Research Methods in Applied Linguistics 4 (2025) 100210

Contents lists available at ScienceDirect

Research Methods in Applied Linguistics

journal homepage: www.elsevier.com/locate/rmal

Large language models fall short in classifying learners' open-ended responses

Atsushi Mizumoto [a,1], Mark Feng Teng [b,2,*]

[a] Kansai University, Japan
[b] Macao Polytechnic University, Macau SAR, PR China

ARTICLE INFO

Keywords:
Research methods
Generative AI
Large language models (LLM)
Qualitative analysis
Coding and classification

ABSTRACT

Generative Artificial Intelligence (GenAI), based on large language models (LLMs), excels in various language comprehension tasks and is increasingly utilized in applied linguistics research. This study examines the accuracy and methodological implications of using LLMs to classify open-ended responses from learners. We surveyed 143 Japanese university students studying English as a foreign language (EFL) about their essay-writing process. Two human evaluators independently classified the students' responses based on self-regulated learning processes: planning, monitoring, and evaluation. At the same time, several LLMs performed the same classification task, and their results were compared with those of the human evaluators using Cohen's kappa coefficient. We established $\kappa \geq 0.8$ as the threshold for strong agreement based on rigorous methodological standards. Our findings revealed that even the best-performing model (DeepSeek-V3) achieved only moderate agreement ($\kappa = 0.68$), while other models demonstrated fair-to-moderate agreement ($\kappa = 0.37$–0.61). Surprisingly, open-source models outperformed several commercial counterparts. These results highlight the necessity of expert oversight when integrating GenAI as a support tool in qualitative data analysis. The paper concludes by discussing the methodological implications for using LLMs in qualitative research and proposing specific prompt engineering strategies to enhance their reliability in applied linguistics.

Introduction

Large Language Models (LLMs) have rapidly emerged as powerful tools capable of generating and analyzing human-like text, prompting researchers across disciplines to explore their potential applications. In applied linguistics, where researchers regularly analyze large volumes of textual data—including student responses, interview transcripts, and forum posts—LLMs offer particularly

* Corresponding author.
E-mail addresses: mizumoto@kansai-u.ac.jp (A. Mizumoto), markteng@mpu.edu.mo (M.F. Teng).

[1] Atsushi Mizumoto holds a Ph.D. in Foreign Language Education and is a professor in the Faculty of Foreign Language Studies and the Graduate School of Foreign Language Education and Research at Kansai University, Japan. His current research interests include corpus use for pedagogical purposes, learning strategies, language testing, and research methodology.

[2] Mark Feng Teng, Ph.D., is Associate Professor at Macao Polytechnic University. His-research portfolio mainly focuses on L2 vocabulary acquisition and L2 writing. His-publications have appeared in international journals, including Applied Linguistics, TESOL Quarterly, Language Teaching Research, System, Applied Linguistics Review, Computer Assisted Language Learning, Computers & Education, Foreign Language Annals, and IRAL, among others. His-recent monographs were published by Routledge, Springer, and Bloomsbury. He serves as editor-in-chief for International Journal of TESOL Studies (IJTS) and Digital Applied Linguistics (DAL).

https://doi.org/10.1016/j.rmal.2025.100210
Received 16 February 2025; Received in revised form 11 April 2025; Accepted 11 April 2025
Available online 21 April 2025

promising possibilities for enhancing research efficiency.

Recent studies have demonstrated LLMs' proficiency in various language-related tasks (Barros et al., 2024), suggesting their potential to transform methodological approaches to linguistic data analysis. Researchers are increasingly exploring the applications of LLMs for traditionally manual research tasks, such as coding data, identifying themes, and categorizing responses (e.g., Kim & Lu, 2024; Yu, 2025). This growing interest stems from LLMs' ability to process and analyze text at scale, potentially reducing the time and resources required for qualitative data analysis.

However, the integration of LLMs into qualitative research methods presents significant challenges that require careful consideration. While these models can automate labor-intensive processes and produce consistent outputs, they may struggle to interpret contextual meanings and identify subtle relationships between concepts (Lee et al., 2024). This limitation is particularly relevant in applied linguistics, where analysis often requires a deep understanding of context-dependent language use and learner data. Additionally, ensuring that an LLM's analysis aligns with theoretical frameworks and captures real-world meanings remains a substantial challenge. Thus, it is evident that intrinsic limitations exist in the capabilities of LLMs, necessitating careful consideration in their application.

In response to these inherent limitations, recent qualitative research methodology emphasizes a collaborative model that integrates human expert judgment with AI-driven efficiency. In practice, this means leveraging LLMs for their speed and consistency while retaining human researchers' interpretive oversight (Barros et al., 2024). Such a partnership ensures that AI-assisted analysis benefits from the strengths of both parties: the model can rapidly process large volumes of text, and the human expert can contextualize and validate the results. This perspective on human–AI collaboration provides a theoretical backdrop for our study, underscoring the notion that LLMs are best used as supportive tools within qualitative research workflows rather than independent judges of data.

Given these challenges, empirical research examining the effectiveness of LLMs in qualitative research methods within applied linguistics is crucial. However, such research remains surprisingly scarce. This gap is particularly notable in one key area: the use of GenAI to support the analysis and interpretation of open-ended responses and interviews. While LLMs show promise for partially automating aspects of qualitative analysis, their accuracy and reliability in tasks such as response classification and thematic analysis remain largely untested within the field.

The present study makes a significant contribution to research methods in applied linguistics qualitative research by examining the performance of LLMs in classifying open-ended data. Specifically, it investigates their ability to categorize student responses based on established theoretical frameworks, a task traditionally performed manually by researchers. By directly comparing LLM-generated classifications with those of human researchers, this study provides valuable empirical evidence on the potential and limitations of LLMs as tools for qualitative data analysis. The findings aim to advance methodological approaches in applied linguistics, offering insights into how LLMs can complement or enhance traditional qualitative research practices.

