QRCBN 62-6861-8367-215
https://proceeding.unnes.ac.id/eltlt

Customizing mistral 7B large language model for qualitative research: A feasibility study

Hendi Pratama[1]

Author Affiliations
[1]*Universitas Negeri Semarang , Indonesia*

Abstract. In qualitative linguistic research, particularly within the domain of discourse analysis, the manual identification of pragmatic features such as Grice's conversational maxims can be time-consuming and cognitively demanding. This feasibility study investigates the potential of using the Mistral 7B large language model (LLM) to support such analysis by automating the classification of Gricean maxims, Quantity, Quality, Relevance, and Manner, and identifying corresponding illocutionary acts in Instagram captions. A dataset comprising 88 bilingual captions (primarily English with several in Indonesian) from Samsung Indonesia's official Instagram account was used. The model was prompted to analyze each caption, score the observance of the four maxims, assign an illocutionary act type, and provide justification for its classifications. The outputs were compared to a previously published human-coded analysis. Results showed that Mistral could produce accurate classifications for most captions, particularly in identifying directives and informative acts, and provided plausible justifications. However, the model displayed a bias toward higher maxim observance scores (3 and 4), showing reluctance to assign lower ratings such as "barely observed" or "not observed," which human coders used more readily. Mistral also failed to parse a syntactically complex caption, indicating limitations in handling mixed or informal structures. Overall, the findings highlight Mistral's potential as a fast, accessible tool for supporting qualitative linguistic inquiry, especially in large-scale or exploratory settings. While its accuracy and interpretive depth require refinement, Mistral offers a promising starting point for integrating AI into pragmatic analysis workflows. Further development in prompt design and model calibration is recommended.

Keywords*:* Automated Analysis, Mistral 7B, LLMs, Pragmatics, Gricean Maxims

INTRODUCTION

The increasing sophistication of large language models (LLMs) has ushered in new possibilities across various fields of inquiry, including qualitative research. Traditionally, qualitative linguistic research, particularly in the domain of discourse analysis and pragmatics, has relied heavily on manual coding processes. This reliance demands substantial time and interpretive labor, especially when identifying nuanced phenomena such as conversational implicatures and the application of Grice's Cooperative Principle (Grice, 1975). As research questions grow more complex and datasets expand, there is an urgent need for scalable tools that can support, rather than replace, human judgment in these interpretive tasks.

Large language models like OpenAI's GPT-4 and open-source alternatives such as Mistral 7B have demonstrated remarkable capabilities in natural language understanding and generation. These models are trained on extensive corpora and have shown promise in identifying rhetorical structures, predicting communicative intent, and replicating stylistic features of human dialogue. Given these developments, an important question emerges: "can LLMs be adapted to assist in qualitative tasks traditionally considered outside the realm of automation?"

This paper presents a feasibility study of customizing the Mistral 7B model for discourse-pragmatic analysis. Specifically, the study evaluates the model's ability to classify Instagram marketing captions based on Grice's conversational maxims Quality, Quantity, Relevance, and Manner and to identify the primary illocutionary act conveyed. By building an accessible workflow using Google Colab and Python, the study explores how LLMs can function as assistants to qualitative researchers, helping reduce the workload associated with interpretive coding while maintaining analytical rigor.

Gricean Pragmatics and Qualitative Analysis

Grice's Cooperative Principle has long served as a foundational framework in pragmatics and discourse analysis. His theory posits that effective communication relies on adherence to four conversational maxims: Quality (truthfulness), Quantity (informativeness), Relevance, and Manner (clarity) (Grice, 1975). Identifying whether these maxims are observed or flouted has helped linguists interpret speakers' intentions and the pragmatic implications of their statements.

In applied settings, such as marketing or digital discourse, identifying these maxims is crucial for understanding how brands convey information, build rapport, and persuade audiences. However, manual coding of these maxims is labor-intensive and inherently subjective. This has led researchers to explore whether computational tools can assist in or augment such interpretive processes.

