

How AI Coders Discuss, Disagree, and Reach Consensus: Challenges and Opportunities for LLM-Based Qualitative Coding

JEONGYEON KIM, Stanford University, USA

JOHN MITCHELL, Stanford University, USA

The utility of AI in multi-coder qualitative coding has been widely discussed, yet little empirical evidence exists to delineate the contexts in which it performs reliably. We address this gap by quantifying the effectiveness of multi-agent LLM coding across varied qualitative datasets, revealing key contextual and structural factors that mediate coding outcomes. We developed a literature-informed baseline pipeline that enables AI agents to independently code, debate, and reconcile disagreements. Results revealed that coding accuracy depends on factors such as codebook length, qualitative data similarity, and agent disagreement. Notably, intense and unresolved debates between agents led to higher accuracy. Our analysis showed that while LLMs emulate many human discussion behaviors, they lack adaptive responsiveness to context. From these findings, we offer design recommendations for building automated coding systems. Our open-source AI discussion dataset and methodological framework lay the groundwork for advancing the design of AI-mediated automated thematic analysis.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing systems and tools**; **HCI design and evaluation methods**; • **Computing methodologies** → **Discourse, dialogue and pragmatics**.

1 Introduction

Deductive qualitative coding is the process of applying predefined frameworks to categorize and interpret qualitative data. It is a time-consuming and labor-intensive task even for experts, demanding meticulous examination and contextual understanding of nuanced data [77, 80]. The central role of qualitative coding in social science has led to a growing body of recent work on automating qualitative coding through artificial intelligence (AI) and natural language processing (NLP) techniques. Motivated by our own interest in qualitative research with large amounts of data, we investigate AI-based coding methods modeled on human practices, such as multi-coder comparison and consensus. Our study quantifies the effectiveness of multi-agent Large Language Model (LLM) coding across varied qualitative datasets, revealing key contextual and structural factors that mediate coding outcomes.

In leveraging AI for multi-coder qualitative coding, we investigate how effectively AI can engage with the complexities of data, identifying specific conditions — when and in which contexts — AI demonstrates consistent reliability. Our analysis is intended to help researchers using our methods avoid both *over-reliance* that can amplify automation bias by sidelining necessary human interpretation and *under-trust* that could preclude their potential utility. Although AI tools may perform adequately in structured cases, they often falter in interpreting the nuanced subtleties of qualitative data like interview transcripts. This form of data may resist rigid categorization, requiring interpretive judgment that extends beyond mechanical rule-based classification. Even among experienced human coders, variability in interpretation is common, which is why measures like intra-rater reliability and consensus-building are indispensable. Qualitative coding is thus not merely a technical task but a deeply interpretive practice grounded in human insight.

Despite these inherent complexities of qualitative coding, recent advancements in LLMs have accelerated efforts to automate this process. They explored prompt engineering [31, 183] and incorporated human feedback loops into the automation process [14, 179]. However, these efforts remain confined to narrowly defined contexts, addressing limited or domain-specific datasets. Moreover, the prevailing focus on coding accuracy marginalizes user-centered

Authors' Contact Information: Jeongyeon Kim, jeongyeon.kim@stanford.edu, Stanford University, Stanford, CA, USA; John Mitchell, John.Mitchell@stanford.edu, Stanford University, Stanford, CA, USA.

Manuscript submitted to ACM

considerations and broader system design concerns, such as transparency and interpretability. As a result, system designers are left without comprehensive, practically grounded frameworks to guide the development of automated systems that align with the needs and expectations of end users. Our research bridges these gaps by (1) investigating user needs with automated coding systems through conducting an exploratory literature review, (2) evaluating LLM capabilities across diverse datasets, and (3) proposing actionable design guidelines for future systems.

The findings of the exploratory literature review emphasize the significance of transparency and reliability in automated coding systems. The reviewed studies demonstrate that users seek to understand the underlying coding process, including why codes are assigned and how disagreements among coders are resolved. The literature also stresses the necessity of communicating system reliability, allowing users to set realistic expectations about system outputs. Another recurrent theme across the literature is a lack of explicit optimization strategies or actionable guidance to enhance system performance, which often leads to user frustration. Drawing upon the findings from the literature review, we proposed a set of design goals aimed at improving the transparency and reliability of automated qualitative coding systems.

Following a synthesis of user needs derived from our review, we investigated the capacity of LLMs with the aim of translating identified user needs into actionable design recommendations. Based on the extracted design goals, we developed a baseline pipeline for qualitative coding, which was applied across four distinct qualitative datasets. This pipeline adopts a multi-agent architecture where AI agents simulate the collaborative coding process of human coders. The experiment results demonstrated how factors such as dataset characteristics, labeling uncertainty, and coder consensus levels influence the effectiveness of LLM-based systems. For example, coding accuracy improves when the codebook is more concise and when AI coders engage in more intense disagreements during the coding process. Additionally, by analyzing and comparing the behaviors of LLMs and human coders, we identified strategies for developers and designers to adjust LLMs for automated coding tasks. The analysis revealed that while LLMs replicate basic human interaction patterns, such as acknowledging other discussants’ perspectives and proposing alternatives, they fall short in nuances, such as questioning to clarify uncertainties or drawing on personal experiences.

These findings culminated in a set of design recommendations for creating LLM-based qualitative coding systems. The recommendations introduce strategies for designers, including guiding users of systems through data preprocessing, leveraging uncertainty and conflict as diagnostic signals for accuracy, and tuning agent discussion styles based on task ambiguity. These recommendations serve as a conceptual toolkit that helps users better understand and interact with model outputs. To advance the field, we made our baseline pipeline and the LLM agents’ discussion dataset publicly available.

The contributions of our work can be summarized as below:

- Design goals for automated qualitative coding systems based on exploratory literature review
- Design of baseline pipeline for LLM-based qualitative coding
- Discussion dataset of LLM coders across four qualitative coding tasks
- Data analysis of LLM capabilities and factors influencing coding accuracy
- Design recommendations for LLM-based qualitative coding systems

2 Related Work

Even for experts, qualitative coding involves considerable cognitive and temporal investment, owing to the need for detailed scrutiny and deep contextual understanding of nuanced data [77, 80]. Specifically, our work focused on

deductive coding, which applies predefined theoretical frameworks to categorize and interpret qualitative data. This section addresses two dimensions in the evolution of qualitative coding systems: the automation of qualitative coding through NLP techniques and the role of human users in engaging with these automated systems.

2.1 Automation of Qualitative Coding

Efforts to automate qualitative coding in NLP have evolved from simple rule-based approaches to interactive machine-learning systems. Marathe and Toyama [102] proposed a semi-automated tool using basic NLP methods like keyword matching, while Chen et al. [27] introduced SVM models with a visualization on coder disagreement. Several studies broadened the scope of AI-assisted automated coding to encompass workflows that involve collaboration between multiple human coders [52, 61]. Building on these efforts, the advancement of LLMs has accelerated the development of automated qualitative coding techniques. For example, Dunivin et al. [54] and Ashwin et al. [7] demonstrated that chain-of-thought reasoning enhances coding accuracy alongside a few-shot learning method [171]. Several papers [31, 183] introduced stepwise prompt-guided coding and Liu et al. [94] highlighted the role of contextual data, such as dialogue flow and timestamps. Other studies have shown that prompt engineering can greatly enhance the accuracy of coding with LLMs. Examples include using precise codebook definitions in social science domains [69], adding coding rules informed by debates over borderline cases [106], and including the reasoning behind code assignments [149]. Following these advances in prompt engineering, researchers have explored ways to make the LLM-based coding process more rigorous and systematic to improve the precision and adaptability of coding. Subramaniam et al. [154] proposed DebateGPT, a multi-agent system refining outputs through summarization and confidence scoring. Bryda and Sadowski [21] iteratively adjusted prompt design to guide AI in deductive coding with clarity and consistency. Meanwhile, recognizing the need for methodological rigor and reproducibility, Tai et al. [156] introduced a metric to measure coding consistency over repeated runs, and several other studies [17, 84] also ran multiple repeated LLM experiments to ensure reliability and robustness in their findings.

Existing research is confined to experiments on individual datasets, leaving a gap in providing a comprehensive guide for designing scalable and generalizable automated qualitative coding systems. Furthermore, as far as we are aware, there is no existing research that takes an analytical approach to how LLMs perform under varying conditions (e.g., codebook length, complexity of data to be coded). This lack of analysis of the factors influencing LLM performance makes it difficult for system developers and designers to optimize LLM performance, leaving it as an unguided trial-and-error process. Our study seeks to bridge this gap by performing a statistical analysis of LLM capabilities in qualitative coding across diverse datasets. By analyzing the relationship between dataset features and LLM performance, our work offers a foundation for understanding and optimizing LLM capabilities in qualitative coding.

2.2 Humans' Roles in Automated Qualitative Coding

Despite advancements in technologies for automating qualitative coding, practical adoption may remain limited if the roles and needs of human users are overlooked. In this context, prior research highlights efforts to address users' needs and challenges in automating the process of coding. Through interviews, Zhang et al. [182] uncovered several user concerns regarding the use of LLMs in qualitative coding, such as a lack of transparency and limited understanding of LLMs' capabilities. Similarly, Chen et al. [28] highlighted users' concerns regarding the black-box nature and lack of transparency in automation models.

Jiang et al. [77], based on interviews with 17 qualitative researchers, suggested that AI-based tools should embrace ambiguity, enhance human agency, and encourage serendipitous discoveries instead of removing uncertainty. In light of

these findings, further research has emphasized the active role of users in addressing limitations and refining automation processes. For instance, a few studies stressed the indispensability of user input in automated qualitative coding [35], such as reflecting and adapting the automated results [83] and interpreting nuances of data [21, 23, 109]. Preserving user agency [80, 153] and ensuring transparency [57, 74, 136] have also been identified as key features of automated systems. These perspectives collectively highlight the importance of designing systems that not only streamline the coding process but also foster human-in-the-loop collaboration and maintain user control. In this thread, Zambrano et al. [179] have explored the use of AI agents as collaborative coding companions, and another study [14] showed that human-AI collaboration achieves higher accuracy compared to using AI or human coders alone. Meanwhile, interactive tools like PaTAT [64] and CollabCoder [62] exemplified mixed-initiative approaches, allowing users to iteratively guide AI while retaining autonomy.

These approaches align with the primary goal of qualitative research, which is to foster a deeper understanding of qualitative data rather than simply speeding up the coding process. Basit et al. [15] further underscored that the goal of qualitative coding in thematic analysis is not merely to label data but to support researchers in discovering, supporting, and confirming their theories through the data. Meanwhile, challenges of automated systems remain as Bano et al. [12] noted the limitations of LLMs in understanding context and nuances, cautioning against AI agents fostering a single, echo-chamber perspective. Ziemis et al. [187] evaluated LLMs' capabilities on diverse benchmarks, showing potential to enhance computational social science but highlighting the need for human oversight in the process.

Building on previous studies, our exploratory literature review investigates user requirements and expectations in automating qualitative coding. We then designed LLM experiments based on these findings. By comparing the findings of the exploratory literature review with the outcomes of the experiments, we formulate concrete design recommendations for automated qualitative coding systems.

By addressing these gaps, our study lays the groundwork for an integration of LLMs into qualitative research practices through the investigation of user needs, statistical analysis, and design recommendations.

3 Exploratory Literature Review

Design Goals (DG)	Supporting Literature
DG1. Provide reasons and evidence for how codes are assigned.	[2, 6, 7, 11, 12, 16, 19, 22, 26, 27, 31, 34, 35, 40, 42, 44, 45, 57, 63, 64, 70, 71, 74, 77, 78, 81, 84, 87, 89, 98, 104, 116, 118–121, 125, 127, 128, 131, 135, 136, 140, 147, 149, 156, 159, 162–165, 172, 174, 175, 181, 182, 186]
DG2. Illuminate uncertainties and interpretative variability in code assignments.	[2, 5, 8, 10, 11, 21, 23, 26, 27, 31–33, 35, 35, 36, 40–46, 48, 54, 55, 58, 60–62, 64, 66, 70, 72, 73, 76, 77, 80, 83–86, 91, 97, 98, 103–105, 109–121, 123–127, 131–133, 135, 136, 138–140, 147, 149–151, 153, 156, 159, 163, 169, 174, 177, 185, 186]
DG3. Enable users to calibrate expectations for coding accuracy.	[10, 35, 60, 61, 64, 71, 74, 78, 85, 103, 119, 120, 171, 177, 181, 182]
DG4. Facilitate users' understanding of how data characteristics affect coding accuracy.	[2, 5, 7, 8, 12, 17, 20, 21, 31, 38, 43, 54, 57, 60, 62, 65, 68, 70, 72, 84, 93, 103, 107, 110, 118, 119, 128, 130, 136, 149, 163, 165, 175]

Table 1. An overview of the design goals for automated qualitative coding systems and literature that informed their formulation.

The exploratory literature review explores user needs and expectations for automated qualitative coding systems, aiming to derive design goals that inform the development of a baseline LLM-based coding pipeline. We adopted the PRISMA statement [108] for review.

3.1 Review Methodology

3.1.1 Identification Process. For the identification of relevant literature, we conducted a database search using the following keywords: (“AI” OR “artificial intelligence” OR “LLMs” OR “large language models” OR “automated”) AND (“qualitative coding” OR “qualitative analysis” OR “thematic analysis” OR “thematic coding”). This search yielded an initial set of 988 articles, which were sorted and selected based on their relevance to the specified keywords.