Literature review

A brief history of GenAI in applied linguistics

Since OpenAI's release of ChatGPT on November 30, 2022, the field of applied linguistics has experienced a surge in research examining the applications and implications of generative AI (GenAI). Although GenAI tools were available prior to ChatGPT's launch, their capabilities were relatively limited. ChatGPT's remarkable ability to understand and generate language effectively across diverse contexts revolutionized the landscape, spurring the development of other AI applications such as Google Bard (now Gemini) and Perplexity.ai, and compelling educators and educational institutions to take notice. The release of the DeepSeek-R1 model in January 2025 intensified competition among open AI models, prompting OpenAI to respond by launching more advanced models, further reshaping the dynamics of the AI ecosystem.

Beginning with Kohnke et al. (2023), a growing number of studies have emerged in applied linguistics and English language teaching and learning (e.g., Moorhouse et al., 2024). However, concerns surrounding GenAI have persisted, particularly regarding issues such as privacy, accuracy and potential bias, despite ongoing efforts by AI developers to address these challenges. In the education domain, concerns were initially centered on the potential for student cheating, which led to changes in assessment practices, including a return to in-class paper-and-pencil tests and an emphasis on oral assessments (Kohnke et al., 2023), for which Todd (2025) argued that GenAI is a disruptive technology. A significant shift soon followed, moving from perceiving GenAI as an obstacle to recognizing its potential as a facilitator of learning. Bans on GenAI use in educational settings were largely rescinded, and the discourse shifted toward exploring ways for teachers and students to leverage GenAI for educational benefit. Research began highlighting its potential to enhance student motivation (Huang & Mizutomo, 2024), engagement (Huang & Teng, 2025), and metacognitive awareness (Teng, 2025). The proliferation of research on ChatGPT grew dramatically, particularly in 2024, with a notable focus on its applications in second language education. This surge in interest has already resulted in multiple systematic reviews (e.g., Lo et al., 2024; Teng, 2024; Yang & Li, 2024), many of which examine its role in English as a Foreign Language (EFL) contexts, underscoring its transformative potential in language education.

A systematic review by Lo et al. (2024) examined the effects on learners from behavioral, emotional, and cognitive perspectives, with particularly notable findings regarding anxiety reduction. One prominently researched area has been the use of ChatGPT for writing feedback, with studies indicating its viability as a support tool for instructors (Allen & Mizumoto, 2024; Steiss et al., 2024). These studies suggest that GenAI tools like ChatGPT can bring significant benefits to language learning and instruction, provided there is careful attention to potential issues such as over-reliance, privacy concerns, and risks of plagiarism and passive learning without critical thinking (Crosthwaite & Baisa, 2023). As assessment practices continue to evolve in response to these challenges, the

integration of GenAI into assessment design has become increasingly relevant. GenAI offers opportunities to develop more personalized, adaptive, and formative assessment methods.

Potentials and limitations of LLMs in qualitative analysis

Beyond educational applications, the L2 field has seen GenAI integration in various research contexts, including automated essay scoring and accuracy assessment (e.g., Mizumoto & Eguchi, 2023; Mizumoto et al., 2024; Uchida & Negishi, 2025) and materials and test development (Aryadoust et al., 2024; Lin, 2023; Shin, 2023; Xin, 2024). The high linguistic capabilities and performance of ChatGPT and similar GenAI tools have positioned them as increasingly valuable support tools for researchers and practitioners in the field.

These broad applications of GenAI in applied linguistics raise the question of how effectively LLMs can perform more specialized qualitative analysis tasks, such as detailed linguistic annotation. However, in applied linguistics research specifically, empirical studies examining LLMs' ability to perform detailed linguistic annotation remain scarce. Two notable exceptions are Kim and Lu (2024) and Yu (2025), who investigated the use of ChatGPT for move analysis in academic texts. Kim and Lu (2024) evaluated ChatGPT's performance in annotating rhetorical moves and steps in research article introductions, finding that fine-tuning significantly improved the model's accuracy (92.3 % for moves and 80.2 % for steps). Similarly, Yu (2025) explored ChatGPT-4's capability to analyze rhetorical moves in corporate social responsibility reports, achieving 87.14 % accuracy with a fine-tuned model but noting the necessity of human verification for inconsistent cases. Although these studies demonstrate promising potential for AI-assisted annotation in applied linguistics, they also emphasize that LLMs should complement rather than replace human expertise in detailed linguistic analysis.

This growing interest in AI-assisted qualitative analysis is reflected in the integration of GenAI capabilities into major Computer Assisted Qualitative Data Analysis Software (CAQDAS) tools such as ATLAS.ti, MAXQDA, and NVivo. However, this integration raises critical concerns regarding the reliability and validity of such analysis. Despite their potential, LLMs face significant challenges in qualitative research, particularly in contexts where context and nuance are crucial. These limitations stem from three primary factors: dependency on training data, a tendency to overgeneralize, and a lack of deep contextual understanding.

Several studies have systematically evaluated LLMs' capabilities in qualitative data analysis, revealing consistent patterns in their strengths and limitations. Morgan (2023) compared ChatGPT's performance with traditional manual coding across focus group datasets, finding that while it effectively identified concrete, descriptive themes, it struggled with capturing abstract, interpretative themes that human coders readily recognized. Wachinger et al. (2024) confirmed these patterns in interview data analysis, noting similar limitations in theoretical engagement. These difficulties stem from two key factors: LLMs' reliance on training data—potentially misinterpreting or overlooking underrepresented perspectives (Schroeder et al., 2024)—and their tendency to overgeneralize based on surface-level patterns, associating certain phrase structures with specific categories regardless of content (Barros et al., 2024; Nielbo et al., 2024). Despite these limitations, some practical applications show promise. Fuller et al. (2024) reported significant time savings when using AI to analyze open-ended student evaluations, while noting the need for human oversight. Collectively, these studies emphasize that LLMs should serve as assistive tools rather than replacements for human analysis, especially when interpretative depth is required.

Overall, while LLMs have shown promising potential, their role in qualitative research in applied linguistics remains underexplored. While studies such as Kim and Lu (2024) and Yu (2025) have demonstrated the capabilities of LLMs in performing tasks like rhetorical move analysis with impressive accuracy, they also highlight the need for human oversight in addressing inconsistencies and ensuring nuanced interpretation. Fundamentally, LLMs lack true contextual understanding—their "knowledge" is based on statistical correlations rather than lived experience or interpretive insight, making them prone to missing deeper contextual elements that human analysts intuitively grasp.