Large Language Models in Discourse and Pragmatics

Recent advancements in artificial intelligence, particularly in LLMs, have revitalized interest in automating linguistic tasks previously considered too nuanced for machines. GPT-4, for example, has demonstrated strong capabilities in interpreting pragmatics, even outperforming human subjects in certain dialogue-based tasks that rely on Grice's principles (Gebreegziabher et al, 2023). In this study, GPT-4 outperformed humans in identifying implicature and context-based meaning, suggesting that LLMs may have the potential to support or accelerate pragmatic analysis in research.

Mistral 7B has been explored for its capability to provide reasoned judgments in text synthesis and evaluation tasks. Evans et al. (2024) compared Mistral to GPT-4 in evaluating the quality of scientific summaries and found that while both models provided logical explanations, their alignment with human ratings varied. The findings suggest that LLMs can produce coherent reasoning but still require human oversight to ensure interpretive accuracy.

Further evidence of LLMs' discourse capabilities comes from research on dialogue breakdown detection. Vatsal and Dubey (2024) evaluated models like Mistral and GPT-4 in identifying incoherence and irrelevance in dialogue. While GPT-4 consistently performed better, prompting techniques such as chain-of-thought improved Mistral's performance significantly, indicating that careful prompt engineering can extend the usefulness of smaller models in qualitative analysis.

LLMs in Qualitative Data Analysis

Beyond pragmatics, LLMs have also been deployed for broader qualitative data analysis tasks. For instance, Long et al. (2024) explored the use of GPT-4 and Mistral for coding requirements engineering data. Their findings revealed that detailed, context-specific prompts could significantly enhance annotation accuracy, particularly in deductive settings. These results are echoed by Randerson et al. (2025), who assessed GPT-4's ability to identify inductive codes in a discourse network analysis. The model successfully retrieved over two-thirds of human-identified codes but also introduced noise, emphasizing the need for human review in final analysis stages.

These studies highlight both the opportunities and constraints of using LLMs for qualitative tasks. While they can streamline the identification of patterns and support the creation of codebooks, they still require domain knowledge and careful calibration to avoid interpretive errors.

Feasibility of Open-Source Models in Humanities Research

Despite the capabilities of commercial models like GPT-4, their proprietary nature and resource demands limit their accessibility to many researchers. In contrast, open-source models like Mistral offer a more viable option for integration into research workflows, particularly in settings with limited computational resources or funding.

Kreikemeyer et al. (2025) demonstrated how a 7B-parameter Mistral model, fine-tuned on domain-specific data, could perform reasonably well in translating natural language to formal simulation models. Although it did not outperform larger models, the open-source Mistral's results were considered practically sufficient and replicable for academic use.

Those findings support the feasibility of using smaller models for qualitative research. With well-crafted prompts and contextual framing, models like Mistral 7B can assist in interpretive tasks while remaining computationally lightweight and cost-effective.

METHODS

This study explores the feasibility of using a large language model (LLM), specifically Mistral 7B, to conduct pragmatic analysis of social media captions within a qualitative linguistic research framework. The research was motivated by the need to reduce the time-intensive nature of traditional discourse analysis, especially when analyzing texts using Grice's Cooperative Principle. The study was conducted as a comparative evaluation in which the output of the Mistral 7B model was assessed against a set of human-coded annotations previously published by Pratama et al. (2025). This dual approach enabled a systematic analysis of alignment between human interpretive judgment and machine-assisted annotation using a replicable, lightweight computational process.

The dataset consisted of 88 Instagram captions published by Samsung Indonesia during August 2024. These captions were selected based on their communicative variety, ranging from promotional calls to action to purely descriptive statements, which made them suitable for assessing the four Gricean maxims: Quantity, Quality, Relevance, and Manner (Grice, 1975). In the human-coded baseline, each caption had previously been analyzed using a 4-point scale to determine the degree to which each maxim was observed or flouted, and categorized according to its illocutionary act type using Searle's (1976) taxonomy of speech acts. The human annotation process employed a three-coder consensus model, ensuring intersubjective validity and interpretive rigor (Pratama et al., 2025). This made the dataset ideal for testing the effectiveness of automated linguistic analysis using a transformer-based model.