3.1.2 Screening Process. One of the authors reviewed the titles and abstracts of all retrieved papers. Papers were excluded if their focus was unrelated to automated qualitative coding or if they were not written in English. For example, papers using the term “coding” in the context of computer programming rather than thematic analysis were removed. This screening procedure resulted in a final set of 180 relevant papers.

3.1.3 Assessing Eligibility. To determine the eligibility of relevant literature, one of our authors read full articles of 180 papers. Through an iterative process of reading and analysis, the author developed the following exclusion criteria:

- Papers were excluded if the primary focus was not on automated qualitative coding. For instance, studies that employed qualitative methods such as interviews or surveys as part of broader research themes (e.g., AI in Education: A Qualitative Analysis) were omitted.
- Papers that discussed solely human-driven coding practices without reference to automated systems were excluded (e.g., Analyzing Qualitative Data via Collaborative Coding: A Practical Guide).
- Papers were excluded if they lacked explicit discussions of user expectations or design considerations in automated qualitative coding systems. This includes works that merely presented use cases of technologies or technical architectures without addressing user-side design requirements.

The exclusion criteria were refined throughout the iterative review process. As a result of this eligibility evaluation, a final corpus of 118 papers was selected for analysis. The complete list of reviewed papers is provided in the Supplementary Materials.

3.2 Analysis

Given that our aim was to inform design goals for our baseline pipeline of LLM-based qualitative coding, we conducted an exploratory review rather than a comprehensive, systematic literature review. One of our authors generated preliminary codes through a full reading of the corpus, followed by a second round of analysis to finalize the codebook. We then invited a qualitative research expert with a PhD in sociology to independently assess the codebook against the full corpus.

3.3 Review Results

We present the design goals identified through the review of prior literature. Table 1 provides an overview of these goals and their supporting studies.

3.3.1 Transparency. Coding automation should help users make informed decisions by presenting outputs that are transparent in their reasoning and limitations, including an explanation of how results are derived and any associated uncertainties.

DG1. Provide reasons and evidence for how codes are assigned. Prior work emphasizes that automated coding systems should offer explanations for their coding outputs [136, 181] to support user understanding of AI reasoning [35, 64]. To reflect this, we prompted the LLM to include justifications when generating codes.

DG2. Illuminate uncertainties and interpretative variability in code assignments. Literature highlights the importance of exposing uncertainties in AI-generated codes, often represented through confidence scores [64, 85, 91]. Furthermore, prior work underscores interpretive variability, across both human and AI coders [41, 149]. Based on this, we used two LLM coders to perform coding independently and resolve disagreements through discussion, mimicking human coding processes. We instructed the models to estimate confidence level, allowing LLM coders to assign an ‘Undecidable’ label when uncertain.

3.3.2 Reliability. Automated coding systems should enable users to better understand the system’s capabilities and develop strategies to optimize performance.

DG3. Enable users to calibrate expectations for coding accuracy. Users often struggle when they cannot align expectations with system performance [35, 120]. In line with established approaches [12, 177], we used coder disagreements during discussions as a proxy for evaluating automated coding accuracy.

DG4. Facilitate users’ understanding of how data characteristics affect coding accuracy. Literature suggests that factors such as codebook complexity [21], the abstractness of language [118, 130], and data similarity [2, 17] affect coding accuracy. Our study investigated how variations in data and codebook (e.g., length, similarity, and specialization) impact coding reliability.

4 Large Language Model Experiments and Data Analysis

In this section, we report the pipeline and results of LLM experiments designed based on the design goals. Our goal is to examine the capabilities of LLMs in qualitative coding by evaluating their performance across different datasets.

4.1 Datasets

We conducted experiments using LLMs across four distinct domains: education, law, sociology, and medicine, utilizing datasets from prior studies. The datasets include a dialogue corpus in learning [166], cases from the European Court of Human Rights [24], open-ended interview responses [141], and medical paper abstracts [144]. To standardize dataset sizes, we randomly selected five labels from each corpus and sampled 500 instances per label. Meanwhile, we have released the prompt design, coding pipeline, and LLM-generated discussions as open-source materials¹.

4.2 Design of Coding Pipeline

Based on the exploratory literature review, we designed an automated coding pipeline powered by AI agents. The process mirrors human coding in that the AI coders first independently code the entire dataset and two AI coders engage in discussions about conflicting cases (DG2). In the discussion process, they aim to resolve conflicts and derive new labeling rules based on the discussions.

¹<https://anonymous.4open.science/r/qual-coding-llm-experiment-1DA3/README.md>

An important note is that, before any of these steps, an initial phase of extracting excerpts to be coded was conducted. This data preprocessing step aimed to standardize the units of analysis. In this step, the LLM was supplied with the codebook and the data to be coded, and it was instructed to extract data excerpts likely to correspond to the codes in the codebook. As a result, only the extracted excerpts, rather than the entire qualitative dataset, were used for coding. This approach aligns with methodologies discussed in recent studies, which emphasize the importance of preprocessing to improve the reliability of qualitative coding by ensuring only pertinent excerpts are subjected to analysis [60, 146, 159].

The prompts used in the coding process were structured as follows. AI coders were instructed to justify their coding decisions (**DG1**) and had three possible response options: the code is present in the data, absent from the data, or undecidable (**DG2**). We grounded the prompt design in prior work, adhering to principles of constructive discussions [49, 142, 155] and efficient AI-agent dialogue frameworks [25, 88, 152, 168]. To reduce ambiguity in the coding task for AI models, we followed established conventions in the literature by including only one code per coding cycle and repeating the process for each code individually. As the discussion progressed through multiple rounds and turns, LLM coders were encouraged to formulate their answers by considering both their previous reasoning and the other coder’s perspectives. If disagreements arose initially, but consensus was achieved through rounds of discussions, the coders collaboratively developed a new labeling rule to resolve similar disputes in the future. Each back-and-forth exchange constituted one round of discussion. If no agreement was reached after three rounds, the case was marked as “disagreed,” and the process moved on to the next case. To avoid ordering bias, we adhered to methodologies established in reasoning and QA frameworks for AI-agent discussions [53]. We utilized OpenAI’s ChatGPT with the ‘gpt-4o-mini’ model for this process. The temperature settings were 0 for coding and 0.7 for discussions. These configurations align with previous studies on qualitative coding, but future researchers may adjust them to meet their objectives. After completing the discussion phase, both AI coders re-coded the entire dataset using shared knowledge derived from discussions and newly established labeling rules. At each coding round, we calculated inter-rater reliability (IRR) using Cohen’s Kappa.

4.3 Evaluation Method

Following the pipeline design, we addressed data imbalance using an undersampling technique [95] and computed F1 scores to assess coding accuracy. We then examined the relationship between coding accuracy and factors related to datasets and AI-agent discussions (**DG4**). These factors included dataset characteristics, such as length of codebooks, semantic similarity of content within codebooks, length of data excerpt to be coded, semantic similarity of content within excerpts, and disparity of specialization degree between codebooks and excerpts. Additionally, labeling dynamics, such as labeling uncertainty, degree of consensus and controversy between coders, were also considered (**DG3**), drawing upon the literature detailed in Table 2, 3, 4, and 5.

We quantify corpus’ semantic proximity using cosine similarity, one of the most common similarity metrics, enabling a scale-invariant comparison of textual representations [134]. Meanwhile, to measure the degree of specialization of corpus, we adopted a frequency-based approach that follows practice in corpus linguistics and psycholinguistics, lower reference-corpus frequency indicates higher lexical difficulty [18, 29]. Specifically, we used the Corpus of Contemporary American English (COCA) to compute a Word Difficulty Score (WDS) for each word, defined by the inverse frequency of the word [39, 157].

For statistical analysis, we adopted Mixed-Effects Models, an approach designed to handle both fixed and random effects, making it particularly suited for grouped or hierarchical data [1, 100, 143]. Given the grouping structure of our data — comprising four datasets with five labels in each dataset — we modeled these groupings as controlled

random effects. Model appropriateness was verified through assessments of VIF, ICC, and residual normality. In addition, we conducted a qualitative evaluation of AI coder discussion patterns. Using the Thomas-Kilmann Conflict Mode Instrument [160] and frameworks for discursive moves [167, 178], we compared the behaviors of human and AI discussions. One of our authors iteratively coded the patterns in AI-generated discussions and resolved disagreements.

4.4 Statistical Analysis Results

Across all datasets and labels, the IRRs between two AI coders exceeded 0.85. The F1 scores ranged from 0.31 to 0.89 depending on dataset and label, with an average of 0.68 and a standard deviation of 0.16. From this data, we specifically analyzed cases where either one of the two AI coders assigned the ‘Undecidable’ label or where the two coders provided conflicting results, estimating the accuracy of these cases. The average accuracy was 0.60 with a standard deviation of 0.25, highlighting the tradeoff between coverage and accuracy. Lower confidence levels and consensus degrees resulted in reduced accuracy. Full evaluation details are available in the Supplementary Material.

The Mixed Effects Model analysis revealed how coding accuracy is influenced by data characteristics and coding dynamics, as summarized in Table 2, 3, 4, and 5. Detailed statistics, including minimum, maximum, average, and standard deviations for the analysis, are available in the Supplementary Materials.

4.4.1 Relationship between Coding Accuracy and Dataset Characteristics.

Metric	Dataset Characteristic	P-value (T-statistic, Effect Size)	Relationship	Reason and Interpretation	Relevant Literature
Accuracy of Initial Coding	Length of Codebook	< 0.0001 **** (-11.702, -0.287)	The shorter the codebook, the higher the initial accuracy of automated coding.	LLMs’ ability to retrieve and retain relevant information from the input declines as the context length increases.	- The performance of LLMs degrades when identifying relevant information within lengthy contexts [90, 92]. - As context length increases, the model’s ability to focus on the most relevant information can decrease, leading to a “dilution” of attention across the input [173]. - To create a good codebook in qualitative data analysis, ensure it is clear and concise [14, 47, 137].

Metric	Dataset Characteristic	P-value (T-statistic, Effect Size)	Relationship	Reason and Interpretation	Relevant Literature
	Cosine Similarity of Code-book	< 0.0001 **** (-11.738, -0.276)	The less similar the semantics of words in each codebook, the higher the initial accuracy of automated coding.	Describing code's meaning in varied language provides multiple perspectives and contexts for the code compared to using overly specific or narrow language.	Not Applicable
	Length of Excerpt	< 0.0001 **** (-9.249, -0.438)	The shorter the excerpt to code, the higher the initial accuracy of automated coding.	Concise data contains less noise from extraneous details, making it easier for AI to focus on the main idea.	- The performance of LLMs degrades when identifying relevant information within lengthy contexts [90, 92]. - The long and unsegmented text segments made coding decisions more complex [75].
	Cosine Similarity of Excerpt	< 0.0001 **** (7.372, 0.189)	The more similar the semantics of words in excerpt to code, the higher the initial accuracy of automated coding.	Limited or consistent topics of data to code make it easier for the AI to identify patterns and relationships within the data.	- LLMs perform better in cases where fewer, more concentrated topics allow the model to capture relevant information easily [51].

Metric	Dataset Characteristic	P-value (T-statistic, Effect Size)	Relationship	Reason and Interpretation	Relevant Literature
	Gap of Degree of Specialization Between Codebook and Excerpt	< 0.0001 **** (-7.629, -0.331)	The smaller the gap of degree of specialization between codebook and excerpt to code, the higher the initial accuracy of automated coding.	Text alignment in NLP involves mapping corresponding segments between two texts by measuring their similarity and connecting the most similar parts. Qualitative coding requires similar alignment between the codebook and data. A gap in technical terminology use between the two lowers the similarity score, which can reduce coding accuracy.	- Text alignment relies on the assessment of similarity to establish accurate correspondences across different text segments [82, 180]. - Significant terminological differences between sources — such as technical versus lay language — impair the model’s ability to form coherent personas. Lacking shared features like vocabulary and style, the model struggles to maintain cross-domain accuracy and truthfulness [79].

Table 2. Dependency on Dataset Characteristics. This table summarizes how different dataset characteristics affect the accuracy of AI agents in initial coding. The summary of the analysis, conducted using a Mixed Effects Model, includes statistical significance, reasons and interpretations of the findings, and connections to relevant literature. Green cell indicates positive, red negative, and gray no relationship with coding accuracy. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$.

Table 2 illustrates the relationship between dataset characteristics and coding accuracy. Lengthier codebooks, higher cosine similarity within the codebook, lengthier excerpts, and greater gaps in specialization between codebooks and excerpts all negatively affect accuracy. In contrast, higher cosine similarity within the excerpt positively impacts coding accuracy.

We observed that concise codebooks improve AI coders’ labeling accuracy, as LLMs struggle to retrieve and retain information with extended input contexts (Table 2, Row 1). This finding aligns with previous studies emphasizing the efficiency of concise inputs for LLMs [14, 47, 90, 92, 137, 173]. One should be cautious in interpreting this finding, as straightforward codes often require less detailed descriptions, whereas complex codes necessitate nuanced, lengthy explanations. This means that the observed high accuracy with concise inputs may partly reflect the complexity of the code content itself, rather than the conciseness alone. To account for this, we incorporated specialized terminology and Word Sense Disambiguation (WSD) to minimize the complexity’s influence. This further analysis mitigated the potential confounding effects of complexity, which our analysis found to be statistically insignificant. The same pattern was also observed in the data to be coded (Table 2, Row 3), where AI coders performed more accurately with concise data.