To better understand LLMs' unique position in qualitative research, it is instructive to compare them with traditional Computer Assisted Qualitative Data Analysis Software (CAQDAS). When compared with established tools like NVivo, ATLAS.ti, or MAXQDA, significant methodological differences emerge. Traditional CAQDAS tools function as extensions of the researcher's analytical process, keeping the human at the core of interpretation. In contrast, LLMs can act as semi-autonomous analysts, potentially reducing researchers' deep engagement with data. While LLMs offer practical advantages—intuitive interfaces, rapid processing, and integrated features like translation—they may not align with the epistemological values of qualitative research that emphasize multiple interpretations and researcher reflexivity.

Current best practices suggest using LLMs to augment rather than replace human analysis—for example, employing them for preliminary coding or synthesizing large datasets, followed by human verification. This hybrid approach leverages LLMs' efficiency while ensuring analyses remain trustworthy, nuanced, and contextually grounded. This growing body of evidence points to an urgent need for further exploration of LLMs in qualitative applied linguistics research, particularly regarding their methodological implications and appropriate integration within established qualitative frameworks.

Theoretical framework: principles of qualitative coding

In addition to content-specific theories, our study is informed by fundamental principles of qualitative coding from research methodology. In qualitative research, coding open-ended data involves clear category definitions, careful training, and iterative consensus-building among human coders to ensure reliability and validity of interpretations. Researchers are expected to consider context and nuance when assigning codes, often referring back to a coding manual or theoretical model that guides the analysis. In our case, the three processes of the metacognitive aspect of self-regulated learning theory–planning, monitoring, and evaluation–were

defined a priori (drawn from the literature such as Teng, 2020; Teng et al., 2022; Teng & Zhang, 2016) and serve as a structured coding scheme. Grounding our approach in these coding principles means that we evaluate LLM performance by asking: can the model apply the codes consistently and accurately in line with human coders, given the same definitions and examples? By situating the study within the established practice of qualitative coding, we align our work with qualitative research standards and can better interpret where the LLM's classifications converge with or diverge from expected human coding behavior.

The present study

As reviewed above, the use of GenAI in qualitative research has been increasing. However, despite its growing presence, the accuracy and reliability of GenAI in qualitative research in applied linguistics remain largely unexamined. While some studies have demonstrated its potential for detecting linguistic patterns (e.g., Kim & Lu, 2024; Yu, 2025), the extent to which large language models (LLMs) can classify open-ended responses with precision and consistency is still unclear.

To address this gap, this study focuses on the question: "How well do LLMs classify open-ended data?" Specifically, it examines whether GenAI can accurately assign pre-established categories to qualitative data, a task that involves structured classification rather than open-ended thematic generation. By evaluating its performance in this specific use case, this study aims to contribute to the ongoing investigation of the role of AI in qualitative research and its methodological implications.

Methods

Participants

This study examined a total of 143 first-year university students enrolled at a private university located in the western region of Japan. The cohort consisted of 34 male and 109 female students, all of whom belonged to the faculty of foreign language studies with a specialization in English. As part of their academic curriculum, these students were required to take a mandatory course designed to strengthen their grammar and vocabulary skills. The primary goal of this course was to equip them with the necessary language proficiency to successfully participate in a one-year study abroad program scheduled for the following academic year.

To collect data for this study, a convenience sampling method was employed, as writing assignments were an integral component of the course. The English proficiency levels of the participants were evaluated using the TOEFL ITP test, with the majority of students scoring within the B1 to B2 range according to the Common European Framework of Reference (CEFR). Additionally, all participants had undergone at least eight years of compulsory English education within the Japanese school system, which includes primary and secondary education.

Procedure

Participants were asked to respond in an open-ended manner to the question: "*You have been assigned an essay-writing task in English. How will you approach this task? Please explain in one sentence.*" Their responses were collected through the university's Learning Management System (LMS). The decision to restrict responses to a single sentence was made to streamline the subsequent categorization process, ensuring consistency and clarity in data analysis. Prior to data collection, participants were informed that their responses would not affect their grades, and written informed consent was obtained.

The gathered text data were then carefully reviewed and categorized by two researchers holding Ph.D. in Applied Linguistics. Both researchers had prior experience teaching similar learner populations at Japanese universities and in conducting qualitative research. Grounded in established theoretical frameworks, they classified the responses into three distinct self-regulated learning processes—planning, monitoring, and evaluation. These three categories were selected based on Zimmerman's (2000) widely accepted cyclical model of SRL, which has been extensively validated across educational contexts, including second language acquisition. To operationalize these categories, we used specific item definitions from validated self-regulated learning questionnaires developed in previous studies (e.g., Teng et al., 2022; Teng & Zhang, 2016). These established instruments provided clear criteria for distinguishing between planning behaviors (setting goals and organizing ideas before writing), monitoring activities (adjusting and checking one's writing during the composition process), and evaluation processes (reflecting on and reviewing completed writing). To ensure reliability, each researcher independently assigned responses to categories. We acknowledge that many responses contained elements that could potentially span multiple SRL processes—a common challenge in qualitative coding of complex cognitive behaviors. In such cases, we focused on identifying the dominant process in each response by carefully referring to the specific questionnaire items from the validated instruments. This allowed us to determine which SRL phase was most prominently reflected in the student's description of their writing approach. In cases where their classifications differed, the first author of this paper joined and facilitated discussions, referring to the specific questionnaire items to reach a consensus on the final categorization.

Table 1 presents the number of responses assigned to each category along with illustrative examples. Because the primary objective of this study was to examine the feasibility of using LLMs for categorizing responses based on pre-defined classifications, each response was assigned to only one category. In practice, some responses encompassed elements of multiple processes (e.g., planning and monitoring), making it challenging to assign them to a single, clear-cut category. However, in such instances, the researchers discussed and jointly decided on the most appropriate classification for each response.