The computational side of the study was carried out using the Mistral 7B-Instruct v0.3 model, an open-source LLM known for its balance between performance and resource efficiency. The model was deployed via Hugging Face's Transformers library in Google Colab and loaded using 4-bit quantization to optimize memory usage. This configuration enabled fast and accessible execution on cloud GPUs without requiring advanced hardware or technical expertise, an important consideration for qualitative researchers unfamiliar with machine learning infrastructure. Researchers have noted the increasing accessibility of open-source LLMs like Mistral as a way to democratize NLP tools for use in the humanities and social sciences (Kreikemeyer et al., 2025; Long et al., 2024).

Each Instagram caption was fed into the model via a structured prompt that explained the task using clear definitions and a fixed annotation schema. The prompt instructed the model to score each of the four maxims on a scale from 1 to 4, assign one or more illocutionary act types (e.g., directive, informative, expressive), and provide a short justification explaining its reasoning. The prompt was formulated in natural language and designed to be both comprehensive and self-contained, so that each caption could be analyzed independently. Prompt engineering has been shown to significantly affect the performance of LLMs in interpretive tasks, particularly when using few-shot or zero-shot instruction settings (Vatsal & Dubey, 2024).

```
Analyzing caption 1/88: Bringing #GalaxyZFold6 to life with a stylish new case!
Analyzing caption 2/88: 📺Catch all the action of the Samsung Odyssey Cup Dota 2
Analyzing caption 3/88: Match your room interior with making your own TV. Whethe
Analyzing caption 4/88: Taking wellness to new heights! 🌿 ✨ Our Wellness Getaw
Analyzing caption 5/88: YOU STILL GOT THE TIME TO WIN 5 #ENHYPEN SIGNED T-SHIRT
Analyzing caption 6/88: Mau cerita sedikit, dari dulu tuh SkyAvenue menjadi sala
Analyzing caption 7/88: Go beyond of your hobby and entertainment with AI Pictur
Analyzing caption 8/88: Where technology meets art, it's possible to browse on 2
Analyzing caption 9/88: Siapa yang sekarang mulai rutin Yoga buat achive hidup s
Analyzing caption 10/88: See how #JAY, #JAKE and #JUNGWON get delicious dessert
Analyzing caption 11/88: Did #HEESEUNG pick out the right shirt for #NI_KI? See
Analyzing caption 12/88: Zoom in, snap a shot, then let #GalaxyAI turn it into a
Analyzing caption 13/88: Udara di luar lagi buruk dan kamu pengen udara di rumah
Analyzing caption 14/88: Take a wind down-time after a tiring day with your favo
Analyzing caption 15/88: Get your Limited Awesome Gaming Package #GalaxyA55 5G n
Analyzing caption 16/88: Activate Listening Mode on Interpreter to travel with e
Analyzing caption 17/88: AI Picture transforms travel vlog, action movies, docum
Analyzing caption 18/88: With Active Voice Amplifier Pro and 8K AI Upscaling Pro
Analyzing caption 19/88: Speak like locals with the help of Galaxy Z Flip6 Dual
Analyzing caption 20/88: Obsessed with how #GalaxyS24 Series captured the #ENHYP
Analyzing caption 21/88: Experience the new era of AI TV, where your front-row t
Analyzing caption 22/88: Now #SamsungAITV is here, meet the innovation that will
Analyzing caption 23/88: Time may have passed, but #NeoQLED8K instantly takes me
```

Figure 1. Mistral 7B doing caption-per-caption analysis

To extract the model's outputs accurately, regular expressions were used to parse scores, classification labels, and justifications from each response. These were compiled into a structured DataFrame and exported as a CSV file. If the model output was invalid, a placeholder was recorded; however, all 88 captions yielded valid, parsable responses.

The resulting dataset included 88 rows with seven key variables: caption text, maxim scores, speech act classification, and justification. These outputs were compared with human-coded annotations through a three-part analysis: (1) quantitative comparison of maxim scores, (2) alignment of speech act classifications, and (3) review of model-generated justifications against those from Pratama et al. (2025) for interpretive depth.