The analysis revealed opposing effects of cosine similarity within the codebook (Table 2, Row 2) versus the data to be coded (Table 2, Row 4). Codebooks employing diverse languages improved accuracy, as variety aids in contextual

understanding of the code. Conversely, consistency in topics within the data to be coded improved accuracy, as fewer topics made it easier for the model to capture relevant information easily [51]. To summarize, coding accuracy improves when codebooks employ diverse language while the target data remains consistent. Since the coding process requires determining whether individual data points fall within codebook categories, this difference in linguistic breadth — varied vocabulary in the codebook and the homogeneity of languages in data — leads to more accurate coding results. It suggests that designing a corpus requires balancing diversity with specificity, depending on the role of each corpus. This balance highlights the complementary nature of broad comprehension (for codebooks) and targeted application (for data).

Additionally, drawing parallels from text alignment tasks in NLP domain, which involves mapping corresponding text segments across sources, we explored the impact of similarity alignment between codebooks and data. The result showed that significant disparities in technical terminology diminished similarity alignment, impairing coding accuracy (Table 2, Row 5). This result corroborates existing research showing that differences in terminology reduce LLM efficacy [79, 82, 180].

4.4.2 Relationship between Coding Accuracy Increase and Labeling Uncertainty.

Metric	Labeling Uncertainty	P-value (T-statistic, Effect Size)	Relationship	Reason and Interpretation	Relevant Literature
Accuracy Increase via Discussions	Number of Undecidable Labels in Initial Coding	< 0.01 ** (3.274, 0.123)	The higher the number of ‘Undecidable’ labels before discussion, the greater the accuracy improvement through AI agent discussions.	Dialogue process naturally clarifies uncertainties and builds a precise, shared understanding, thereby improving coding accuracy.	Language model performance is enhanced by emphasizing the uncertain data during the training process [148, 176].
	Number of Undecidable Labels in Discussion-Prompted Coding	< 0.0001 **** (-14.476, -0.369)	The lower the number of ‘Undecidable’ labels after discussion, the greater the accuracy improvement through AI agent discussions.	Low frequency of ‘Undecidable’ labels serves as an indicator of the effect of AI agent discussions.	Low confidence and agreement among coders are indicators of poor performance [59].

Table 3. Dependency on Labeling Uncertainty. This table summarizes how labeling uncertainty among AI agents affects the accuracy improvements achieved through AI agent discussions. The summary of the analysis, conducted using a Mixed Effects Model, includes statistical significance, reasons and interpretations of the findings, and connections to relevant literature. Green cell indicates positive, red negative, and gray no relationship with coding accuracy. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$.

Table 3 presents the relationship between coding accuracy improvement through discussions and labeling uncertainty, measured by the frequency of ‘Undecidable’ labels assigned by AI coders. The results reveal a positive relationship between coding accuracy increase via discussions and the ratio of ‘Undecidable’ labels during initial coding before

discussions. In contrast, after discussion-prompted coding (i.e., coding performed by LLM coders using a prompt that included the two coders’ discussion process and the new labeling rule derived from it), this trend reverses, showing a negative relationship. The positive relationship suggests that greater ambiguity in the initial coding task enhances the effectiveness of AI agent discussions, as the dialogue process naturally clarifies uncertainties and builds a shared, precise understanding (Table 3, Row 1). These findings corroborate the studies that revealed the positive relationship between clarification of labeling uncertainty and language model performance [148, 176]. On the other hand, the frequency of ‘Undecidable’ labels persisted following the discussions serves as an indicator of the effect of AI agent discussions – more occurrences indicate a small improvement in accuracy through discussions (Table 3, Row 2). This result parallels findings from human coding studies, where low coder confidence levels often signal poor coding performance [59].

4.4.3 Relationship between Coding Accuracy Increase and Consensus Level of AI Coders.

Metric	Consensus among Coders	P-value (T-statistic, Effect Size)	Relationship	Reason and Interpretation	Relevant Literature
Accuracy Increase via Discussions	Number of Conflicting Labels in Initial Coding	0.059 (-1.888, -0.037)	No relationship between the number of conflicting labels before discussion and the accuracy improvement through AI agent discussions.	There is no evidence that the number of conflicting labels before discussion serves as an indicator of the accuracy improvement through AI agent discussions.	Not Applicable
	Number of Conflicting Labels in Discussion-Prompted Coding	< 0.0001 **** (-6.861, -0.147)	The lower the number of conflicting labels after discussion, the greater the accuracy improvement through AI agent discussions.	Low frequency of conflicting labels serves as an indicator of the effect of AI agent discussions.	Low confidence and agreement among coders are indicators of poor performance [59].

Table 4. Dependency on Consensus Among Coders. This table summarizes how consensus among AI agents during coding affects the accuracy improvements achieved through AI agent discussions. The summary of the analysis, conducted using a Mixed Effects Model, includes statistical significance, reasons and interpretations of the findings, and connections to relevant literature. Green cell indicates positive, red negative, and gray no relationship with coding accuracy. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$.

As shown in Table 4, coding accuracy improvement varies depending on the agreement levels between AI coders. Our analysis revealed that the number of conflicting labels before discussion did not influence the accuracy gains achieved through AI agent discussions (Table 4, Row 1). In contrast, a low frequency of conflicting labels after discussion is indicative of the positive impact of AI agent discussions (Table 4, Row 2). This is consistent with existing literature that identifies low agreement among human coders as a marker of poor qualitative coding accuracy [59].

4.4.4 Relationship between Coding Accuracy and Controversy Level of AI Coders.

Metric	Controversy in Discussions	P-value (T-statistic, Effect Size)	Relationship	Reason and Interpretation	Relevant Literature
Accuracy of Discussion-Prompted Coding	Number of Turns in Discussions	< 0.0001 **** (7.997, 0.168)	The more controversial the discussions, the higher the accuracy of discussion-prompted coding.	Discussions about controversial cases reveal nuanced details that clarify the scope of the code's application.	To create a good codebook in qualitative data analysis, include clear inclusion and exclusion criteria for each code [101, 137].
	Ratio of Resolved Conflicts	< 0.0001 **** (-6.300, -0.139)	The lower the agreement ratio in the discussions, the higher the accuracy of discussion-prompted coding.	Debates with disagreements reveal nuanced details that clarify the scope of the code's application.	
	Ratio of Correctly Resolved Conflicts	< 0.0001 **** (-9.720, -0.169)	The lower the number of correctly resolved conflicting cases via discussions, the greater the accuracy of discussion-prompted coding.	Small number of correctly resolved conflicting cases indicates that the AI agent discussions involved a high proportion of challenging and borderline cases. These discussions about borderline cases reveal nuanced details that clarify the scope of the code's application.	Not Applicable
	Ratio of Competing Modes in Discussions	< 0.01 ** (2.789, 0.054)	No relationship between the ratio of competing modes of AI agents in discussions and the accuracy of discussion-prompted coding.	There is no evidence that the ratio of competing modes of AI agents in discussions serves as an indicator of the accuracy of discussion-prompted coding.	Not Applicable

Metric	Controversy in Discussions	P-value (T-statistic, Effect Size)	Relationship	Reason and Interpretation	Relevant Literature
	Ratio of Compromising Modes in Discussions	< 0.0001 **** (-4.251, -0.066)	The less compromising the discussions, the higher the accuracy of discussion-prompted coding.	Discussions about controversial cases, which provokes less compromising discussions among AI agents, reveal nuanced details that clarify the scope of the code's application.	- Addressing conflicts in a collaborative, not competitive manner turns disagreements into chances for improved decision-making [161]. - Collaborating mode is desired in coding discussion, seeking for consensual decision [160].
	Ratio of Collaborating Modes in Discussions	< 0.0001 **** (-8.420, -0.255)	The less collaborating the discussions, the higher the accuracy of discussion-prompted coding.	Discussions about controversial cases, which provokes less collaborating discussions among AI agents, reveal nuanced details that clarify the scope of the code's application.	

Table 5. Dependency on Controversy in Discussions. This table summarizes how the level of controversy in AI agent discussions affects the accuracy of AI agents when prompted with the discussion process. The summary of the analysis, conducted using a Mixed Effects Model, includes statistical significance, reasons and interpretations of the findings, and connections to relevant literature. Green cell indicates positive, red negative, and gray no relationship with coding accuracy. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$. Note that the term “competing/compromising/collaborating mode” relates to the content’s characteristics, not the discussion styles of the AI agents, as we did not adjust the AI agents’ discussion personas.

Table 5 illustrates the relationship between accuracy of discussion-prompted coding and the degree of controversy between AI coders. Accuracy was positively related to the number of discussion turns, with contentious disagreements extending debates and leading to an increase in turns. In contrast, accuracy showed a negative relationship with the ratio of resolved conflicts, correctly resolved conflicts, as well as discussions conducted in compromising and collaborating modes. There was no significant relationship observed with the ratio of competing modes.

The high coding accuracy observed in controversial discussions can be attributed to the nuanced understanding generated during debates. Lengthier debates allowed for clarification of the inclusion and exclusion criteria for each code, aligning with principles for designing effective codebooks by exposing clear inclusion and exclusion criteria [101, 137] (Table 5, Row 1). Similarly, these effects were also prominent in discussions without consensus (Table 5, Row 2). Interestingly, a smaller number of correctly resolved conflicts correlated with higher accuracy, indicating that debates

about challenging or borderline cases uncovered subtle details that revealed the scope of code application (Table 5, Row 3). This underscores that the value of engaging in nuanced debates may outweigh the importance of achieving correct resolutions, as the process itself drives a deeper understanding of ambiguous cases.

Contrary to findings from human discussions, where collaborative modes are typically desired, which provide emotional support and psychological safety [160, 161], less compromise and collaboration among AI coders produced better results (Table 5, Rows 5 and 6). This discrepancy can be explained by the nature of LLM agents, which operate free of emotional and social dynamics. Competitive discussions, which might induce stress in humans, function as optimization tools for LLMs as they can drive more rigorous refinement of reasoning.

Additionally, LLMs typically share the same foundational architecture and training data, thereby sometimes generating similar opinions among agents. Accordingly, collaborative modes can lead to overly convergent answers and reinforce biases [3, 12, 37].

4.5 Qualitative Observation of AI-Agent Discussions

Discursive Move of LLMs	Description	Discursive Move of Humans in Literature	Examples from AI Agent Discussions
Stance Maintenance	Reaffirming a viewpoint to emphasize commitment to a particular perspective.	Clarification [9, 25, 67]	“I maintain that it should be labeled as 1 for the presence of a check question.” “However, I still believe it primarily serves a rhetorical purpose rather than acting as a direct check question.” “I still assert that the claim fundamentally advocates for the inclusion of minority rights within a multicultural framework.”

Discursive Move of LLMs	Description	Discursive Move of Humans in Literature	Examples from AI Agent Discussions
Acknowledgment Before Disagreement	Acknowledging the opposing coder's viewpoint before clearly expressing disagreement.	Tolerant Divergence [4, 30, 56, 99], Partial Concurrence [4, 30, 56, 99]	"I understand Coder 2's concerns about the nuances of the phrase ... However, I still assert that this utterance functions as a check question." "I understand your perspective, but I would argue that the mere presence of poor conditions alone does not meet the threshold of a violation of the right to life as defined in Article 2." "While I acknowledge Coder 2's point regarding 'severe cardiorespiratory collapse,' it is essential to clarify that this term alone does not constitute a cardiovascular disease as defined by the provided description, which focuses on chronic or specific conditions related to the circulatory system."
Softening Language	Using softer and cooperative language to maintain a positive and constructive atmosphere.	Affirming Communication [4, 30, 56, 99]	"I appreciate Coder 2's reevaluation and the acknowledgment of ambiguity." "Coder 1's assessment correctly identifies it as a form of agreement that confirms the sentiment previously discussed."
Reassessing Evidence	Revisiting the description of code, often mentioning specific lines to reinforce their position or make it more convincing.	Referenced Argumentation [4, 30, 56, 99]	"According to the requirements of D01, the utterance must specifically ask a check question." "The right to life [which is one of the codes in the codebook] encompasses not only protection from direct life-threatening actions but also the provision of a safe and healthy environment." "The essence of Article 5 [which is one of the codes in the codebook] is focused on actual deprivation of liberty through arrest or detention."

Discursive Move of LLMs	Description	Discursive Move of Humans in Literature	Examples from AI Agent Discussions
Direct Challenge	Confronting the opposing coder's perspective head-on, with clear language indicating a strong contrasting view.	Explicit Rejection [13, 56, 96], Constructive Disagreement [13, 56, 96]	"Coder 1's interpretation of the question as a check question misrepresents the essence of what constitutes a check question." "Coder 1's perspective overlooks this critical relationship, which justifies the labeling of Wilson's disease under digestive system diseases."
Alternative Proposal	Suggesting an alternative viewpoint.	Alternative Proposal [9, 25, 122], Reframing Perspective [9, 122]	"Therefore, it might be better categorized as an open-ended question [a third alternative option distinct from the views of Coder1 and Coder2] instead of a clear check question." "Therefore, I propose labeling it as undecidable case [a third alternative option distinct from the views of Coder1 and Coder2] due to this ambiguity."
Tentative Stance	Expressing an opinion with caution, acknowledging ambiguity or lack of certainty.	Tentative Expression [4, 30, 56, 99]	"Thus, it is ambivalent; it does not definitively confirm or refute the existence of guiding principles." "However, the ambiguity surrounding the intention behind 'right?' makes it difficult to firmly categorize this utterance as either a check question or not." "The lack of clear legal standards or precedents regarding the threshold for inhumane treatment in this context leaves room for interpretation."
Seeking Consensus	Showing willingness to align or compromise with others.	Mutual Agreement [25, 50, 145, 158, 184]	"I agree with Coder 2's revised assessment that the case presents complexities that make it difficult to categorize definitively under Article 5." "I concur with Coder 1's conclusion that this utterance should be labeled as 0."