As part of this categorization process, the original Japanese responses provided by participants were translated into English by the two researchers responsible for classification. This translation step was deemed necessary because prior research indicates that LLMs

generally exhibit stronger comprehension and processing capabilities in English than in Japanese. To maintain consistency and minimize potential discrepancies, all responses were converted into English before being processed by the LLM for classification. Furthermore, a back-translation procedure was conducted to verify that the meaning and intent of the original Japanese responses remained intact in the English versions. During this process, the researchers also performed a double-check to confirm that the assigned categories remained unchanged between the original and translated versions, ensuring alignment in the data interpretation.

Analysis with LLMs

To categorize responses using large language models (LLMs), we employed seven models in Table 2: three from the GPT series (GPT-4o, GPT-o1, GPT-o3mini), Llama3.3–70B, Gemini2.0-Flash, Claude3.5-Sonnet, and DeepSeek-V3. Developed by different entities, these models are considered among the best-performing as of February 2025. This study aimed to evaluate the feasibility of using LLMs for classification tasks by comparing their performance.

The application programming interface (API) for each LLM was obtained from its respective developer. An API is a system that enables applications to interact, allowing programs to send instructions to an LLM and receive responses. The prompt embedded in the code instructed the model to categorize each comment from 143 participants into one of three categories: Planning (before writing—organizing ideas, setting goals, e.g., "*I will write a draft in Japanese and then translate it.*"), Monitoring (while writing—adjusting grammar, vocabulary, coherence, e.g., "*I will avoid repeating words and use varied vocabulary.*"), or Evaluation (after writing—revising, proofreading, reflecting, e.g., "*I will check my grammar with Grammarly and make corrections.*"). The model was required to return only the category name (see the online supplementary material for reference). However, at the time of analysis, the API for DeepSeek-V3 was temporarily unavailable. Consequently, we manually entered each of the 143 responses into the browser-based version of the model one by one.

Initial testing with zero-shot classification (where LLMs were asked to classify without examples) yielded unsatisfactory results. We therefore developed a structured prompt that included category definitions and examples. The same prompt was used across all LLMs to ensure consistency and fair comparison. The prompt was embedded in our API calls as follows:

Please categorize each comment as Planning (before writing: organizing ideas, setting goals, e.g., "I will write a draft in Japanese and then translate it"), Monitoring (while writing: adjusting grammar, vocabulary, coherence, e.g.,"I will avoid repeating words and use varied vocabulary"), or Evaluation (after writing: revising, proofreading, reflecting, e.g., "I will check my grammar with Grammarly and make corrections"), and return only one category name.

This prompt design incorporated three key elements: (1) clear definitions of each category, (2) representative examples for each category, and (3) explicit instructions to return only the category name. We found that including examples for each category significantly improved classification accuracy compared to definitions alone, while the instruction to return only the category name ensured consistent output formatting across all models.

To assess the accuracy of the classifications, we compared the LLM-generated categories with a human gold standard (benchmark). In addition to calculating the simple matching rate, we computed Cohen's kappa coefficient (κ). Cohen's kappa coefficient is a statistic that is used to measure inter-rater reliability for qualitative (categorical) items. Cohen (1960) suggested that kappa values are interpreted as follows: $\kappa \leq 0$ indicates no agreement, 0.01–0.20 represents none to slight agreement, 0.21–0.40 is fair, 0.41–0.60 is moderate, 0.61–0.80 is substantial, and 0.81–1.00 is almost perfect agreement. This places $\kappa = 0.8$ at the border between substantial and almost perfect agreement. While values in the 0.61–0.80 range are traditionally considered substantial, many fields adopt more stringent standards, particularly in research contexts where high reliability is crucial. McHugh (2012) notes that in several disciplines, especially medical and reliability studies, researchers often treat $\kappa \geq 0.8$ as a benchmark for strong reliability suitable for drawing definitive conclusions. Given the objective of this study—to assess whether LLMs could approach expert-level (human) performance in qualitative coding—we adopted this more stringent interpretation, using $\kappa \geq 0.8$ as our target for strong agreement while recognizing that values in the 0.60–0.79 range still indicate moderate to strong agreement with potential practical utility.

All text analyses in this study were conducted using Python (version 3.11.7). For statistical analysis and visualization, we used R (version 4.4.1). To ensure reproducibility and transparency in our data analysis (In'nami et al., 2022), all data and Python/R code used in this study are shared on OSF (https://osf.io/6bnjy/).

Table 1
Number of categories assigned ($N = 143$).

Category	Count	Key Concepts	Example
Planning	78	Before writing: organizing ideas, setting goals	"I will write a draft in Japanese and then translate it."
Monitoring	54	While writing: adjusting grammar, vocabulary, coherence	"I will avoid repeating words and use varied vocabulary."
Evaluation	11	After writing: revising, proofreading, reflecting	"After drafting my essay, I will check my grammar with Grammarly and make corrections."

Table 2
Description of LLMs used for analysis.

Model Name	Developer	Characteristics
GPT-4o	OpenAI	Flagship model used in ChatGPT (as of February 2025). Large context window (128 K tokens), strong reasoning, multimodal capabilities, and suitable for general text tasks.
GPT-o1	OpenAI	Specialized for complex logical reasoning. Performs well on structured outputs and multi-step problem-solving but has a higher cost.
GPT-o3mini	OpenAI	Efficient balance of speed and accuracy. Offers strong reasoning at lower cost and faster response times than larger GPT models.
Llama3.3–70B	Meta	Open-source, large-scale model with strong performance. Supports flexible fine-tuning for custom applications.
Gemini2.0-Flash	Google	Optimized for speed with multimodal input processing. Handles extremely long contexts with fast inference.
Claude3.5-Sonnet	Anthropic	High accuracy in language and reasoning tasks. Offers a very long context window but is relatively expensive compared to other models.
DeepSeek-V3	DeepSeek	Open-source mixture-of-experts model. Provides strong performance at low cost. While its inference speed is slower than other models, it remains competitive with GPT-4 class models.

Results

Overall performance of LLMs

Table 3 summarizes the results of the analysis, including the percentage agreement between each LLM and the human gold standard, as well as Cohen's kappa coefficient (κ) with 95 % confidence intervals. The percentage agreement represents the proportion of responses classified identically by the LLM and human annotators, while Cohen's kappa accounts for chance agreement, providing a more robust measure of classification reliability.