This mixed-methods approach follows best practices in computational linguistics, combining metrics with human-in-the-loop validation (Evans et al., 2024; Randerson et al., 2025). As shown in related studies (e.g., Long et al., 2024), LLMs are not expected to replicate human judgment but can support qualitative analysis by reducing time and cognitive effort.

The technical setup was lightweight and fully executable in a single Google Colab notebook, using only Python, ensuring accessibility for non-technical users. Open-source tools and cloud platforms have made AI integration in humanities research more feasible (Gebreegziabher et al., 2023; Kreikemeyer et al., 2025).

All data were public, involved no human subjects, and complied with ethical AI guidelines (Bender et al., 2021). The study used the Mistral 7B model in a prompt-based workflow that paralleled traditional qualitative coding, evaluating the model's interpretive accuracy and utility for discourse-pragmatic research.

FINDINGS AND DISCUSSION

The initial outputs of the Mistral 7B model demonstrate its clear capability in performing key discourse-pragmatic tasks on real-world marketing content. Specifically, the model was able to extract and interpret Instagram captions accurately, assigning scores for each of Grice's conversational maxims, Quality, Quantity, Relevance, and Manner, with consistency and precision. In all five example captions, the model correctly identified the extent to which each maxim was observed, with most receiving full adherence scores. This suggests the model can recognize informational completeness, contextual relevance, clarity, and truthfulness within the concise format of social media posts.

Beyond scoring, the model successfully categorized each caption's illocutionary act, aligning with classifications typically assigned in human-led speech act analysis. For instance, it distinguished between purely informative captions, clear directives, and mixed-purpose utterances that served both to inform and prompt user engagement. The nuanced labeling, such as "Mixed (Informative & Directive)," indicates the model's sensitivity to multimodal communicative intent, a core challenge in pragmatic analysis.

Table 1. Sample of The First Five Analysis by Generated by Mistral 7B

Data	Caption	Qual	Quan	Rele	Mann	Illocut Act	Justification
1	*Bringing #GalaxyZFold6 to life with a stylish new case! Tap the link in bio to explore.*	4	4	4	4	Mixed	The caption has two main goals: informing the audience about a product and prompting them to take action by clicking the link.
2	*📺Catch all the action of the Samsung Odyssey Cup Grand Finals – streaming live now!*	4	4	4	4	Directive	The caption is a clear and direct command for the user to watch a product showcase event.

3	*Match your room interior with making your own MyBespoke panels!*	4	4	4	4	Informative	The caption provides information about the customizability of a product without explicitly instructing the user.
4	*Taking wellness to new heights! 🌿✨ Our Wellness Hub is now live.*	4	4	4	4	Mixed (Informative & Directive)	The caption provides information about a wellness feature and subtly encourages engagement by highlighting availability.
5	*YOU STILL GOT THE TIME TO WIN 5 #ENHYPEN SIGNED ALBUMS!*	3	4	4	3	Mixed (Directive & Informative)	The caption is a mix of informative (providing the opportunity) and directive (urging immediate participation), though its tone is emphatic.

Mistral 7B provided justifications for each classification, demonstrating its ability to reflect on the reasoning behind its output. These justifications were clear, contextually relevant, and echoed the interpretive strategies used by trained human coders. In particular, the explanations for mixed acts revealed an understanding of dual-purpose messaging in marketing language. Overall, the model showed promise not only in producing structured annotations but also in supporting them with interpretive reasoning, affirming its potential as a qualitative research assistant.

Comparing between Mistral Classification and Human Analysis

The comparison between Mistral 7B and human coders across the four Gricean maxims, Quality, Quantity, Relevance, and Manner, reveals a consistent pattern: Mistral is significantly more lenient and skewed toward higher observance ratings, while human evaluators utilize a broader range of the scale, including lower observance categories.