Discursive Move of LLMs	Description	Discursive Move of Humans in Literature	Examples from AI Agent Discussions
Collaborative Reflection	Explicitly referencing how the interaction with the other coder's viewpoint has led to a new understanding or shift in thought.	Dual Perspective Recognition [4, 30, 56, 99]	"After considering Coder 1's perspective that the utterance does not ask a check question and instead presents a rhetorical comparison, ..." "Upon reevaluating the concerns raised in Coder 1's position, I acknowledge that hepatoblastoma, being a malignancy originating in the liver, is indeed related to the digestive system as the liver plays a critical role in digestion and metabolism."
Redundant Restatement	Restating an earlier argument without introducing new information or distinct perspectives.	Repetition [178]	Coder 1: For the code "Person" to be applicable, there should be a sense of emotional resonance or significance linked to that person in relation to the music, which is absent here. Coder 2: ... Coder 1: For the "Person" code to apply, there needs to be a layer of emotional resonance or significance that connects the person to the music, which is not evident in this instance.
Explicit Contrast	Directly contrasting their own reasoning with the other coder's point, highlighting the differences before affirming a stance.	Arguments by Comparison [167]	"While Coder 1 argues it lacks the explicit structure of a check question [the other coder's opinion], it does in fact seek clarification about the action [their own opinion]." "The utterance 'Could you clarify what you mean by 'having the waves together'?' does not ask a check question [the other coder's opinion]; instead, it requests clarification [their own opinion]."
Reflective Re-consideration	Indicating a review of prior discussion before concluding.	Partial Concurrence [4, 30, 56, 99]	"After reconsidering the earlier discussion ..." "After reviewing both opinions, ..." "After reflecting on the ongoing discussion, ..."

Discursive Move of LLMs	Description	Discursive Move of Humans in Literature	Examples from AI Agent Discussions
-------------------------	-------------	---	------------------------------------

Table 6. Comparison of Discursive Moves Between LLM Coders and Humans. This table summarizes the discursive moves observed in LLM coders’ discussions to those reported for humans in existing research. The table also includes detailed descriptions and examples for each move.

We observed the discussion behaviors of LLM agents without additional fine-tuning or instruction about discursive moves and compared them with human discursive moves, providing a baseline that other researchers can expand and tailor for diverse applications.

Table 6 contains the comparison between human discursive moves from the literature and the ones that are observed from AI coders’ discussions. The discursive moves that are absent from the AI coders but still used by humans are organized in the table in Appendix A. LLMs replicate many of the discussion patterns of human coders, though key distinctions remain evident. For example, LLMs emulate basic patterns of interaction, such as maintaining stances, acknowledging others’ perspectives, reassessing evidence, explicitly contrasting ideas, expressing a tentative stance, and proposing alternatives. These behaviors highlight their capacity for logical and systematic reasoning. Additionally, LLMs are capable of supporting essential aspects of constructive discussions, such as fostering collaborative reflection and seeking consensus.

However, certain key aspects of human discursive behavior are absent in LLMs. Human discussants actively influence the flow and direction of discussions through dynamic interactions, such as raising questions to address uncertainties (e.g., clarifying ambiguities or challenging evidence), expanding on others’ ideas, sharing personal experiences, and moderating discussions based on evolving conversational contexts. These actions reflect humans’ ability to adjust their communication based on subtle shifts in the discussion and the reactions of other participants. In contrast, LLMs tend to operate within a more fixed framework, often missing the fluidity and adaptability required to respond to implicit cues or evolving conversational dynamics.

Moreover, LLMs do not replicate the emotional expressiveness inherent in human discourse. Human discussants often incorporate sarcasm and complaints, which can disrupt the focus of discussions or lead to unnecessary tension. This absence of emotionally charged or overly critical tones in LLMs can sometimes foster a more neutral and focused environment for logical reasoning. Instead, agents use softening language to foster a cordial atmosphere during the discussion. However, LLMs do share certain non-constructive discursive behaviors with humans, such as redundant restatement, which reiterates their earlier arguments without introducing new perspectives.

Meanwhile, humans use rhetorical strategies, such as analogies, to enhance persuasiveness and concretize abstract concepts. By comparison, while LLMs are proficient in generating well-structured and logical narratives, they frequently lack the creative subtlety needed to craft compelling narratives.

It is worth noting that LLMs engage in significantly less questioning compared to humans. While humans frequently ask clarifying or probing questions to refine understanding or actively explore the perspectives of other participants, LLMs tend to prioritize delivering coherent and direct answers over engaging in exploratory dialogue and mutual learning.

The differences between humans and LLMs are not limitations but complementary features that can be harnessed for mutual reinforcement in collaborative settings. Researchers can design human-AI collaborations by integrating the

structured and efficient reasoning of LLMs with the creativity, emotional intelligence, and contextual awareness of humans. The discussion dataset we created enables the evaluation of LLMs’ baseline discussion capabilities without fine-tuning, offering a foundation for guiding and benchmarking automated qualitative coding methodologies.

5 Design Recommendations for Automated Qualitative Coding Systems

Based on results from LLM experiments and data analysis, we propose several design recommendations for developers and designers of automated qualitative coding systems. The quantitative findings of experiments, including mean, standard deviation, minimum, and maximum values, are available in the Supplementary Materials. However, the primary aim of design recommendations is not to specify rigid thresholds but to offer a conceptual and methodological foundation upon which future research may further extend by focusing more directly on quantitative metrics.

5.1 Design Recommendation for Designers 1: Provide users with guidance on preprocessing their datasets to optimize system performance.

One key recommendation is that designers should provide users with targeted guidance on how to preprocess their datasets to optimize system performance. This includes embedding features in the system that offer diagnostic feedback on dataset properties and provide actionable suggestions. For instance, systems can notify users when certain data characteristics are likely to result in reduced performance, such as when particular data for coding or code definitions may yield lower accuracy. In such cases, users can be advised to allocate additional scrutiny to the associated segments, or the system can generate preprocessing recommendations accordingly.

The example suggestions may address codebook development such as encouraging concise codebook creation. Below are representative examples from the data analysis.

<Example of Short Codebook>

To be labeled as “Article03”, the court case should violate the Article03. Article03: Prohibition of torture No one shall be subjected to torture or to inhuman or degrading treatment or punishment.

<Example of Long Codebook>

To be labeled as “Article06”, the court case should violate the Article06. Article06: Right to a fair trial

- 1. In the determination of his civil rights and obligations or of any criminal charge against him, everyone is entitled to a fair and public hearing within a reasonable time by an independent and impartial tribunal established by law. Judgment shall be pronounced publicly but the press and public may be excluded from all or part of the trial in the interests of morals, public order or national security in a democratic society, where the interests of juveniles or the protection of the private life of the parties so require, or to the extent strictly necessary in the opinion of the court in special circumstances where publicity would prejudice the interests of justice.*
- 2. Everyone charged with a criminal offence shall be presumed innocent until proved guilty according to law.*
- 3. Everyone charged with a criminal offence has the following minimum rights: ...*

Another recommendation for the codebook can include avoiding repetitive or circular code definitions, which result in a high similarity between words within a codebook, and instead employing multiple, distinct descriptions that have a low similarity between words within a codebook.

<Example of Low Similarity Codebook>

To be labeled as “Person”, the respondent should associate the track with a person such as a family member, a friend, the loss of a person, a group of people, or not being with a person or being alone.

<Example of High Similarity Codebook>

To be labeled as “Negative Evaluations”, the respondent should make a negative evaluation about the track or an aspect of the track.

Users may also be prompted to split data for coding into chunks of one to two sentences.

<Example of Short Excerpt>

So do you know what virtual reality is?

<Example of Long Excerpt>

This is a great question because it really gets to the difference between some of our different tools in our machine learning tool belt in addressing problems like this. So if we were to use a supervised learning classic classification approach, a person would need to think about those features and creatively come up with them in approach we call the kitchen sink approach, which is just try everything you can possibly think of and see what works.

The qualitative data for coding should then be organized around coherent or limited topics using NLP techniques such as TextTiling.

<Example of Low Similarity Excerpt>

Slightly like slipknot. It reminds me of the game “need for speed underground 2” a little. The repetitive riffs make me think of grinding the same missions or tasks at work or in certain games. There is a certain anger, or energy brought by the vocals that makes me knuckle down, but the guitar is mostly happy and upbeat. I went through a phase of listening to the GazettE, but I still think it’s good. Energetic and upbeat. I don’t get too distracted by the lyrics as I don’t speak Japanese. It helps me focus and power through. Nope, it’s just a song I quite like.

<Example of High Similarity Excerpt>

A sense of calmness and serenity. Because it puts my life into perspective when I feel anxious or worried about something. Because it calms me down and makes me think of only positive things, all negative things just fall away. Yes, it’s a Christian track, so when I play it, it makes me think of God and his promises, so it gives me a perspective and reminder that I am loved.

Furthermore, designers should encourage users to match the degree of specialization in their codebooks with that in their datasets.

5.2 Design Recommendation for Designers 2: Allow users to leverage (1) the “Undecidable” label and (2) the conflicting label as a performance indicator to predict potential accuracy improvements resulting from AI agent discussions.

Beyond preprocessing, systems should enable users to understand and use performance signals. In particular, the “Undecidable” and conflicting labels should be made interpretable as indicators of potential accuracy gains via AI agent discussions. For example, users can be guided to interpret the number of uncertainty prior to AI agent discussions as a potentially positive sign, and the number of uncertain labels after discussion as a negative indicator of limited benefit

from the process. Likewise, persistent conflict between agents following discussions should be treated as a potential signal of limited effects of AI agent discussions. Incorporating such diagnostic tools allows users to develop informed mental models of the system’s behavior and adjust their expectations.

5.3 Design Recommendation for Designers 3: Allow users to leverage the level of controversy within AI agent discussions as a performance indicator for evaluating the expected accuracy of discussion-prompted coding.

Another major recommendation addresses the value of the degree of agent disagreement. While disagreements might be intuitively understood as limitations, our findings suggest that they are often more informative than their resolution. Designers should reconceptualize agent disagreement during coding discussions not as a failure but as a signal of interpretive richness. Our analysis indicates that prolonged or unresolved disagreements between agents often yield clarification of coding criteria and foster a more nuanced understanding of edge cases. In some cases, the resolution of disagreements may correspond with low information gain, suggesting that users should capitalize on the presence and structure of disagreement rather than its resolution alone. Specifically, in gauging the expected accuracy of discussion-prompted coding, designers can guide users to interpret high volumes of conversational exchange between agents as a positive sign while viewing high proportions of resolved or even correctly resolved conflicts as potential negative indicators. Below are representative examples from the data analysis.

<Example of Controversial Discussion>

...

Coder 1: I appreciate Coder 2’s interpretation that “what about negative harmony” could serve as a discussion prompt. However, the nature of the phrase “what about” often leans towards rhetorical questioning rather than a genuine inquiry seeking specific information. ...

Coder 2: After considering Coder 1’s arguments, I maintain that the phrase “what about negative harmony” does indeed present a question. While Coder 1 suggests it functions more as a prompt, I argue that it invites exploration of a specific topic, which aligns with the essence of D02. ...

Coder 1: I appreciate Coder 2’s perspective that “what about negative harmony” could be seen as an invitation for discussion; however, I still contend that it lacks the qualities of a genuine inquiry. ...

Coder 2: I understand Coder 1’s view regarding the rhetorical nature of “what about negative harmony,” yet I continue to assert that it functions as an inquiry. ...

<Example of Non-Controversial Discussion>

...

Coder 1: While I understand Coder 2’s perspective that the excerpt discusses potential and challenges, I believe it still conveys specific information about opportunities in synthetic worlds and the limitations of current simulation technologies. ...

Coder 2: Upon reevaluating my position and considering Coder 1’s reasoning, I recognize that the utterance does indeed convey substantial information about the opportunities and limitations involved in creating synthetic fantasy worlds. Phrases like “there’s tons of opportunities for the synthetic case” and “we still don’t do people well simulated” provide clear insights into the current state of technology in this area. Therefore, I agree with Coder 1 that this aligns with the requirements for D09, as it effectively informs the listener about specific aspects of the topic.

5.4 Design Recommendation for Designers 4: Adjust AI agents' discussion style based on the intended objectives of the coding task.

A final recommendation focuses on discussion dynamics and the adjustment of agent behaviors according to the objectives of the coding task. Designers may consider providing users with the option to activate a discussion mode that deliberately introduces controversy in tasks where ambiguity is prevalent in the dataset. When a label is undecidable, confidence is low, or prior performance has been unreliable, the system can shift into a debate-enhancing mode that fosters disagreement, prompting agents to adopt explicitly opposing perspectives. Our findings suggest that introducing structured opposition within agent discussion improves the accuracy of discussion-prompted coding.

<Example of Discussion with Competing Mode>

Upon further reflection, I maintain my stance that the excerpt contains positive evaluations. ... While Coder 1 argues these are not explicit evaluations, they do reflect a positive judgment about the track and its context, fulfilling the criteria for Positive Evaluations as outlined in the description.

