon, Sreecharan Sankaranarayanan, Elham Tajik, Conrad Borchers, Bahar shahrokhian, Francesco Balzan, et al. Comparing a Human's and a Multi-Agent System's Thematic Analysis: Assessing Qualitative Coding Consistency. Lecture notes in computer science. 2025. doi:10.1007/978-3-031-98420-4_5.
47. Jakob Krause-Jensen, Mark Friis Hau. Chatbots and the Craft of Ethnography: Exploring AI's Impact on Anthropological Teaching and Practice. Teaching Anthropology. 2025. doi:10.22582/ta.v14i2.783.
48. Rebecca Marrone, Abhinava Barthakur, David Randall, Nazanin Mottaghi, Vitomir Kovanović, Maarten De Laat. Evaluating Generative AI as a Supportive Analytic Partner in Qualitative Coding of Metacognitive Student Reflections. Zenodo (CERN European Organization for Nuclear Research). 2026. doi:10.5281/zenodo.20367266.
49. Tiatemsu Longkumer. Generative AI in Anthropology: Redefining Fieldwork, Textualisation, and Collaborative Knowledge Production. Teaching Anthropology. 2025. doi:10.22582/ta.v14i2.778.
50. Jessica H Zhu, Shayla Stringfield, Vahe Zaprosyan, Michael Wagner, Michel Cukier, Joseph Richardson. Can LLMs Understand the Impact of Trauma? Costs and Benefits of LLMs Coding the Interviews of Firearm Violence Survivors. ArXiv. 2026. doi:10.18653/v1/2026.findings-acl.591.
51. Yiqiu Zhou, Jina Kang, Ha Nguyen. Can We Trust Large Language Models for Video Analysis: An Exploration of Hallucination in Multimodal LLMs. Proceedings. 2025. doi:10.22318/icls2025.704957.

52. Madeleine Sorapure, Seth Erickson, Sarah Hirsch, Kenny Smith. Using AI to understand students' self-assessments of their writing. Journal of Writing Research. 2026. doi:10.17239/jowr-2026.17.03.07.
53. Saríah López-Fierro, Ha Xuan Nguyen. Making Human-AI Contributions Transparent in Qualitative Coding. Computer-supported collaborative learning/The Computer-Supported Collaborative Learning Conference. 2024. doi:10.22318/cscl2024.352932.
54. Corrie Wilder, Shannon Calderone. Empowering Educational Leadership Research with Generative AI. Impacting Education Journal on Transforming Professional Practice. 2025. doi:10.5195/ie.2025.489.
55. Arina Kostina, Marios Dikaiakos, Alejandro Porcel, Tassos Stassopoulos. Can LLMs Capture Expert Uncertainty? A Comparative Analysis of Value Alignment in Ethnographic Qualitative Research. arXiv (Cornell University). 2026.
56. Stephen Dewitt, Alice Liefgreen, Nine Adler, Laura Elaine Strittmatter. 'Please explain your response': A guide to uncovering cognitive processes from open-text box data using pragmatic and reflexive content analysis. Judgment and Decision Making. 2025. doi:10.1017/jdm.2025.10010.
57. Thomas Übellacker. AcademiaOS: Automating Grounded Theory Development in Qualitative Research with Large Language Models. arXiv (Cornell University). 2024. doi:10.48550/arxiv.2403.08844.

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Maicol Stiven Vanegas-Nieto, Sofia Belen Castro Brenes.
Writing–review and editing: Maicol Stiven Vanegas-Nieto, Sofia Belen Castro Brenes.