Model comparison

Among the models, DeepSeek-V3 achieved the highest accuracy, with a percentage agreement of 83.2 % and a κ value of 0.68 (95 % CI: 0.58–0.79), indicating moderate agreement bordering on strong, based on McHugh's (2012) scale. Llama3.3–70B and GPT-o3mini also demonstrated relatively strong performance, both achieving an agreement rate of 77.6 %, with κ values of 0.61 (95 % CI: 0.50–0.72) and 0.60 (95 % CI: 0.48–0.71), respectively, reaching the moderate agreement range.

Other models, including GPT-4o, GPT-o1, Gemini2.0-Flash, and Claude3.5-Sonnet, exhibited lower levels of agreement, with κ values ranging from 0.37 to 0.49, indicating minimal-to-weak agreement according to McHugh's (2012) standard. Surprisingly, Claude3.5-Sonnet—despite being known for its strong performance in language and reasoning tasks—produced the lowest κ value (0.37, 95 % CI: 0.25–0.48) among the seven models tested. These findings are visually illustrated in Fig. 1, where DeepSeek-V3 shows the highest κ value, approaching 0.8, while Claude3.5-Sonnet exhibits the lowest agreement among all models.

These results led to several unexpected findings. First, DeepSeek-V3, the most recently released model in our comparison, demonstrated the highest classification accuracy, approaching $\kappa = 0.8$, which is often considered the benchmark for strong agreement. While its strong performance was noteworthy, it was particularly surprising given that DeepSeek is an open-source model, freely accessible to users, unlike GPT-o1, GPT-o3mini, and Claude3.5-Sonnet, which require paid access. Similarly, Llama3.3–70B, another open-source model, outperformed several proprietary models, suggesting that open-source LLMs are becoming increasingly competitive with their commercial counterparts.

Note. Cohen's kappa values are interpreted as follows: 0–.20 = None, 0.21–.39 = Minimal, 0.40–.59 = Weak, 0.60–.79 = Moderate, 0.80–.90 = Strong, and Above 0.90 = Almost Perfect (McHugh, 2012).

Misclassification patterns

Conversely, the relatively poor performance of Claude3.5-Sonnet raises interesting questions. Despite its reputation for strong natural language understanding and reasoning tasks, it performed the worst in classifying responses into the three predefined categories (planning, monitoring, and evaluation). One possible explanation is that this particular classification task—assigning open-

Table 3
Classification accuracy of LLMs.

LLM Model	Percentage agreement	Cohen's kappa coefficient (95 % CI)		
		Lower	Estimate	Upper
GPT-4o	68.5	0.27	0.39	0.51
GPT-o1	71.3	0.33	0.45	0.57
GPT-o3mini	77.6	0.48	0.60	0.71
Llama3.3–70B	77.6	0.50	0.61	0.72
Gemini2.0-Flash	72.0	0.38	0.50	0.61
Claude3.5-Sonnet	67.2	0.25	0.37	0.48
DeepSeek-V3	83.2	0.58	0.68	0.79

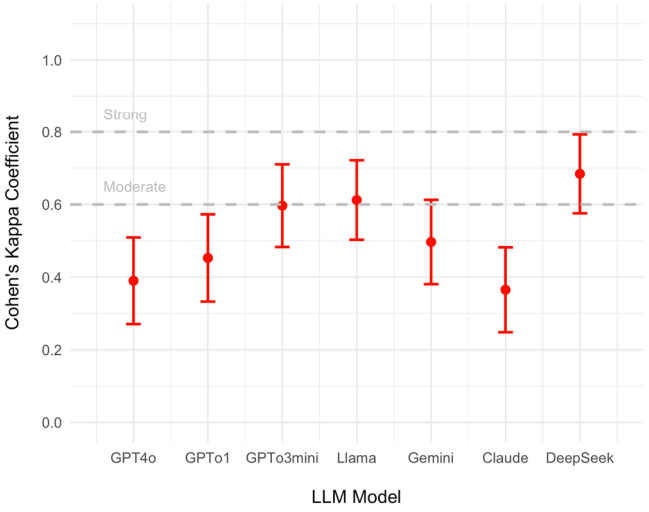

Fig. 1. Visual representation of the results.

ended text responses from learners to predefined self-regulatory learning strategy categories—may not align well with the model's strengths.

A notable example is the following response, which human annotators categorized as "monitoring," but all LLMs misclassified as "planning": "*I think everything through and write it down first, then look up the parts I'm not confident about online.*" However, the definition of "monitoring" involves checking one's output (e.g., writing or speaking), identifying errors or uncertainties, and making corrections or adjustments accordingly. The sequence "write everything down → look up the uncertain parts" suggests that the learner is actively recognizing and addressing weaknesses in their output. This process aligns more closely with self-checking and self-correction, both of which are key characteristics of monitoring.

One possible reason for this misclassification is that LLMs rely heavily on pattern recognition. If a sentence contains structures such as "first… then…," the model may overgeneralize and assume that any structured sequence of actions represents a planning process, even when it more accurately fits the definition of monitoring. A similar pattern of misclassification appeared in other responses. For instance, the sentence "*Write as much as I can in English, and then look up the phrases I cannot express*" was categorized as "monitoring" only by Llama3.3–70B, while all other models classified it as "planning." Likewise, the response "*I choose a topic from something familiar and repeatedly revise my writing*" should have been categorized as "monitoring," yet all models except DeepSeek-V3 misclassified it as "evaluation." This misclassification suggests that LLMs may associate "revision" with evaluation processes, particularly in academic contexts where revision often occurs after assessment. However, revision during the writing process is a monitoring strategy rather than an evaluation strategy. This indicates that LLMs may misinterpret revision as a final evaluative process rather than an ongoing monitoring activity.