<Example of Discussion with Compromising Mode>

While I understand Coder 2's perspective regarding closed head injury being a significant condition, I believe we are dealing with an undecidable case here. ... Therefore, based on the provided data, it is difficult to definitively categorize it as either present or absent under this specific label [, suggesting the Undecidable label as an alternative].

<Example of Discussion with Collaborating Mode>

Upon reevaluating Coder 1's reasoning, I recognize that the term "nostalgic" serves as a descriptive adjective that evokes specific emotional qualities associated with the track. ... Therefore, I change my initial assessment to agree with Coder 1.

6 Discussion

This section examines the key findings of our work, acknowledges the study's limitations, and outlines directions for future research.

6.1 Summary of Key Findings and Implications

Our study reveals that LLM "coders" can engage in a structured dialogue to perform qualitative coding, with performance influenced by specific dataset characteristics and discussion dynamics. First, we found that coding accuracy is shaped by dataset's length and semantic similarity, and the gap of degree of specialization between codebook and data for coding. These findings highlighted that preprocessing codebooks and qualitative data for coding can boost AI coding performance.

Second, our results show that the frequency of "Undecidable" label and conflict between AI agents can be utilized as a performance indicator. These signals assist users in calibrating their expectations of automated systems. We discovered that instances of disagreement and debate between the AI coders can be a positive signal for high accuracy of coding. Rather than viewing disagreement as a failure, our study indicates that when the two AI coders disagreed on how to code a segment, the conversation often increased interpretive depth.

The AI's convergence is a double-edged sword: on one hand, it led to consistent consensus in most cases, but on the other hand, we suspect it might sometimes mask uncertainty. For instance, in a few discussions we examined, one agent proposed a code interpretation that the other accepted quickly. Such instances resulted in a consensus label

that both agents agreed on, yet the underlying ambiguity of the data might not have been fully resolved, leading to an incorrect labeling outcome. This behavior underscores a design consideration on how to interpret automated outcomes: simply because two AI agents agree after a short discussion does not guarantee the decision was correct or well-founded. Furthermore, prolonged exchanges were associated with a more nuanced understanding of edge cases, whereas quick resolutions sometimes signaled that less information was gained through discussions. Together, these findings ground our discussion in concrete evidence from the experiment, moving beyond abstract statements to actionable recommendations for design.

Third, our qualitative analysis of LLM discourse behavior versus human coder discussions uncovered both parallels and gaps. On one hand, the AI agents replicated many constructive discussion patterns observed in human teams. They maintained or defended their initial stances, acknowledged each other’s perspectives, contrasted differing interpretations explicitly, reassessed evidence, and proposed alternatives. These behaviors suggest that LLMs are capable of logical, systematic reasoning. On the other hand, we identified several human discussion strategies that were absent or weaker in the AI conversations. For example, human coders often interject questions to clarify uncertainties, challenge assumptions, or seek additional evidence; in our study, the LLM agents did not engage in clarifying or probing questions. Humans also dynamically adjust the discussion flow, moderate the direction of debate, share personal anecdotes, and resort to figurative language and rhetorical strategies to enhance engagement and clarity. In contrast, the AI discussions followed a more fixed pattern, not fully adapting to nuanced conversational cues or shifting strategy mid-discussion. Emotional or stylistic nuances present in human dialogue, such as sarcasm or frustration, were essentially absent from the AI’s exchanges. This lack of emotional reactivity in LLMs kept the tone neutral and focused, potentially creating a more consistently cordial environment for deliberation. Meanwhile, the AI agents did exhibit some less productive tendencies akin to humans, such as repetitive restatement of points without adding new information. In summary, LLM-based coders can mirror basic human reasoning moves but miss the adaptive and inquisitive depth of human interaction, suggesting that while LLMs excel at structured reasoning, they may fall short in demonstrating the more affective and socially attuned aspects of human discourse without additional fine-tuning.

6.2 Limitations

While our study provides encouraging evidence for LLM-based qualitative coding, several limitations temper our conclusions. First, the AI model we used was not fine-tuned for qualitative coding tasks or any domain-specific data. All agent behaviors were elicited through prompting alone. This choice allowed us to examine out-of-the-box capabilities, but it also means performance might be improved with customization. A fine-tuned model or one explicitly instructed in effective discussion strategies could produce different (potentially better) outcomes.

Second, the evaluation of coding accuracy and consensus has inherent challenges. We treated the original human-coded dataset labels or expert consensus as the “ground truth” for accuracy calculations. However, in qualitative research, coder disagreements can sometimes reflect multiple valid interpretations rather than clear-cut errors. Our use of accuracy as a metric, while necessary for quantitative analysis, oversimplifies this nuance. An AI-assigned code different from the ground truth was counted as incorrect, even if it might be a justifiable interpretation in context.

Third, the rapid pace of LLM development introduces another layer of complexity. New releases of LLMs may significantly alter the dynamics of agent discussions, potentially changing both the quality and style of coding collaboration. This evolving baseline makes it challenging to draw stable conclusions from experiments tied to a particular model version, and highlights the need for ongoing evaluation as the technology advances. Meanwhile, although our experiments were tied to GPT-4o-mini, the underlying dynamics — such as the tendency for accuracy gains to emerge

from prolonged debate rather than rapid consensus, with controversy about borderline cases clarifying inclusion and exclusion criteria for codes and — represent model-agnostic mechanisms. These dynamics are likely to persist across LLM families, since they stem from the interactional structure of multi-agent debate rather than from model-specific capabilities. At the same time, given the rapid evolution of LLMs, we frame our findings less as static benchmarks and more as design-oriented insights that persist across model generations.

Finally, our study setting was a simulated one: AI agents discussed with each other, and we compared their behavior to documented human coding discussions from literature. We did not conduct a user study of human analysts actually working with these AI agents. As a result, we have not directly measured human researchers’ trust, satisfaction, or workflow integration when partnering with AI coders. These human factors are crucial for real-world adoption and were beyond our current scope. The absence of direct human-in-the-loop experimentation is a limitation that we acknowledge and aim to address in future work.

6.3 Theoretical Implications: Rethinking Reliability and Agreement

Beyond practical design, our work invites HCI and computational social science researchers to reconsider some theoretical assumptions in qualitative research when one of the coders is an AI. In traditional qualitative coding, intercoder reliability (e.g., Cohen’s Kappa) between human coders serves as a check on consistency and code clarity. But what does reliability mean when a coder is an AI? We found that measuring agreement between two AI coders was useful for our analysis, yet it differs from human scenarios.

For one, AI coders (especially if they are instances of the same LLM model) may have correlated errors or shared biases, violating the assumption of independent judgments. In our study, both AI coders were based on the same underlying model architecture and were exposed to the same training data distribution, which is unlike two independent human coders with distinct backgrounds. This means a high agreement between AI coders might not guarantee the validity of the coding — they could be consistently wrong in the same way, implying that reliability statistics need reinterpretation.

Inter-rater reliability with an AI partner could be overly optimistic (if the AI’s decisions are rubber-stamped by a human or vice versa) or overly pessimistic (if the AI systematically challenges a human coder on edge cases that humans might ordinarily skip as noise). One theoretical extension suggested by our work is to develop new metrics that account for the discussion process. Instead of just percent agreement on final codes, one could measure the proportion of cases that required discussion to reach agreement, or the stability of the AI’s decisions before and after debate. These measures might capture the nuance of AI coding interactions better than a single agreement coefficient.

Additionally, we propose rethinking the role of codebooks and coding guidelines for AIs. When an AI is involved, perhaps the codebook should be written not just for human interpretability but also optimized for machine interpretability.

Another implication is how we define consensus. In human teams, consensus is often the end of discussion; in human–AI teams, we might deliberately choose not to fully automate consensus. It could be acceptable, or even desirable, to have some lasting disagreements, if those indicate areas for further human analysis. Theorists and practitioners should be open to the idea that a discord between AI coder or between a human and an AI coder could spark deeper insights.

Meanwhile, the distinction between what is ‘reasonable practice’ and what is ‘inherent to qualitative analysis’ illuminates a question: is qualitative coding reducible to a standardized analytic procedure, or is it constitutively interpretive? We take the latter view. The presence of ambiguity and disagreement is not a byproduct of methodological insufficiency of imperfect method but a constitutive condition of qualitative knowledge production. Our contribution is

to formalize its role in computational pipelines. By operationalizing disagreement and undecidability, we reposition what qualitative research has treated as unavoidable and generative into a measurable feature that informs the design of automated systems.

In summary, our research calls for an expanded notion of reliability and consensus in coding when integrating AI – one that values the process of reaching agreement (or not reaching it) as much as the final agreement itself.

6.4 Design Implications: LLM-Based Coding Systems

Our findings suggest several important design implications for systems that incorporate LLMs into qualitative coding workflows. One key design recommendation is to offer tunable AI behavior modes, specifically a toggle between a “strict” mode and a “debate” mode. In the debate-enhancing mode, the system could detect low-confidence scenarios – such as undecidable cases or potential coding conflicts – and then trigger divergent perspectives among AI agents. By encouraging structured disagreement or devil’s advocacy, the system can prompt deeper analytical engagement and help surface nuanced considerations that may be lost in overly harmonious exchanges. Our results indicate that such structured debate is especially beneficial for ambiguous or complex coding situations, improving the overall accuracy and depth of analysis. On the other hand, for routine or well-defined coding tasks, a concise exchange may be preferable. In these scenarios, a “strict” mode could streamline interactions. This dynamic control over the discussion mode offers alignment with the user’s task goals. For example, a complicated qualitative analysis might benefit from competitive and discursive outputs clarifying challenging cases, whereas quick coding for a simple survey might prefer efficiency.

While our study focused on AI–AI coder pairs, a future direction is the integration of human coders directly into the discussion loop. A human-in-the-loop design could better align the coding process to users’ goals and intent by allowing human users to guide and challenge AI-generated codes in real time.

Qualitative coding has long been theorized as more than a procedural act of labeling; it is an interpretive practice in which researchers mediate between data, theory, and their own positionality. Reflexivity – acknowledging how one’s standpoint and assumptions shape interpretation – is central to ensuring validity in this tradition. While our study demonstrates that LLMs reproduce discursive moves reminiscent of human coders (e.g., stance maintenance, acknowledgment, tentative stances), these moves operate as surface-level simulations of interpretive dialogue. They do not embody the meta-cognitive work of interrogating one’s assumptions, nor the positional disclosure that grounds qualitative validity. As a result, AI “agreement” may project an illusion of rigor while bypassing the epistemic labor that human researchers contribute through reflexive engagement.

For automated qualitative coding, this implies that validity must be reconceptualized as a co-constructed property of human – AI interaction. AI coders can highlight ambiguity, surface discursive conflicts, and generate systematic consistency, but reflexivity remains the province of human researchers. Rather than viewing LLM consensus as a substitute for interpretive rigor, we argue that its true value lies in provoking reflexive engagement: Why did the models disagree? What kinds of ambiguity do their debates reveal? How does a researcher’s theoretical stance shape the interpretation of such ambiguity? Designing systems that scaffold these questions – by embedding reflexive checkpoints, prompting positional notes, and visualizing where AI outputs diverge from human interpretive expectations—can transform automation from a tool of decontextualization into a partner in reflexive inquiry. In this way, the uniquely human aspects of coding are not displaced but re-centered, with AI disagreements serving as catalysts for deeper interpretive practice.

We also encourage longitudinal studies and domain-specific deployments of LLM-based coding to see how our findings hold in real-world practice. For instance, deploying an AI coding assistant in a social science research lab

for several months could uncover how human coders adapt their strategies when an AI is part of the team — do they come to trust the AI for certain codes more than others? Does the presence of an AI change how teams construct their codebook? Different domains (e.g., healthcare interviews vs. Twitter data vs. HCI user study feedback) may present unique challenges for LLM coders. Future work should examine a variety of contexts, potentially leading to specialized design guidelines.

Several directions emerge from our study that could extend and deepen this line of inquiry:

- **Richer analysis dimensions:** Future studies could vary the number of AI agents, assign them distinct personas or expertise areas, and experiment with multiple iterations of discussion to observe how arguments evolve or converge over time.
- **Tracking dynamics of iterative discussions:** Our study involved a single iteration of discussion. Meanwhile, investigating how the structure and tone of AI discussions change across multiple iterations could offer insights into discussion quality and decision-making patterns.
- **Customized discussion behaviors:** Based on observed behaviors, systems could adapt the agents’ conversational styles or debate strategies to match the user’s preferences or the coding context.
- **Human-in-the-loop pipeline design:** Designing workflows where human coders participate proactively — adding feedback to AI debates, inserting their own counterpoints, or steering the discussion toward specific frames — can make the process better align with the user’s intent and goals. Future tools could allow humans to reflect their opinions within AI debates, not just through final selections but as part of the ongoing discussion, enabling richer co-creation and mutual learning between humans and AI.

7 Conclusion

Our work contributes a foundational step toward understanding and improving automated qualitative coding using LLMs. By simulating coder discussions, examining the impact of data characteristics, and analyzing AI behaviors qualitatively, we provide a nuanced account of the strengths and limitations of LLMs in replicating human coding dynamics. Our findings emphasize that disagreement and ambiguity are not merely challenges but opportunities for deeper insight, and they should be strategically harnessed in system design. From these findings, we propose a set of design recommendations for designers: maintain concise and diverse codebooks, structure excerpts around coherent topics, use coding uncertainty like “Undecidable” labels as diagnostic tools for coding accuracy, and tune agent discussion style to suit task goals.

