Can Large Language Models Handle Discourse Particles? A Case Study of Colloquial Malay

Anonymous ACL submission

Abstract

Discourse particles, such as *well* and *kind of*, are crucial components that enable LLMs to "speak" more like humans. They are used to convey emotions, intentions, and interpersonal meanings. However, existing studies have not yet built a comprehensive understanding of LLMs' capabilities in handling discourse particles. Moreover, the limited number of research focuses primarily on high-resource languages such as English, with little attention paid to Southeast Asian languages. In this paper, we (1) propose MALAYPRAG, a benchmark designed to systematically evaluate and analyze LLMs' capabilities in handling discourse particles in colloquial Malay; (2) introduce five attributes that provide a linguistically-grounded, unified framework for interpreting pragmatic functions of discourse particles. Applying these two, we prompt ten off-the-shelf LLMs to perform three predication tasks. The experimental results reveal substantial challenges for current LLMs to accurately connect discourse particles and their pragmatic functions in Malay. The provision of the five attributes designed in this study is found to significantly improve the connections, highlighting the need for structured scaffolding for models' pragmatic competence. *Our benchmark can be accessed via the link.*

1 Introduction

As the demand for more *human-like* communication in Large Language Models (LLMs) continues to grow, there is increasing interest in enabling LLMs to express phenomena such as hesitation, uncertainty, and nuanced sentiments. Consequently, discourse particles and their pragmatic functions[1] have emerged as an increasingly important research topic (Sheffield et al., 2025; Sadlier-Brown et al., 2024; Wang et al., 2025; Rocha et al., 2025; Ein-Dor et al., 2022). Discourse particles are unbound morphemes that encode implicit interpersonal dynamics, signaling multiple aspects of a speaker's intentions (Grzech, 2021). Typical examples of English particles include *well*, *like*, and *kind of*. They do not contribute to the literal meanings of an utterance but are, nevertheless, important in organization (e.g., pause, turn-taking) and meaning negotiations (e.g., mitigation, politeness). In addition, they index identities, regions, genders, and other social qualities (Schiffrin, 1987; Fraser, 1999).

Understanding the capabilities and mechanisms of LLMs in handling discourse particles is essential for the development of more human-like language models given the well-known pragmatic gaps in LLMs when interpreting meanings that arise through language use in context (Liu et al., 2025). Recent studies have sought to bridge these gaps by developing pragmatics-oriented datasets and benchmarks (Sravanthi et al., 2024; Ruis et al., 2023; Cong, 2024). However, these efforts have primarily focused on English and other high-resource languages, leaving low-resource Southeast Asian languages largely unexplored (Ma et al., 2025). To date, no prior work has investigated LLMs' understanding of discourse particles in Malay or its neighboring languages, despite their rich systems of discourse particles.

In the meanwhile, LLMs have been found to struggle to capture pragmatic functions of discourse particles (Sheffield et al., 2025), as each discourse particle tends to have an excessive range of functions that vary by context, for example, the phrase "how-to-say" in Chinese alone has 15 different functions (Chen and Ren, 2023). A potential solution to this issue is to explore the first-order epistemic, emotional, and linguistic attributes that underlie the varied pragmatic functions, hence translating the abstract notions of functions into

[1] In linguistics, pragmatic function refers to the communicative role or purpose that a linguistic expression serves in a particular context, especially in relation to speakers' intentions, social interaction, discourse structure, and inferred meaning.