Summary of findings

All in all, these findings suggest that while LLMs show promise in classification tasks, they have not yet reached the level required to fully replace human annotation. Notably, none of the models achieved $\kappa \geq 0.8$, which we established as our stringent benchmark for strong agreement based on McHugh's (2012) recommendations. However, it is worth emphasizing that several models, particularly DeepSeek-V3 ($\kappa = 0.68$), achieved moderate-to-strong agreement that approaches our high threshold. This level of performance, while falling short of our stringent criterion for standalone reliability, nonetheless suggests meaningful alignment with human judgments and indicates potential practical value for researchers using LLMs as preliminary coding aids or assistive tools in qualitative analysis. These results indicate that, in their current state, LLMs cannot fully replace human judgment in categorizing free-text responses for research purposes, but they may serve as valuable complementary tools. Further investigation into fine-tuning models or refining prompt design may be necessary to improve classification accuracy toward the higher threshold of $\kappa \geq 0.8$.

Discussion

The findings of this study contribute to the growing body of research on the use of LLMs in qualitative data classification, highlighting both their potential and limitations. While some models, particularly DeepSeek-V3 and Llama3.3–70B, demonstrated moderate agreement (as defined by McHugh, 2012) with human classifications ($\kappa = 0.68$ and $\kappa = 0.61$ respectively), none reached the widely accepted $\kappa \geq 0.8$ threshold for strong inter-rater reliability. These results suggest that although LLMs can support qualitative analysis, they are not yet sufficiently reliable to fully replace human annotators in tasks requiring nuanced interpretation. This is in line with Nielbo et al.'s (2024) comprehensive review of quantitative text analysis, which emphasizes that while modern generative models

have made significant advances, they should be viewed as tools to enhance rather than replace human expertise. This view is further supported by Wachinger et al.'s (2024) empirical investigation of ChatGPT in qualitative analysis, which found that while the model could identify descriptive themes effectively, it struggled with more interpretative analyses requiring context-dependent understanding.

Our findings directly connect to the theoretical perspectives introduced earlier regarding the principles of qualitative coding and human-AI collaboration. Notably, the fact that no model achieved the $\kappa \geq 0.8$ "strong agreement" threshold reinforces the view that human expertise remains indispensable—a point consistent with the human–AI collaboration paradigm in qualitative research. For instance, when classifying responses containing sequential actions, LLMs consistently misinterpreted planning versus monitoring processes due to their reliance on pattern recognition rather than deeper understanding of the activities' temporal context. This alignment between our empirical results and the theoretical framework reinforces the conclusion that LLMs, in their current state, function best as assistive tools working alongside human analysts rather than autonomous coders—supporting the human-AI collaborative approach advocated in qualitative methodology literature.

Despite this limitation, the observed levels of agreement still offer practical insights for real-world applications. While we adopted the stringent $\kappa \geq 0.8$ criterion to assess whether LLMs could approach expert-level reliability, the moderate agreement achieved by models like DeepSeek-V3 ($\kappa = 0.68$) and Llama3.3–70B ($\kappa = 0.61$) suggests these tools may still offer considerable value in applied research contexts. For instance, researchers working with large datasets might use these LLMs for preliminary coding to accelerate the process, followed by human verification. In contexts where perfect reliability is less critical or where LLMs serve as assistive rather than autonomous tools, κ values in the 0.60–0.79 range may provide sufficient practical utility while significantly reducing the time and resources required for qualitative analysis. This perspective aligns with Fuller et al.'s (2024) finding that AI-assisted analysis of student evaluations offered significant time savings while maintaining reasonable agreement with human coding.

One key finding from our study is the divergence in classification strategies between human raters and LLMs. As previous research has shown, humans rely on contextual understanding and real-world knowledge, allowing them to interpret implicit meanings and resolve ambiguities that LLMs may struggle with (Wachinger et al., 2024). In contrast, LLMs categorize responses based on statistical correlations learned from training data, which can lead to consistent but sometimes superficial classifications. For instance, our results indicated that LLMs occasionally misclassified responses that contained multiple overlapping elements of planning, monitoring, and evaluation—an issue that human raters resolved through context-driven decision-making.

This tendency was also evident in our results and mirrors findings in prior research, where LLMs tend to over-rely on surface cues, such as specific keywords, rather than deeper contextual meaning. While this approach can yield high agreement for clear-cut classifications, it also introduces systematic errors, particularly when faced with ambiguous or novel responses. This limitation underscores the importance of retaining human oversight in LLM-assisted qualitative coding, particularly in research contexts where interpretative accuracy is critical.

To address this issue, it is worth exploring how specific interventions—such as prompt refinement—might enhance model performance. Research has shown that prompt engineering—particularly the use of few-shot prompting and enriched task instructions—can enhance LLM accuracy in qualitative classification tasks. Our study used a structured prompt specifying category definitions and examples, yet results indicate that additional refinements may be necessary.

To address these limitations, we propose several specific strategies for enhancing LLM classification accuracy in qualitative research contexts. First, researchers could implement structured few-shot prompting by including 3–5 deliberately selected examples that represent edge cases or commonly misclassified responses. For instance, in our study, providing examples like "Write as much as I can in English, and then look up the phrases I cannot express" (monitoring) and "I choose a topic from something familiar and repeatedly revise my writing" (monitoring) could help models distinguish between planning and monitoring processes. Second, chain-of-thought prompting could improve classification by explicitly instructing the LLM to reason through specific criteria: "First, identify whether the action occurs before writing (planning), during writing (monitoring), or after completion (evaluation). Then, check if the response describes organizing ideas, adjusting content, or reviewing completed work." Third, researchers might employ contrastive examples—deliberately pairing similar responses with different classifications to highlight subtle distinctions, such as "I will plan my essay structure before writing" (planning) versus "I will check my essay structure as I write" (monitoring). Additionally, implementing a confidence scoring system where LLMs assign probability values to each category could help researchers identify ambiguous cases requiring human verification. For example, researchers could use a predefined confidence threshold (e.g., 70 %) to flag low-confidence classifications for manual review. Confidence scores could also be visualized across the dataset to identify patterns of ambiguity, which may inform revisions to coding schemes or prompt design. These approaches warrant systematic investigation in future research to determine which yields the highest classification accuracy for specific qualitative analysis tasks.