References

- [1] Alan Agresti, James G Booth*, James P Hobert*, and Brian Caffo*. 2000. Random-effects modeling of categorical response data. *Sociological Methodology* 30, 1 (2000), 27–80.
- [2] Sara Amani, Lance White, Trini Balart, Karan Watson, and Kristi Shryock. 2025. Applying Generative AI for Thematic Analysis: A Model for Researchers Exploring Qualitative Methods. *Available at SSRN 5128905* (2025).
- [3] Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th Conference on Creativity & Cognition*. 413–425.
- [4] Jo Angouri and Theodora Tseliga. 2010. “you HAVE NO IDEA WHAT YOU ARE TALKING ABOUT!” From e-disagreement to e-impoliteness in two online fora. (2010).
- [5] Shafullah Anis and Juliana A French. 2023. Efficient, explicatory, and equitable: Why qualitative researchers should embrace AI, but cautiously. *Business & Society* 62, 6 (2023), 1139–1144.
- [6] Clarissa Sabrina Arlinghaus, Charlotte Wulff, G unter W Maier, C Arlinghaus, C Wulff, and G Maier. 2024. Inductive coding with chatgpt-an evaluation of different gpt models clustering qualitative data into categories.

- [7] Julian Ashwin, Aditya Chhabra, and Vijayendra Rao. 2023. Using large language models for qualitative analysis can introduce serious bias. *arXiv preprint arXiv:2309.17147* (2023).
- [8] Amine Hatun Atas, Berkan Celik, Francesco Balzan, and Bahar Shahrokhian. [n. d.]. Automating Thematic Analysis with Multi-Agent LLM Systems. ([n. d.]).
- [9] Mikhail Mikhailovich Bakhtin. 2010. *The dialogic imagination: Four essays*. University of texas Press.
- [10] Muneera Bano, Rashina Hoda, Didar Zowghi, and Christoph Treude. 2024. Large language models for qualitative research in software engineering: exploring opportunities and challenges. *Automated Software Engineering* 31, 1 (2024), 8.
- [11] Muneera Bano, Didar Zowghi, and Jon Whittle. 2023. AI and human reasoning: Qualitative research in the age of Large Language Models. *The AI Ethics Journal* 3, 1 (2023).
- [12] Muneera Bano, Didar Zowghi, and Jon Whittle. 2023. Exploring qualitative research using LLMs. *arXiv preprint arXiv:2306.13298* (2023).
- [13] Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 251–261.
- [14] Amanda Barany, Nidhi Nasir, Chelsea Porter, Andres Felipe Zambrano, Alexandra L Andres, Dara Bright, Mamta Shah, Xiner Liu, Sabrina Gao, Jiayi Zhang, et al. 2024. ChatGPT for education research: exploring the potential of large language models for qualitative codebook development. In *International conference on artificial intelligence in education*. Springer, 134–149.
- [15] Tehmina Basit. 2003. Manual or electronic? The role of coding in qualitative data analysis. *Educational research* 45, 2 (2003), 143–154.
- [16] Issam Bennis and Safwane Mouwafaq. 2025. Advancing AI-driven thematic analysis in qualitative research: a comparative study of nine generative models on Cutaneous Leishmaniasis data. *BMC Medical Informatics and Decision Making* 25, 1 (2025), 1–14.
- [17] Rimke Bijker, Stephanie S Merkouris, Nicki A Dowling, and Simone N Rodda. 2024. ChatGPT for Automated qualitative research: content analysis. *Journal of medical Internet research* 26 (2024), e59050.
- [18] Hunter M Breland. 1996. Word frequency and word difficulty: A comparison of counts in four corpora. *Psychological Science* 7, 2 (1996), 96–99.
- [19] Edward Brent and Pawel Slusarz. 2003. “Feeling the Beat” Intelligent Coding Advice from Metaknowledge in Qualitative Research. *Social Science Computer Review* 21, 3 (2003), 281–303.
- [20] Mario Brondani, Claudia Alves, Cecilia Ribeiro, Mariana M Braga, Renata C Mathes Garcia, Thiago Ardenghi, and Komkham Pattanaporn. 2024. Artificial intelligence, ChatGPT, and dental education: Implications for reflective assignments and qualitative research. *Journal of Dental Education* 88, 12 (2024), 1671–1680.
- [21] Grzegorz Bryda and Damian Sadowski. 2024. From words to themes: AI-powered qualitative data coding and analysis. In *World Conference on Qualitative Research*. Springer, 309–345.
- [22] Harold Cameron. 2025. Breaking Up with NVivo: AI Delivers Faster, Deeper, and Better Qualitative Insights. (2025).
- [23] Ana Carolina Carius and Alex Justen Teixeira. 2024. Artificial Intelligence and content analysis: the large language models (LLMs) and the automatized categorization. *AI & SOCIETY* (2024), 1–12.
- [24] Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. *arXiv preprint arXiv:2103.13084* (2021).
- [25] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201* (2023).
- [26] John Chen, Alexandros Lotsos, Lexie Zhao, Caiyi Wang, Jessica Hullman, Bruce Sherin, Uri Wilensky, and Michael Horn. 2024. A Computational Method for Measuring “Open Codes” in Qualitative Analysis. *arXiv preprint arXiv:2411.12142* (2024).
- [27] Nan-Chen Chen, Margaret Drouhard, Rafal Kocielnik, Jina Suh, and Cecilia R Aragon. 2018. Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 1–20.
- [28] Nan-chen Chen, Rafal Kocielnik, Margaret Drouhard, Vanessa Peña-Araya, Jina Suh, Keting Cen, Xiangyi Zheng, Cecilia R Aragon, and V Peña-Araya. 2016. Challenges of applying machine learning to qualitative coding. In *ACM SIGCHI Workshop on Human-Centered Machine Learning*.
- [29] Xiaobin Chen and Detmar Meurers. 2016. Characterizing text difficulty with word frequencies. In *Proceedings of the 11th workshop on innovative use of nlp for building educational applications*. 84–94.
- [30] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 1217–1230.
- [31] Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. 2023. LLM-assisted content analysis: Using large language models to support deductive coding. *arXiv preprint arXiv:2306.14924* (2023).
- [32] Felix Chopra and Ingar Haaland. 2023. Conducting qualitative interviews with AI. (2023).
- [33] P Christou. 2025. Looking Beyond Numbers in Qualitative Research: From Data Saturation to Data Analysis. *The Qualitative Report* 30, 1 (2025), 3088–3100.
- [34] Prokopis A Christou. 2023. A critical perspective over whether and how to acknowledge the use of artificial intelligence (AI) in qualitative studies. *The Qualitative Report* 28, 7 (2023), 1981–1991.
- [35] Prokopis A Christou. 2023. How to use artificial intelligence (AI) as a resource, methodological and analysis tool in qualitative research? *Qualitative Report* 28, 7 (2023).
- [36] Prokopis A Christou. 2024. Thematic analysis through artificial intelligence (AI). *Qualitative Report* 29, 2 (2024).

- [37] Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. 2023. Simulating opinion dynamics with networks of llm-based agents. *arXiv preprint arXiv:2311.09618* (2023).
- [38] Laura Ann Chubb. 2023. Me and the machines: Possibilities and pitfalls of using artificial intelligence for qualitative data analysis. *International journal of qualitative methods* 22 (2023), 16094069231193593.
- [39] Brown Corpus. [n. d.]. Corpus of Contemporary American English. ([n. d.]).
- [40] António Pedro Costa. 2023. Qualitative Research Methods: do digital tools open promising trends? *Revista Lusófona de Educação* 59, 59 (2023).
- [41] Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging large language model for thematic analysis. *arXiv preprint arXiv:2310.15100* (2023).
- [42] Robert M Davison, Hameed Chughtai, Petter Nielsen, Marco Marabelli, Federico Iannacci, Marjolein van Offenbeek, Monideepa Tarafdar, Manuel Trenz, Angsana A Techatassanasoontorn, Antonio Diaz Andrade, et al. 2024. The ethics of using generative AI for qualitative data analysis. *Information Systems Journal* 34, 5 (2024).
- [43] Matheus de Moraes Leça, Lucas Valença, Ronnie de Souza Santos, and Reyndre Santos. 2024. Applications and Implications of Large Language Models in Qualitative Analysis: A New Frontier for Empirical Software Engineering. *arXiv e-prints* (2024), arXiv-2412.
- [44] Stefano De Paoli. 2023. 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 preprint arXiv:2305.13014* (2023).
- [45] Stefano De Paoli. 2024. Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review* 42, 4 (2024), 997–1019.
- [46] Stefano De Paoli and Walter S Mathis. 2024. Reflections on inductive thematic saturation as a potential metric for measuring the validity of an inductive thematic analysis with LLMs. *Quality & Quantity* (2024), 1–27.
- [47] Jessica T DeCuir-Gunby, Patricia L Marshall, and Allison W McCulloch. 2011. Developing and using a codebook for the analysis of interview data: An example from a professional development research project. *Field methods* 23, 2 (2011), 136–155.
- [48] Andreas Dengel, Rupert Gehrlein, David Fernes, Sebastian G^orlich, Jonas Maurer, Hai Hoang Pham, Gabriel Großmann, and Niklas Dietrich genannt Eisermann. 2023. Qualitative research methods for large language models: Conducting semi-structured interviews with ChatGPT and BARD on computer science education. In *Informatics*, Vol. 10. MDPI, 78.
- [49] Morton Deutsch, Peter T Coleman, and Eric C Marcus. 2011. *The handbook of conflict resolution: Theory and practice*. John Wiley & Sons.
- [50] Dejana Diziol, Erin Walker, Nikol Rummel, and Kenneth R Koedinger. 2010. Using intelligent tutor technology to implement adaptive support for student collaboration. *Educational Psychology Review* 22 (2010), 89–102.
- [51] Tomoki Doi, Masaru Isonuma, and Hitomi Yanaka. 2024. Comprehensive Evaluation of Large Language Models for Topic Modeling. *arXiv:2406.00697* [cs.CL]
- [52] Margaret Drouhard, Nan-Chen Chen, Jina Suh, Rafal Kocielnik, Vanessa Pena-Araya, Keting Cen, Xiangyi Zheng, and Cecilia R Aragon. 2017. Aeonium: Visual analytics to support collaborative qualitative coding. In *2017 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 220–229.
- [53] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325* (2023).
- [54] Zackary Okun Dunivin. 2024. Scalable qualitative coding with llms: Chain-of-thought reasoning matches human performance in some hermeneutic tasks. *arXiv preprint arXiv:2401.15170* (2024).
- [55] James Eschrich and Sarah Stermann. 2024. A framework for discussing llms as tools for qualitative analysis. *arXiv preprint arXiv:2407.11198* (2024).
- [56] Mark Felton, Merce Garcia-Mila, and Sandra Gilabert. 2009. Deliberation versus dispute: The impact of argumentative discourse goals on learning and reasoning in the science classroom. *Informal Logic* 29, 4 (2009), 417–446.
- [57] Jessica L Feuston and Jed R Brubaker. 2021. Putting tools in their place: The role of time and perspective in human-AI collaboration for qualitative analysis. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.
- [58] Carli Friedman, Aleksa Owen, and Laura VanPuymbrouck. 2024. Should ChatGPT help with my research? A caution against artificial intelligence in qualitative analysis. *Qualitative Research* (2024), 14687941241297375.
- [59] Abbas Ganji, Mania Orand, and David W McDonald. 2018. Ease on Down the Code: Complex Collaborative Qualitative Coding Simplified with 'Code Wizard'. *Proceedings of the ACM on human-computer interaction* 2, CSCW (2018), 1–24.
- [60] Jie Gao, Junming Cao, ShunYi Yeo, Kenny Tsu Wei Choo, Zheng Zhang, Toby Jia-Jun Li, Shengdong Zhao, and Simon Tangi Perrault. 2023. Impact of Human-AI Interaction on User Trust and Reliance in AI-Assisted Qualitative Coding. *arXiv preprint arXiv:2309.13858* (2023).
- [61] Jie Gao, Kenny Tsu Wei Choo, Junming Cao, Roy Ka-Wei Lee, and Simon Perrault. 2023. CoAICoder: Examining the effectiveness of AI-assisted human-to-human collaboration in qualitative analysis. *ACM Transactions on Computer-Human Interaction* 31, 1 (2023), 1–38.
- [62] Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2024. CollabCoder: a lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–29.
- [63] Jie Gao, Zhiyao Shu, and Shun Yi Yeo. 2025. Using Large Language Model to Support Flexible and Structural Inductive Qualitative Analysis. *arXiv preprint arXiv:2501.00775* (2025).
- [64] Simret Araya Gebreegziabher, Zheng Zhang, Xiaohang Tang, Yihao Meng, Elena L Glassman, and Toby Jia-Jun Li. 2023. Patat: Human-ai collaborative qualitative coding with explainable interactive rule synthesis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