Table. 2. Comparison of Maxim Classification by Mistral 7B and Human Coder

Observance Level	**Qual (Mistral)**	**Qual (Human)**	**Quant (Mistral)**	**Quan (Human)**	**Relev (Mistral)**	**Relev (Human)**	**Manner (Mistral)**	**Manner (Human)**
Fully Observed	64	47	87	71	87	84	56	61
Partially Observed	23	40	0	12	0	3	31	11
Barely Observed	0	1	0	4	0	1	0	14
Not Observed	0	0	0	1	0	0	0	2

For instance, in the Quantity and Relevance maxims, Mistral rated 87 out of 87 captions as "Fully Observed," leaving no room for partial or minimal adherence. In contrast, human coders marked only 71 and 84 captions respectively as fully observed, while also identifying several instances of "Partially," "Barely," or "Not Observed." The disparity is particularly evident in the Manner maxim, where Mistral assigned 56 captions as "Fully Observed" and 31 as "Partially Observed," but none in the lower two categories. Human coders, however, classified 14 captions as "Barely Observed" and 2 as "Not Observed," highlighting their more critical assessment of clarity, conciseness, or linguistic precision.

This skew in Mistral's scoring may stem from prompt design or the model's inherent bias toward optimistic evaluations in the absence of strong counterexamples. Unlike human coders, who rely on nuanced interpretive judgments and may penalize vague or overly promotional language, Mistral appears to default to assuming cooperative communication unless explicitly told otherwise. Additionally, Mistral's tendency to operate in a limited scoring band (3 - 4) reduces the granularity needed for fine qualitative distinctions.

These findings suggest that while Mistral is competent in recognizing and classifying maxims at a general level, its output lacks the interpretive subtlety found in human annotation. For more rigorous qualitative applications, future prompt engineering or fine-tuning may be needed to encourage the model to utilize the full evaluative spectrum.

Parsing Error in a Single Case

In this case, the Mistral 7B model failed to generate a complete or parsable analysis of the provided Instagram caption, which included a recipe and product endorsement in Indonesian. Despite Mistral's proven capacity to understand both English and Indonesian captions elsewhere in the dataset, this specific post caused a parsing error that prevented the automated system from extracting observance scores, illocutionary acts, or justifications. The issue did not appear to stem from a language barrier but more likely from structural complexity or formatting inconsistencies in the caption itself.

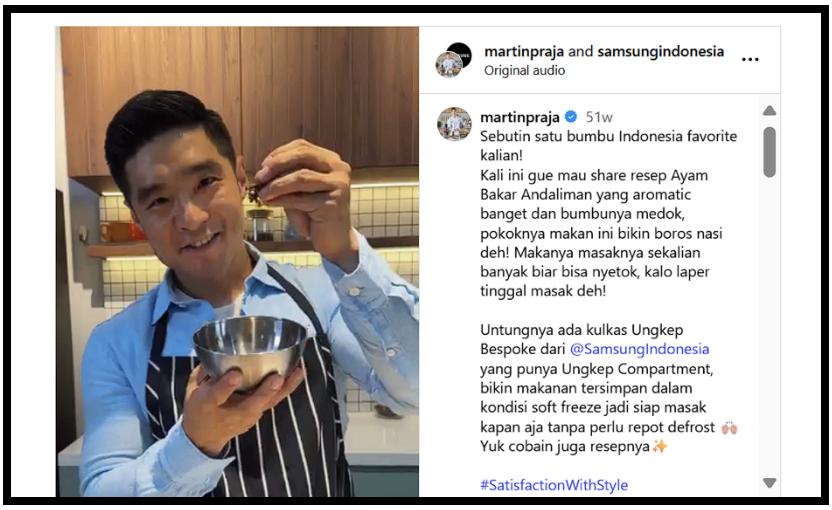

Figure 2. One captions which trigger parsing error in Mistral 7B

One likely cause of this failure is a syntax-related formatting issue. The caption includes multiple elements such as direct engagement ("Sebutin satu bumbu Indonesia favorite kalian!"), informal language ("gue," "pokoknya," "deh"), product integration with emoji and tag mentions ("@SamsungIndonesia", "🙌🏼"), hashtags, and a bulleted recipe list. This mixed content may have disrupted the model's ability to maintain context or distinguish between discourse components (e.g., directive speech acts versus list structures). Furthermore, the input structure likely violated the expected prompt-output format anticipated by the regular expression parser, resulting in a breakdown at the data extraction stage.