Building on these results, we next consider potential reasons why certain models outperformed others, focusing on architectural and training-related factors. Our findings that open-source models like DeepSeek-V3 and Llama3.3–70B outperformed some commercial options such as GPT-4o and Claude3.5-Sonnet warrant further consideration. This performance difference may stem from several factors, including architectural design, training data characteristics, and model recency. DeepSeek-V3's mixture-of-experts architecture appears particularly effective for structured classification tasks, while both DeepSeek-V3 and Llama3.3–70B likely benefit from highly curated, quality-focused training datasets rather than relying solely on scale.

The recency advantage of DeepSeek-V3—being the newest model in our comparison—may also contribute to its superior performance, as it likely incorporates the latest advancements in model architecture and training techniques. Conversely, the unexpectedly poor performance of Claude3.5-Sonnet may reflect optimization for nuanced, explanatory responses rather than direct classification tasks requiring single-label outputs without explanation. These observations highlight the importance of selecting models based on specific task requirements rather than general reputation or commercial status.

From a practical standpoint, our findings indicate that LLMs can serve as a valuable tool for assisting in qualitative classification tasks, particularly in large-scale studies where manual annotation is time-consuming. However, their use should be approached with caution. Nielbo et al. (2024) highlight the importance of maintaining methodological rigor and transparency when using these models, particularly regarding data sources and model limitations. As previous research has highlighted, LLMs can introduce biases based on their training data, potentially skewing classification results, which necessitates careful implementation and human oversight in research contexts.

While our study provides valuable insights, it is not without limitations. First, our classification task involved a relatively small dataset ($N = 143$), making it difficult to generalize the findings to larger or more diverse datasets. Second, because we analyzed the translation of open-ended responses, the results may have been influenced by this factor. Additionally, our study relied solely on a predefined categorization scheme, which may limit applicability to other qualitative analysis contexts. Future research should address these limitations and explore alternative approaches.

Despite these limitations, our findings provide valuable insights into the current capabilities and constraints of LLMs in qualitative data analysis. The following section summarizes our key conclusions and offers practical recommendations for researchers interested in integrating these tools into their qualitative research workflows.

Conclusion

Overall, this study reinforces the potential of LLMs as tools for qualitative data classification while highlighting their current limitations. Although no model achieved the threshold ($\kappa \geq 0.8$) required for expert-level reliability, some models, especially recent open-source ones, demonstrated moderate agreement with human coders, indicating practical value for preliminary coding tasks. Importantly, these findings suggest that the effective integration of LLMs into qualitative research workflows depends not only on model selection, but also on thoughtful human–AI collaboration strategies.

Practical implications

Our results offer several important practical implications for researchers considering the integration of LLMs into qualitative analysis workflows. First, LLMs can serve as valuable assistive tools that significantly reduce the time and resources required for qualitative classification, particularly with large datasets. Models achieving moderate agreement ($\kappa = 0.60$–0.79) may provide sufficient practical utility for preliminary coding followed by human verification. Second, researchers should implement a validation phase before broader deployment, comparing LLM-generated classifications against human-coded benchmarks to identify potential systematic errors. Third, transparency in AI-assisted qualitative analysis is crucial; researchers should explicitly disclose LLM usage in classification tasks to ensure interpretability and reproducibility. Finally, researchers should remain mindful that LLMs may overly depend on surface cues rather than deeper contextual meaning, necessitating human oversight, especially in ambiguous or novel cases.

Future directions

Looking ahead, advancing prompt engineering techniques, implementing confidence-based filtering, and exploring model fine-tuning are promising directions for future research. Structured few-shot prompting with carefully selected edge cases, chain-of-thought prompting to explicitly guide classification reasoning, and the use of contrastive examples to clarify subtle distinctions should be systematically explored. Implementing confidence scoring systems could also help researchers efficiently identify ambiguous classifications requiring further human inspection. Additionally, future studies should examine the impact of domain-specific fine-tuning and assess model performance across larger, more diverse datasets and different qualitative coding frameworks. In the longer term, these developments could pave the way for hybrid coding systems where LLMs and human analysts work in tandem to achieve both scale and interpretative depth—fundamentally transforming qualitative research practices in applied linguistics and beyond.

CRediT authorship contribution statement

Atsushi Mizumoto: Writing – review & editing, Writing – original draft, Validation, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Mark Feng Teng:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

In the preparation of this manuscript, we employed ChatGPT (GPT-4o) to enhance the clarity and coherence of the language, ensuring it adheres to the standards expected in scholarly journals. While ChatGPT played a role in refining the language, it did not

contribute to the generation of any original ideas. The authors alone are responsible for any inaccuracies present in the manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.rmal.2025.100210.

References

Allen, T. J., & Mizumoto, A. (2024). ChatGPT over my friends: Japanese English-as-a-foreign-language learners' preferences for editing and proofreading strategies. *RELC Journal*. https://doi.org/10.1177/00336882241262533

Aryadoust, V., Zakaria, A., & Jia, Y. (2024). Investigating the affordances of OpenAI's large language model in developing listening assessments. *Computers and Education: Artificial Intelligence, 6*, Article 100204. https://doi.org/10.1016/j.caeai.2024.100204

Barros, C.F., Azevedo, B.B., Neto, V.V.G., Kassab, M., Kalinowski, M., Nascimento, H.A.D. do, & Bandeira, M.C.G.S.P. (2024). *Large language model for qualitative research: A systematic mapping study* (Version 3). arXiv. doi:10.48550/ARXIV.2411.14473.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics, 3*(3), Article 100066. https://doi.org/10.1016/j.acorp.2023.100066

Fuller, K. A., Morbitzer, K. A., Zeeman, J. M., Persky, A. M., Savage, A. C., & McLaughlin, J. E. (2024). Exploring the use of ChatGPT to analyze student course evaluation comments. *BMC Medical Education, 24*(1), 423. https://doi.org/10.1186/s12909-024-05316-2

Huang, J., & Mizumoto, A. (2024). Examining the effect of generative AI on students' motivation and writing self-efficacy. *Digital Applied Linguistics, 1*, Article 102324. https://doi.org/10.29140/dal.v1.102324