- [65] Ram Chandra Giri. 2024. Rapid AI: An AI Application for Thematic Analysis and Enhanced Qualitative Research. (2024).
- [66] Manuel Goyanes, Carlos Lopezosa, and Beatriz Jordá. 2024. Thematic analysis of interview data with ChatGPT: Designing and testing a reliable research protocol for qualitative research. *Center for Open Science*. <https://osf.io/8mr2f/download> (2024).
- [67] H Paul Grice. 1990. 1975 Logic and Conversation. *The Philosophy of Language* (1990).
- [68] Timothy C Guetterman, Tammy Chang, Melissa DeJonckheere, Tanmay Basu, Elizabeth Scruggs, and VG Vinod Vydiswaran. 2018. Augmenting qualitative text analysis with natural language processing: methodological study. *Journal of medical Internet research* 20, 6 (2018), e231.
- [69] Andrew Halterman and Katherine A Keith. 2024. Codebook llms: Adapting political science codebooks for llm use and adapting llms to follow codebooks. *arXiv preprint arXiv:2407.10747* (2024).
- [70] Leah Hamilton, Desha Elliott, Aaron Quick, Simone Smith, and Victoria Choplin. 2023. Exploring the use of AI in qualitative analysis: A comparative study of guaranteed income data. *International journal of qualitative methods* 22 (2023), 16094069231201504.
- [71] Abhinav Hasija and Terry L Esper. 2022. In artificial intelligence (AI) we trust: A qualitative investigation of AI technology acceptance. *Journal of Business Logistics* 43, 3 (2022), 388–412.
- [72] Adam S Hayes. 2025. “Conversing” with qualitative data: Enhancing qualitative research through large language models (LLMs). *International Journal of Qualitative Methods* 24 (2025), 16094069251322346.
- [73] Danielle Hitch. 2024. Artificial intelligence augmented qualitative analysis: the way of the future? *Qualitative Health Research* 34, 7 (2024), 595–606.
- [74] Matt-Heun Hong, Lauren A Marsh, Jessica L Feuston, Janet Ruppert, Jed R Brubaker, and Danielle Albers Szafir. 2022. Scholastic: Graphical human-AI collaboration for inductive and interpretive text analysis. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–12.
- [75] Daniel J Hruschka, Deborah Schwartz, Daphne Cobb St. John, Erin Picone-Decaro, Richard A Jenkins, and James W Carey. 2004. Reliability in coding open-ended data: Lessons learned from HIV behavioral research. *Field methods* 16, 3 (2004), 307–331.
- [76] Mohammad S Jalali and Ali Akhavan. 2024. Integrating AI language models in qualitative research: Replicating interview data analysis with ChatGPT. *System Dynamics Review* 40, 3 (2024), e1772.
- [77] Jialun Aaron Jiang, Kandrea Wade, Casey Fiesler, and Jed R Brubaker. 2021. Supporting serendipity: Opportunities and challenges for Human-AI Collaboration in qualitative analysis. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.
- [78] Adam J Johs, Denise E Agosto, and Rosina O Weber. 2022. Explainable artificial intelligence and social science: Further insights for qualitative investigation. *Applied AI letters* 3, 1 (2022), e64.
- [79] Nitish Joshi, Javier Rando, Abulhair Saparov, Najoung Kim, and He He. 2023. Personas as a way to model truthfulness in language models. *arXiv preprint arXiv:2310.18168* (2023).
- [80] Jonathan Kantor. 2024. Best practices for implementing ChatGPT, large language models, and artificial intelligence in qualitative and survey-based research. *JAAD international* 14 (2024), 22–23.
- [81] Andrew Katz, Gabriella Coloyan Fleming, and Joyce Main. 2024. Thematic Analysis with Open-Source Generative AI and Machine Learning: A New Method for Inductive Qualitative Codebook Development. *arXiv preprint arXiv:2410.03721* (2024).
- [82] Martin Kay and Martin Roscheisen. 1993. Text-translation alignment. *Computational linguistics* 19, 1 (1993), 121–142.
- [83] Awais Hameed Khan, Hiruni Kegalle, Rhea D'Silva, Ned Watt, Daniel Whelan-Shamy, Lida Ghahremanlou, and Liam Magee. 2024. Automating Thematic Analysis: How LLMs Analyse Controversial Topics. *arXiv preprint arXiv:2405.06919* (2024).
- [84] Elisabeth Kirsten, Annalina Buckmann, Abraham Mhaidli, and Steffen Becker. 2024. Decoding Complexity: Exploring Human-AI Concordance in Qualitative Coding. *arXiv preprint arXiv:2403.06607* (2024).
- [85] Jan H Klemmer, Stefan Albert Horstmann, Nikhil Patnaik, Cordelia Ludden, Cordell Burton Jr, Carson Powers, Fabio Massacci, Akond Rahman, Daniel Votipka, Heather Richter Lipford, et al. 2024. Using ai assistants in software development: A qualitative study on security practices and concerns. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. 2726–2740.
- [86] V Vien Lee, Stephanie CC van der Lubbe, Lay Hoon Goh, and Jose Maria Valderas. 2024. Harnessing ChatGPT for thematic analysis: Are we ready? *Journal of Medical Internet Research* 26 (2024), e54974.
- [87] Robert P Lennon, Robbie Fraleigh, Lauren J Van Scoy, Aparna Keshaviah, Xindi C Hu, Bethany L Snyder, Erin L Miller, William A Calo, Aleksandra E Zgierska, and Christopher Griffin. 2021. Developing and testing an automated qualitative assistant (AQUA) to support qualitative analysis. *Family medicine and community health* 9, Suppl 1 (2021), e001287.
- [88] Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024. More agents is all you need. *arXiv preprint arXiv:2402.05120* (2024).
- [89] Kevin Danis Li, Adrian M Fernandez, Rachel Schwartz, Natalie Rios, Marvin Nathaniel Carlisle, Gregory M Amend, Hiren V Patel, and Benjamin N Breyer. 2024. Comparing GPT-4 and Human Researchers in Qualitative Analysis of Healthcare Data: Qualitative Description Study. *J Med Internet Res* (2024).
- [90] Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhui Chen. 2024. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060* (2024).
- [91] Zhuofan Li, Daniel Dohan, and Corey M Abramson. 2021. Qualitative coding in the computational era: A hybrid approach to improve reliability and reduce effort for coding ethnographic interviews. *Socius* 7 (2021), 23780231211062345.
- [92] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173.

- [93] Xiner Liu, Andres Felipe Zambrano, Ryan S Baker, Amanda Barany, Jaclyn Ocumpaugh, Jiayi Zhang, Maciej Pankiewicz, Nidhi Nasir, and Zhanlan Wei. 2025. Qualitative Coding with GPT-4: Where it Works Better. *Journal of Learning Analytics* (2025), 1–17.
- [94] Xiner Liu, Jiayi Zhang, and Amanda Barany. [n. d.]. Assessing the Potential and Limits of Large Language Models in Qualitative Coding. ([n. d.]).
- [95] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. 2008. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39, 2 (2008), 539–550.
- [96] Miriam Locher. 2004. Power and Politeness in Action: Disagreements in Oral Communication. *Mouton de Gruyter* (2004).
- [97] Luca Longo. 2019. Empowering qualitative research methods in education with artificial intelligence. In *World conference on qualitative research*. Springer, 1–21.
- [98] Sariah Lopez-Fierro and Ha Nguyen. 2024. Making Human-AI Contributions Transparent in Qualitative Coding. In *Proceedings of the 17th International Conference on Computer-Supported Collaborative Learning—CSCL 2024*, pp. 3–10. International Society of the Learning Sciences.
- [99] Jingyan Lu, Ming Ming Chiu, and Nancy WaiYing Law. 2011. Collaborative argumentation and justifications: A statistical discourse analysis of online discussions. *Computers in Human Behavior* 27, 2 (2011), 946–955.
- [100] Tana M Luger, Jerry Suls, and Mark W Vander Weg. 2014. How robust is the association between smoking and depression in adults? A meta-analysis using linear mixed-effects models. *Addictive behaviors* 39, 10 (2014), 1418–1429.
- [101] Kathleen M MacQueen, Eleanor McLellan, Kelly Kay, and Bobby Milstein. 1998. Codebook development for team-based qualitative analysis. *Cam Journal* 10, 2 (1998), 31–36.
- [102] Megh Marathe and Kentaro Toyama. 2018. Semi-automated coding for qualitative research: A user-centered inquiry and initial prototypes. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.
- [103] David T Marshall and David B Naff. 2024. The ethics of using artificial intelligence in qualitative research. *Journal of Empirical Research on Human Research Ethics* 19, 3 (2024), 92–102.
- [104] Walter S Mathis, Sophia Zhao, Nicholas Pratt, Jeremy Weleff, and Stefano De Paoli. 2024. Inductive thematic analysis of healthcare qualitative interviews using open-source large language models: How does it compare to traditional methods? *Computer Methods and Programs in Biomedicine* 255 (2024), 108356.
- [105] Aida Mehrad, Mohammad Hossein Tahriri Zangeneh, Meriem Bouzedif, Neema George Rweramila, and Hamza Boutament. 2024. Qualitative Methods in Statistics and Methodology: Artificial Intelligence (AI). *Journal of Advanced Studies in Social Sciences* 2, 2 (2024).
- [106] Han Meng, Yitian Yang, Yunan Li, Jungup Lee, and Yi-Chieh Lee. 2024. Exploring the Potential of Human-LLM Synergy in Advancing Qualitative Analysis: A Case Study on Mental-Illness Stigma. *arXiv preprint arXiv:2405.05758* (2024).
- [107] Blaž Mesec. 2023. The language model of artificial intelligence chatGPT-a tool of qualitative analysis of texts. *Authorea Preprints* (2023).
- [108] David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G Altman. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Bmj* 339 (2009).
- [109] David L Morgan. 2023. Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. *International journal of qualitative methods* 22 (2023), 16094069231211248.
- [110] Dimitri Mortelmans. 2024. NVivo and AI:(Semi)-Automatic Coding. In *Doing Qualitative Data Analysis with NVivo*. Springer, 229–250.
- [111] Mohamed Nejjar, Luca Zacharias, Fabian Stiehle, and Ingo Weber. 2025. Llm for science: Usage for code generation and data analysis. *Journal of Software: Evolution and Process* 37, 1 (2025), e2723.
- [112] Caroline A Nelson, Lourdes Maria Pérez-Chada, Andrew Creadore, Sara Jiayang Li, Kelly Lo, Priya Manjaly, Ashley Bahareh Pournamdari, Elizabeth Tkachenko, John S Barbieri, Justin M Ko, et al. 2020. Patient perspectives on the use of artificial intelligence for skin cancer screening: a qualitative study. *JAMA dermatology* 156, 5 (2020), 501–512.
- [113] Laura K Nelson, Derek Burk, Marcel Knudsen, and Leslie McCall. 2021. The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research* 50, 1 (2021), 202–237.
- [114] Kien Nguyen-Trung. 2024. ChatGPT in Thematic Analysis: Can AI become a research assistant in qualitative research. *OSF Preprint, researchgate.net* (2024).
- [115] Kien Nguyen-Trung and Ngoc Lan Nguyen. 2025. Narrative-Integrated Thematic Analysis (NITA): AI-Supported Theme Generation Without Coding. (2025).
- [116] Matthew Nyaaba, Min SungEun, Mary Abiswin Apam, Kwame Owuahene Acheampong, and Emmanuel Dwamena. 2025. Optimizing Generative AI’s Accuracy and Transparency in Inductive Thematic Analysis: A Human-AI Comparison. *arXiv preprint arXiv:2503.16485* (2025).
- [117] Ryan M Omizo. 2024. Automating research in business and technical communication: Large Language models as qualitative coders. *Journal of Business and Technical Communication* 38, 3 (2024), 242–265.
- [118] Cassandra Overney. 2023. *SenseMate: An AI-Based Platform to Support Qualitative Coding*. Master’s thesis. Massachusetts Institute of Technology.
- [119] Cassandra Overney, Belén Saldías, Dimitra Dimitrakopoulou, and Deb Roy. 2024. Sensemate: An accessible and beginner-friendly human-ai platform for qualitative data analysis. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 922–939.
- [120] Dustin Palea, Giridhar Vadhal, and David T Lee. 2024. Annota: Peer-based AI Hints Towards Learning Qualitative Coding at Scale. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 455–470.
- [121] Angelina Parfenova, Alexander Denzler, and Jörgen Pfeffer. 2024. Automating qualitative data analysis with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*. 83–91.
- [122] Walter C Parker. 2006. Public discourses in schools: Purposes, problems, possibilities. *Educational Researcher* 35, 8 (2006), 11–18.