It is important to note that the Mistral model itself likely processed the caption internally but was unable to return its response in the required structured format. This implies that the issue lies more in post-processing or prompt design than in the model's linguistic comprehension. For future iterations, refining the prompt to handle multiline or complex caption structures, and relaxing strict parsing expectations, could help prevent similar errors. Including examples of multilingual, emoji-rich, and multi-format posts during prompt design may also improve the robustness of the output pipeline. This case highlights that technical robustness, not just linguistic understanding, is essential when deploying LLMs in applied qualitative research.

Discussion

This study aimed to assess the feasibility of using Mistral 7B, a large language model (LLM), to automate the identification of Gricean maxims and illocutionary acts in Instagram captions. The findings demonstrate that Mistral was largely successful in identifying the presence and observance of Gricean maxims, classifying speech acts, and providing justifications for its decisions. However, when contrasted with human-coded results, distinct patterns of leniency, limited score range, and sensitivity to caption structure suggest that Mistral's interpretive strategies differ substantially from those used by human analysts.

Mistral's success in extracting captions, assigning scores for each maxim, and categorizing illocutionary acts supports a growing body of literature suggesting that LLMs can be useful for automating aspects of qualitative linguistic research. For example, Krause and Vossen (2024) found that LLMs trained with pragmatic frameworks, particularly those grounded in Grice's Cooperative Principle, could approximate human-like inferences in structured texts, provided they were exposed to sufficient contextual cues and balanced prompt design. This echoes the effectiveness seen in our study, particularly for captions with clear directive or informative purposes. In many instances, Mistral's interpretation of commercial messages, such as promotional calls to action or product announcements, closely aligned with human annotators' readings. While rule-based or formal semantic models of Gricean expectations have been implemented to predict neural or behavioral responses in experimental studies (Augurzky et al., 2019), their translation into generative model outputs remains fragile, especially in informal, real-world discourse.

Differences quickly emerged when examining how each party distributed their assessments. Human coders employed a full range of scores, including lower categories such as "Barely Observed" and "Not Observed." In contrast, Mistral almost exclusively assigned "Fully" and "Partially" observed ratings, avoiding the lower end of the scale altogether. This pattern was most evident in the Manner and Quality maxims, where human coders frequently flagged vague or overly promotional language as violations, whereas Mistral treated these same utterances as generally compliant. This tendency may be explained by the model's lack of exposure to a broad evaluative rubric. Without prompt examples that explicitly differentiate between subtle violations and full observance, Mistral appears to assume cooperative intent, a behavior consistent with prior findings by Vatsal and Dubey (2024), who noted that LLMs often default to positive or neutral assumptions unless trained with counterexamples.

Mistral's inability to utilize the lower half of the scale may be symptomatic of prompt limitations rather than model incapability. Research by Long et al. (2024) emphasized that prompt specificity and calibration are essential when deploying LLMs for interpretive tasks in qualitative contexts. In our case, the prompt directed Mistral to assess adherence to Gricean maxims and assign scores from 1 to 4. However, without sample annotations or negative scoring anchors, Mistral may have lacked the semantic gradient necessary to confidently assign low scores. This reinforces the need for prompt engineering that includes a variety of observance examples and violations to enable finer evaluative distinctions.

The comparison also revealed Mistral's vulnerability to formatting irregularities and syntactic complexity. One caption, written in informal Indonesian, peppered with emojis, hashtags, and a recipe, caused a complete failure in analysis. This failure was not due to a lack of language competence; Mistral successfully processed other captions in Bahasa Indonesia. Rather, the problem appeared to stem from the multi-layered structure of the caption, which contained both narrative and list-based content. Prior studies have observed that LLMs can falter when asked to interpret or segment hybrid content without clear delimiters or hierarchy (Aporbo, 2022). Parsing such content may require not only stronger prompt formatting but also internal logic within the model's tokenizer and decoder to recognize discourse shifts within the same text.