Huang, J., & Teng, M. F. (2025). Peer feedback and ChatGPT-generated feedback on Japanese EFL students' engagement in a foreign language writing context. *Digital Applied Linguistics, 2*, Article 102469. https://doi.org/10.29140/dal.v2.102469

In'nami, Y., Mizumoto, A., Plonsky, L., & Koizumi, R. (2022). Promoting computationally reproducible research in applied linguistics: Recommended practices and considerations. *Research Methods in Applied Linguistics, 1*(3), Article 100030. https://doi.org/10.1016/j.rmal.2022.100030

Kim, M., & Lu, X. (2024). Exploring the potential of using ChatGPT for rhetorical move-step analysis: The impact of prompt refinement, few-shot learning, and fine-tuning. *Journal of English for Academic Purposes*. , Article 101422. https://doi.org/10.1016/j.jeap.2024.101422

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal, 54*(2), 537–550. https://doi.org/10.1177/00336882231162868

Lee, V. V., Van Der Lubbe, S. C. C., Goh, L. H., & Valderas, J. M. (2024). Harnessing ChatGPT for thematic analysis: Are we ready? *Journal of Medical Internet Research, 26*, Article e54974. https://doi.org/10.2196/54974

Lin, P. (2023). ChatGPT: Friend or foe (to corpus linguists)? *Applied Corpus Linguistics, 3*(3), Article 100065. https://doi.org/10.1016/j.acorp.2023.100065

Lo, C. K., Hew, K. F., & Jong, M. S. (2024). The influence of ChatGPT on student engagement: A systematic review and future research agenda. *Computers & Education, 219*, Article 105100. https://doi.org/10.1016/j.compedu.2024.105100

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, 276–282. https://doi.org/10.11613/BM.2012.031

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), Article 100050. https://doi.org/10.1016/j.rmal.2023.100050

Mizumoto, A., Shintani, N., Sasaki, M., & Teng, M. F. (2024). Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment. *Research Methods in Applied Linguistics, 3*(2), Article 100116. https://doi.org/10.1016/j.rmal.2024.100116

Moorhouse, B. L., Wan, Y., Ho, T. Y., & Lin, A. M. Y. (2024). Generative AI-assisted, evidence-informed use of L1 in L2 classrooms. *ELT Journal, 78*, 453–465. https://doi.org/10.1093/elt/ccae033

Morgan, D. L. (2023). Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. *International Journal of Qualitative Methods, 22*, Article 16094069231211248. https://doi.org/10.1177/16094069231211248

Nielbo, K. L., Karsdorp, F., Wevers, M., Lassche, A., Baglini, R. B., Kestemont, M., & Tahmasebi, N. (2024). Quantitative text analysis. *Nature Reviews Methods Primers, 4*(1), 25. https://doi.org/10.1038/s43586-024-00302-w

Schroeder, H., Quéré, M.A.L., Randazzo, C., Mimno, D., & Schoenebeck, S. (2024). *Large language models in qualitative research: Uses, tensions, and intentions* (Version 2). arXiv. doi:10.48550/ARXIV.2410.07362.

Shin, D. (2023). Can ChatGPT make reading comprehension testing items on par with human experts? *Language Learning & Technology, 27*(3), 27–40. https://hdl.handle.net/10125/73530.

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, Article 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Teng, L. S., & Zhang, L. J. (2016). A questionnaire-based validation of multidimensional models of self-regulated learning strategies. *The Modern Language Journal, 100*(3), 674–701. https://doi.org/10.1111/modl.12339

Teng, M. F. (2020). The role of metacognitive knowledge and regulation in mediating university EFL learners' writing performance. *Innovation in Language Learning and Teaching, 14*(5), 436–450. https://doi.org/10.1080/17501229.2019.1615493

Teng, M. F. (2024). A systematic review of ChatGPT for English as a foreign language writing: Opportunities, challenges, and recommendations. *International Journal of TESOL Studies, 6*(3), 36–57. https://doi.org/10.58304/ijts.20240304

Teng, M. F. (2025). Metacognitive awareness and EFL learners' perceptions and experiences in utilising ChatGPT for writing feedback. *European Journal of Education, 60*(1), Article e12811. https://doi.org/10.1111/ejed.12811

Teng, M. F., Wang, C., & Zhang, L. J. (2022). Assessing self-regulatory writing strategies and their predictive effects on young EFL learners' writing performance. *Assessing Writing, 51*(January 2022), Article 100573. https://doi.org/10.1016/j.asw.2021.100573

Todd, R. W. (2025). Generative AI as a disrupter of language education. *International Journal of TESOL Studies*. , Article 250127. https://doi.org/10.58304/ijts.250127

Uchida, S., & Negishi, M. (2025). Assigning CEFR-J levels to English learners' writing: An approach using lexical metrics and generative AI. *Research Methods in Applied Linguistics, 4*(2), Article 100199. https://doi.org/10.1016/j.rmal.2025.100199

Wachinger, J., Bärnighausen, K., Schäfer, L. N., Scott, K., & McMahon, S. A. (2024). Prompts, pearls, imperfections: Comparing ChatGPT and a human researcher in qualitative data analysis. *Qualitative Health Research*. https://doi.org/10.1177/10497323241244669, 10497323241244669.

Xin, J. J. (2024). Investigating EFL teachers' use of generative AI to develop reading materials: A practice and perception study. *Language teaching research*. https://doi.org/10.1177/13621688241303321, 13621688241303321.

Yang, L., & Li, R. (2024). ChatGPT for L2 learning: Current status and implications. *System, 124*, Article 103351. https://doi.org/10.1016/j.system.2024.103351

Yu, D. (2025). Towards LLM-assisted move annotation: Leveraging ChatGPT-4 to analyse the genre structure of CEO statements in corporate social responsibility reports. *English for Specific Purposes, 78*, 33–49. https://doi.org/10.1016/j.esp.2024.11.003

Zimmerman, B. J., Boekaerts, M., Pintrich, P. R., & Zeidner, M. (2000). Attaining self-regulation: A social cognitive perspective. *Handbook of self-regulation* (pp. 13–39). Academic Press. https://doi.org/10.1016/B978-012109890-2/50031-7