- [123] Frédéric Pattyn. 2024. The value of generative AI for qualitative research: A pilot study. *Journal of Data Science and Intelligent Systems* (2024).
- [124] Trena M Paulus and Vittorio Marone. 2024. "In minutes instead of weeks": Discursive constructions of generative AI and qualitative data analysis. *Qualitative Inquiry* (2024), 10778004241250065.
- [125] Mike Perkins and Jasper Roe. 2024. Generative AI tools in academic research: Applications and implications for qualitative and quantitative research methodologies. *arXiv preprint arXiv:2408.06872* (2024).
- [126] Mike Perkins and Jasper Roe. 2024. The use of Generative AI in qualitative analysis: Inductive thematic analysis with ChatGPT. *Journal of Applied Learning and Teaching* 7, 1 (2024).
- [127] Spinoso-Di Piano et al. 2024. Qualitative Code Suggestion: A Human-Centric Approach to Qualitative Coding. (2024).
- [128] Lois Player, Ryan Hughes, Kaloyan Mitev, Lorraine Whitmarsh, Christina Demski, Nicholas Nash, Trisevgeni Papakonstantinou, and Mark Wilson. 2024. The Use of Large Language Models for Qualitative Research: DECOTA. (2024).
- [129] Lotte van Poppel. 2020. The study of metaphor in argumentation theory. In *Argumentation Through Languages and Cultures*. Springer, 177–208.
- [130] Maximo R Prescott, Samantha Yeager, Lillian Ham, Carlos D Rivera Saldana, Vanessa Serrano, Joey Narez, Dafna Paltin, Jorge Delgado, David J Moore, and Jessica Montoya. 2024. Comparing the efficacy and efficiency of human and generative AI: Qualitative thematic analyses. *JMIR AI* 3 (2024), e54482.
- [131] Tingrui Qiao, Caroline Walker, Chris W Cunningham, and Yun Sing Koh. [n. d.]. Thematic-LM: a LLM-based Multi-agent System for Large-scale Thematic Analysis. In *THE WEB CONFERENCE 2025*.
- [132] Varun Nagaraj Rao, Eesha Agarwal, Samantha Dalal, Dan Calacci, and Andrés Monroy-Hernández. 2024. QuaLLM: An LLM-based framework to extract quantitative insights from online forums. *arXiv preprint arXiv:2405.05345* (2024).
- [133] Zeeshan Rasheed, Muhammad Waseem, Aakash Ahmad, Kai-Kristian Kemell, Wang Xiaofeng, Anh Nguyen Duc, and Pekka Abrahamsson. 2024. Can large language models serve as data analysts? A multi-agent assisted approach for qualitative data analysis. *arXiv preprint arXiv:2402.01386* (2024).
- [134] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).
- [135] Tim Rietz. 2021. *Designing AI-based systems for qualitative data collection and analysis*. Ph. D. Dissertation. Dissertation, Karlsruhe, Karlsruhe Institut für Technologie (KIT), 2021.
- [136] Tim Rietz and Alexander Maedche. 2021. Cody: An AI-based system to semi-automate coding for qualitative research. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.
- [137] Mona J Ritchie, Karen L Drummond, Brandy N Smith, Jennifer L Sullivan, and Sara J Landes. 2022. Development of a qualitative data analysis codebook informed by the i-PARIHS framework. *Implementation science communications* 3, 1 (2022), 98.
- [138] Valentin Ritschl, Lars Grespan, and Datum Unterschrift Ort. [n. d.]. Assessing the Effectiveness of Large Language Models in Qualitative Content Analysis of Interview Data. ([n. d.]).
- [139] John Roberts, Max Baker, and Jane Andrew. 2024. Artificial intelligence and qualitative research: The promise and perils of large language model (LLM) 'assistance'. *Critical Perspectives on Accounting* 99 (2024), 102722.
- [140] Soroush Sabbaghan. 2024. Exploring the synergy of human and AI-driven approaches in thematic analysis for qualitative educational research. *Journal of Applied Learning and Teaching* 7, 2 (2024).
- [141] Katie Rose M Sanfilippo, Neta Spiro, Miguel Molina-Solana, and Alexandra Lamont. 2020. Do the shuffle: Exploring reasons for music listening through shuffled play. *PloS one* 15, 2 (2020), e0228457.
- [142] Frank A Scannapieco. 1997. Formal debate: an active learning strategy. *Journal of Dental Education* 61 (1997), 955–961.
- [143] Martin Schmettow. 2015. Tutorial: Modern regression techniques for HCI researchers. In *Human-Computer Interaction—INTERACT 2015: 15th IFIP TC 13 International Conference, Bamberg, Germany, September 14–18, 2015, Proceedings, Part IV* 15. Springer, 651–654.
- [144] Tim Schopf, Daniel Braun, and Florian Matthes. 2022. Evaluating unsupervised text classification: zero-shot and similarity-based approaches. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*. 6–15.
- [145] Baruch B Schwarz, Naomi Prusak, Osama Swidan, Adva Livny, Kobi Gal, and Avi Segal. 2018. Orchestrating the emergence of conceptual learning: A case study in a geometry class. *International Journal of Computer-Supported Collaborative Learning* 13 (2018), 189–211.
- [146] Nicole Schwitter. 2025. Using large language models for preprocessing and information extraction from unstructured text: A proof-of-concept application in the social sciences. *Methodological Innovations* (2025), 20597991251313876.
- [147] Mert Şen, ŞEVVAL Şen, and TUĞRUL Şahin. 2023. A new era for data analysis in qualitative research: ChatGPT! *Shanlax International Journal of Education* 11, 1 (2023).
- [148] Burr Settles. 2009. Active learning literature survey. (2009).
- [149] Leo A Siiman, Meeli Rannastu-Avalos, Johanna Pöysä-Tarhonen, Päivi Häkkinen, and Margus Pedaste. 2023. Opportunities and challenges for AI-assisted qualitative data analysis: An example from collaborative problem-solving discourse data. In *International Conference on Innovative Technologies and Learning*. Springer, 87–96.
- [150] Ravi Sinha, Idris Solola, Ha Nguyen, Hillary Swanson, and LuEtaMae Lawrence. 2024. The role of generative AI in qualitative research: GPT-4's contributions to a grounded theory analysis. In *Proceedings of the 2024 Symposium on Learning, Design and Technology*. 17–25.
- [151] Evgeny Smirnov. 2025. Enhancing qualitative research in psychology with large language models: a methodological exploration and examples of simulations. *Qualitative Research in Psychology* 22, 2 (2025), 482–512.

- [152] Andries Petrus Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D Barrett, and Arnu Pretorius. [n. d.]. Should we be going MAD? A Look at Multi-Agent Debate Strategies for LLMs. In *Forty-first International Conference on Machine Learning*.
- [153] Cesare Spinoso-Di Piano. 2023. *Qualitative Code Suggestion: A Human-Centric Approach to Qualitative Coding*. McGill University (Canada).
- [154] Vighnesh Subramaniam, Antonio Torralba, and Shuang Li. 2024. DebateGPT: Fine-tuning Large Language Models with Multi-agent Debate Supervision. (2024).
- [155] Harrison Boyd Summers, Forest Livings Whan, and Thomas Andrew Rousse. 1950. How to debate: A textbook for beginners. (*No Title*) (1950).
- [156] Robert H Tai, Lillian R Bentley, Xin Xia, Jason M Sitt, Sarah C Fankhauser, Ana M Chicas-Mosier, and Barnas G Monteith. 2024. An examination of the use of large language models to aid analysis of textual data. *International Journal of Qualitative Methods* 23 (2024), 16094069241231168.
- [157] Jingyi Tang, Ting Wang, and Sicheng Zhu. 2023. Assessing Word Difficulty: A Mapping Method. In *2023 International Conference on Blockchain Technology and Applications (ICBTA)*. IEEE, 1–4.
- [158] Stephanie Teasley, Frank Fischer, Pierre Dillenbourg, Manu Kapur, Michelene Chi, Armin Weinberger, and Karsten Stegmann. 2008. Cognitive convergence in collaborative learning. (2008).
- [159] Nga Than, Leanne Fan, Tina Law, Laura K Nelson, and Leslie McCall. 2024. Updating “The Future of Coding”: Qualitative Coding with Generative Large Language Models. (2024).
- [160] Kenneth W Thomas. 2008. Thomas-kilmann conflict mode. *TKI Profile and Interpretive Report* 1, 11 (2008).
- [161] Dean Tjosvold, Alfred SH Wong, and Nancy Yi Feng Chen. 2019. Managing conflict for effective leadership and organizations. In *Oxford research encyclopedia of business and management*.
- [162] Petter Törnberg. 2023. How to use llms for text analysis. *arXiv preprint arXiv:2307.13106* (2023).
- [163] Paul Tschisgale, Peter Wulff, and Marcus Kubsch. 2023. Integrating artificial intelligence-based methods into qualitative research in physics education research: A case for computational grounded theory. *Physical Review Physics Education Research* 19, 2 (2023), 020123.
- [164] Aleksei Turobov, Diane Coyle, and Verity Harding. 2024. Using ChatGPT for thematic analysis. *arXiv preprint arXiv:2405.08828* (2024).
- [165] Jonas Wachinger, Kate Bärnighausen, Louis N Schäfer, Kerry Scott, and Shannon A McMahon. 2024. Prompts, pearls, imperfections: comparing ChatGPT and a human researcher in qualitative data analysis. *Qualitative Health Research* (2024), 10497323241244669.
- [166] Henning Wachsmuth and Milad Alshomary. 2022. “Mama Always Had a Way of Explaining Things So I Could Understand”: A Dialogue Corpus for Learning to Construct Explanations. *arXiv preprint arXiv:2209.02508* (2022).
- [167] DN Walton. 2008. *Argumentation schemes*. Cambridge University Press.
- [168] Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024. Rethinking the Bounds of LLM Reasoning: Are Multi-Agent Discussions the Key? *arXiv preprint arXiv:2402.18272* (2024).
- [169] Kathryn Wheeler. 2025. How to Use Generative AI to Assist the Analysis of Qualitative Data. *Sage Research Methods How to Guides* (2025).
- [170] Alyssa Friend Wise and Ming Ming Chiu. 2011. Analyzing temporal patterns of knowledge construction in a role-based online discussion. *International Journal of Computer-Supported Collaborative Learning* 6 (2011), 445–470.
- [171] Ziang Xiao, Xingdi Yuan, Q Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In *Companion proceedings of the 28th international conference on intelligent user interfaces*. 75–78.
- [172] Huimin Xu, Seungjun Yi, Terence Lim, Jiawei Xu, Andrew Well, Carlos Mery, Aidong Zhang, Yuji Zhang, Heng Ji, Keshav Pingali, et al. 2025. TAMA: A Human-AI Collaborative Thematic Analysis Framework Using Multi-Agent LLMs for Clinical Interviews. *arXiv preprint arXiv:2503.20666* (2025).
- [173] Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Retrieval meets long context large language models. *arXiv preprint arXiv:2310.03025* (2023).
- [174] Lixiang Yan, Vanessa Echeverria, Gloria Milena Fernandez-Nieto, Yueqiao Jin, Zachari Swiecki, Linxuan Zhao, Dragan Gašević, and Roberto Martinez-Maldonado. 2024. Human-AI collaboration in thematic analysis using ChatGPT: A user study and design recommendations. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.
- [175] Yang Yang and Liran Ma. 2025. Artificial intelligence in qualitative analysis: a practical guide and reflections based on results from using GPT to analyze interview data in a substance use program. *Quality & Quantity* (2025), 1–24.
- [176] Donggeun Yoo and In So Kweon. 2019. Learning loss for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 93–102.
- [177] Himanshu Zade, Margaret Drouhard, Bonnie Chinh, Lu Gan, and Cecilia Aragon. 2018. Conceptualizing disagreement in qualitative coding. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–11.
- [178] Stepan Zakharov, Omri Hadar, Tovit Hakak, Dina Grossman, Yifat Ben-David Kolikant, and Oren Tsur. 2021. Discourse parsing for contentious, non-convergent online discussions. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 15. 853–864.
- [179] Andres Felipe Zambrano, Xiner Liu, Amanda Barany, Ryan S Baker, Juhan Kim, and Nidhi Nasir. 2023. From nCoder to ChatGPT: From automated coding to refining human coding. In *International conference on quantitative ethnography*. Springer, 470–485.
- [180] Yuheng Zha, Ruichen Li, and Zhiting Hu. 2023. Text alignment is an efficient unified model for massive nlp tasks. *Advances in Neural Information Processing Systems* 36 (2023), 77942–77968.
- [181] He Zhang, Chuha Wu, Jingyi Xie, ChanMin Kim, and John M Carroll. 2023. QualiGPT: GPT as an easy-to-use tool for qualitative coding. *arXiv preprint arXiv:2310.07061* (2023).

- [182] He Zhang, Chuhao Wu, Jingyi Xie, Yao Lyu, Jie Cai, and John M Carroll. 2023. Redefining qualitative analysis in the AI era: Utilizing ChatGPT for efficient thematic analysis. *arXiv preprint arXiv:2309.10771* (2023).
- [183] He Zhang, Chuhao Wu, Jingyi Xie, Fiona Rubino, Sydney Graver, ChanMin Kim, John M Carroll, and Jie Cai. 2024. When Qualitative Research Meets Large Language Model: Exploring the Potential of QualiGPT as a Tool for Qualitative Coding. *arXiv preprint arXiv:2407.14925* (2024).
- [184] Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2023. Exploring collaboration mechanisms for llm agents: A social psychology view. *arXiv preprint arXiv:2310.02124* (2023).
- [185] Fengxiang Zhao, Fan Yu, and Yi Shang. 2024. A New Method Supporting Qualitative Data Analysis Through Prompt Generation for Inductive Coding. In *2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, 164–169.
- [186] Karolina Zięba-Kulawik. 2025. Generative AI: New Framework of Using Large Language Models for Analysing Descriptive Qualitative Data. (2025).
- [187] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics* 50, 1 (2024), 237–291.

A Comparison of Discursive Moves Between LLM Coders and Humans

Table 7. Comparison of Discursive Moves Between LLM Coders and Humans. This table shows the discursive moves used by humans in the literature, which did not appear in the discussions of LLM coders.

Category of Discursive Moves	Discursive Move of Humans in Literature
Moderating	Moderation [9, 67, 170]
Illustrating	Example and Illustration [129, 167] Metaphor and Analogy [129, 167]
Expansive	Idea Expansion [129, 167, 178]
Personal Story	Personal Narrative [129, 167, 178]
Questioning	Clarification Request [9, 67] Questioning Evidence [129, 167, 178] Answer [129, 167, 178] Critical Question [129, 167, 178]
Aggressive	Confrontational Expression [129, 167, 178] Dismissive Humor [129, 167, 178] Sarcasm [129, 167, 178] Reframing Critique [129, 167, 178]
Disruptive	Relevance Critique [129, 167, 178] Nitpicking [129, 167, 178] Complaint [129, 167, 178] Disruptive Behavior [129, 167, 178] Topic Derailment [129, 167, 178] Non-reasoned Disagreement [129, 167, 178]