This breakdown contrasts sharply with the human approach, where coders navigated informal expressions, slang, and hybrid content with ease. Humans were able to infer implicatures such as exaggeration, humor, or indirectness, especially within the Quality and Manner maxims, whereas Mistral exhibited a flatter interpretive profile. This gap highlights one of the central concerns raised by Martínez Fernández and Fernández-Fontecha (2008), who stressed the challenge of teaching machines to recognize context-sensitive humor and implicature. These findings are echoed in pragmatic studies of humor and conversational flouting, which suggest that maxims are routinely bent in social media discourse not as communicative failures, but as stylistic or rhetorical strategies (Olayemi & Avoaja, 2024). Human coders typically accounted for these subtleties in ways that Mistral, in its current prompt configuration, could not (Korre et al., 2024).

From a theoretical standpoint, this contrast reinforces Grice's original idea that maxims are not rules but guidelines, intentionally flouted to generate conversational implicatures (Grice, 1975). A human interpreter, armed with cultural awareness and pragmatic instinct, understands that marketing language often violates the maxim of Quantity (e.g., vague or inflated descriptions) or Manner (e.g., playful ambiguity) as a means to engage audiences. Mistral, in contrast, is more likely to treat surface-level coherence as a sign of compliance, revealing its reliance on lexical and syntactic cues over deeper pragmatic inference.

The model showed notable strengths in interpretive consistency and processing speed. Once properly prompted, Mistral delivered annotations for over 80 captions in under five minutes, an enormous improvement in efficiency compared to manual coding. It also offered brief, generally coherent justifications for its scores, demonstrating a degree of reflective explanation that could support human coders in preliminary analysis stages. These results mirror the outcomes of prior studies using LLMs for

NLP classification, which found that models can effectively triage data for human review (Gebreegziabher et al., 2023).

Another finding is that Mistral could classify illocutionary acts using Searle's typology, including mixed speech acts. This suggests that the model can understand overlapping communicative intentions, e.g., where a caption both informs and directs, even when not explicitly stated. However, further testing is needed to assess whether the model distinguishes between subtle performatives (e.g., expressive vs. commissive) as clearly as humans. Previous works such as Kong et al. (2023) have shown that automated discourse systems can approximate human speech act judgments but often require fine-tuning and contextual embedding to avoid overgeneralization.

In terms of feasibility, this study reinforces the promise of integrating LLMs into qualitative research workflows. The computational method was accessible, low-cost, and replicable. It used open-source tools and cloud-based deployment, making it viable for researchers with limited technical backgrounds. However, the findings also serve as a caution: while models like Mistral are powerful assistants, they are not replacements for human judgment. Their outputs must be interpreted critically, and their limitations recognized, particularly when analyzing culturally rich or informally structured discourse.

This study demonstrates that LLMs like Mistral 7B can effectively support pragmatic analysis, especially in structured and semi-structured content. However, they currently fall short of matching the interpretive depth, cultural sensitivity, and flexible judgment of human coders. Differences in observance scoring, failure with hybrid content, and constrained output range all point to areas requiring further refinement, particularly in prompt design and model tuning. By building on the model's strengths and addressing its limitations, future research can better integrate LLMs into qualitative methodologies, opening new pathways for scalable and rigorous discourse analysis.

CONCLUSION

This study explored the feasibility of using the Mistral 7B large language model to perform pragmatic analysis on a set of bilingual Instagram captions, primarily in English, with some in Indonesian, through the lens of Grice's conversational maxims. The model was tasked with classifying the observance of the maxims of Quality, Quantity, Relevance, and Manner, as well as identifying illocutionary acts and justifying its decisions. The findings reveal a compelling mix of strengths and limitations that shape the model's potential as a tool in qualitative linguistic research.

One of Mistral's most impressive features is its speed. The model was able to process and annotate over 80 captions, tasks that would require hours of manual work by a human coder, in a matter of minutes. This efficiency makes LLMs highly attractive for researchers who deal with large volumes of qualitative data, such as social media content or digital discourse. The fact that Mistral could handle English and Indonesian captions without language-specific reconfiguration further highlights its adaptability for multilingual analysis.

Mistral's accuracy and interpretive depth remain under development. The model consistently favored higher ratings, rarely assigning lower observance scores, even in captions that human coders judged as vague, exaggerated, or unclear. This indicates a lack of sensitivity to subtle violations, implicatures, or rhetorical strategies, an area where human judgment still excels. In addition, the model failed to analyze at least one syntactically complex caption, likely due to a parsing error caused by mixed formatting (narrative, emojis, and bulleted lists). These errors suggest that the model's success is still heavily dependent on input structure and prompt design.

Despite these limitations, the findings point to a promising future for LLMs in qualitative research. With improved prompts, better training examples, and perhaps fine-tuning for informal and digital language, models like Mistral could offer powerful support in early-stage analysis, data triage, or mixed-methods research. Rather than replacing human analysts, Mistral can complement them, offering speed and consistency where needed, and freeing up researchers to focus on interpretation, theory-building, and nuance. The integration of LLMs into qualitative workflows marks a meaningful advancement in how linguistic insight can be generated at scale.

FUNDING STATEMENT

This study is self-sponsored by the author. I state that there is no conflict of interest in the publication of this work.

REFERENCES

Aporbo, R. (2022). Decoding the linguistic functions of emoji: A discourse analysis of messenger chats. *International Journal of Research Publications, 111*(1), 1–16.

Augurzky, P., Franke, M., & Ulrich, R. (2019). Gricean expectations in online sentence comprehension: An ERP study on the processing of scalar inferences. *Cognitive Science*, *43*(8), e12776.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

Evans, J., D'Souza, J., & Auer, S. (2024). Large language models as evaluators for scientific synthesis. *arXiv preprint arXiv:2407.02977*.

Gebreegziabher, S. A., Zhang, Z., Tang, X., Meng, Y., Glassman, E. L., & Li, T. J. J. (2023, April). Patat: Human-ai collaborative qualitative coding with explainable interactive rule synthesis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-19).

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics* (Vol. 3, pp. 41–58). Academic Press.

Kong, A. P. H., Lau, D. K. Y., & Lai, D. H. Y. (2023). Measuring pragmatic competence of discourse output among Chinese-speaking individuals with traumatic brain injury. *Brain impairment*, *24*(3), 660-678.

Korre, K., Ruggeri, F., & Barrón-Cedeño, A. (2024, October). A Grice-ful examination of offensive language: Using NLP methods to assess the co-operative principle. In *Proceedings of the First LUHME Workshop* (pp. 12-19).

Krause, L., & Vossen, P. T. (2024, September). The gricean maxims in nlp-a survey. In *Proceedings of the 17th International Natural Language Generation Conference* (pp. 470-485).

Kreikemeyer, J. N., Jankowski, M., Wilsdorf, P., & Uhrmacher, A. M. (2025). Using (not so) large language models for generating simulation models in a formal DSL--A study on reaction networks. *arXiv preprint arXiv:2503.01675*.

Long, Y., Luo, H., & Zhang, Y. (2024). Evaluating large language models in analysing classroom dialogue. *npj Science of Learning*, *9*(1), 60.

Martínez Fernández, B., & Fernández Fontecha, A. (2008). The teachability of pragmatics in SLA: *Friends'* humour through Grice. *Porta Linguarum, 10*, 31–43.

Olayemi, O. T., & Avoaja, L. (2024). A pragmatic study of selected humorous discourse by Nigerian stand-up comedians. *NIU Journal of Humanities*, *9*(1), 233-241.

Pratama, H., Setiani, I., Amalia, W., & Putri, S. P. (2025). Cooperative principle maxims mechanisms in Samsung Indonesia Instagram reels' captions. *Journal of English Education Program*, *6*(2).

Randerson, S., Graydon-Guy, T., Lin, E. Y., & Casswell, S. (2025). Exploring the use of a large language model for inductive content analysis in a discourse network analysis study. *Social Science Computer Review*, 08944393251326175.

Vatsal, S., & Dubey, H. (2024). A survey of prompt engineering methods in large language models for different nlp tasks. *arXiv preprint arXiv:2407.12994*.