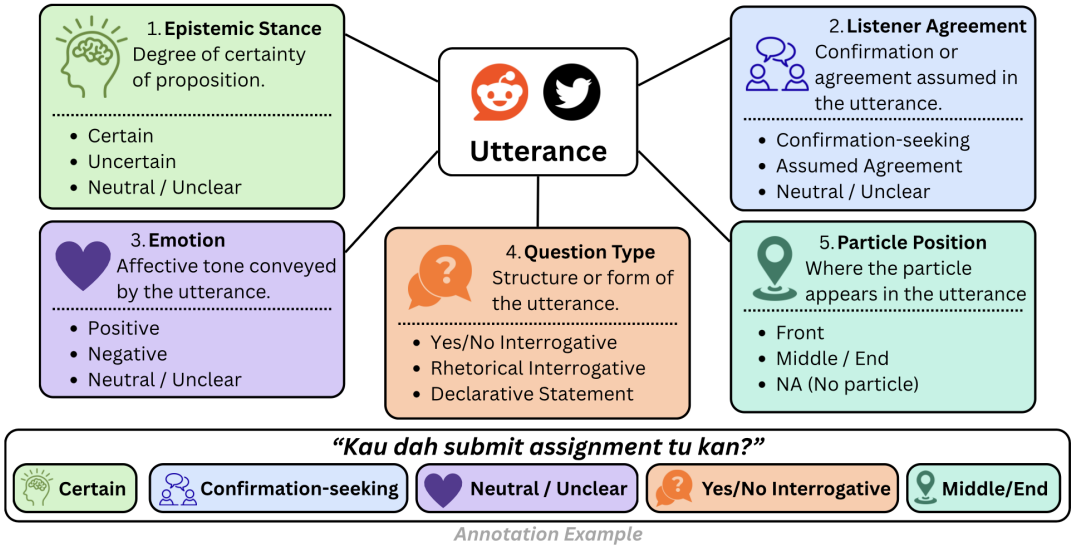

Figure 1: Five-dimensional annotation schema for Malay discourse particle utterances. An utterance is evaluated by a native Malay speaker according to each of the five dimensions and the most appropriate attribute is assigned to the utterance capturing the speaker's intentions and the utterance's syntax. The linguistic theoretical basis for each attribute is available in the Appendix 7.

discrete, computable variables (Hovy and Yang, 2021; Choi et al., 2023). However, to the best of the authors' knowledge, the solution has not yet been experimentally examined.

In this paper, we not only present a new Colloquial Malay dataset of discourse particles (MALAYPRAG), but also develop five attributes for evaluating their pragmatic functions in context. The attribute design is grounded solidly on linguistic theories. Figure 1 demonstrates the five attributes (*Epistemic Stance*, *Listener Agreement*, *Emotion*, *Question Type*, and *Particle Position*), and their classes for annotation.

We apply the dataset and attributes to two Malay discourse particles, *kan* and *ke* (commonly spelled as *ka* or *kah*), because they provide a theoretically meaningful contrast in their pragmatic functions (see Section 2 for their details). We conduct three prediction tasks to assess LLMs' capabilities of handling Malay discourse particles: attribute prediction, pragmatic function prediction, and discourse particle predication, each having a variety of subtasks and measured by model accuracy. We test ten off-the-shelf LLMs.

The experimental results show that **(1)** LLMs exhibit substantial difficulty in interpreting the pragmatic functions of Malay discourse particles and **(2)** performance improves when the five attributes are provided (Figure 1). Accordingly, the contributions of this study are threefold:

1. *MalayPrag: A Novel Low-Resource Benchmark.* We introduce a rigorously verified dataset for Colloquial Malay pragmatics, addressing the critical shortage of Southeast Asian representation in LLM evaluation.
2. *A Theory-Grounded, generalisable framework of five attributes for understanding discourse particle.* We propose five attributes as a unified framework that can be generalised for LLMs to learn pragmatic functions of discourse particles in Southeast Asian languages. *To the best of our knowledge, this is the first attribute framework that enables the unified modeling of pragmatic functions in real-world data.* The attribute-based design also reduces annotator bias and standardizes subjective pragmatic interpretation for computational modeling.
3. *A fine-grained and comprehensive understanding of LLMs' capability of handling discourse particles.* We provide a detailed comparison of how general-purpose and Southeast Asia-focused LLMs perform on Malay discourse particles in three tasks. Our results show that LLMs' pragmatic competence in low-resource Southeast Asian languages is uneven, attribute-sensitive, and can be improved by explicit pragmatic scaffolding.

2 Related Work

In Malay, *kan* and *ke* encode interactional meanings beyond propositional content. *Kan* marks

conjoint knowledge or requests listener agreement (Tay et al., 2016; Wouk, 1998), as in "Dia dah makan, **kan**?" ('He already ate, right?'), while *ke* marks interrogativity, uncertainty, confirmation-seeking, or rhetorical challenge, as in "Awak marah **ke**?" ('Are you mad?'). Since these particles vary across epistemic, interpersonal, affective, structural, and positional dimensions, we operationalize their pragmatic functions through five computationally measurable attributes: *Epistemic Stance*, *Listener Agreement*, *Emotion*, *Question Type*, and *Particle Position*. Among them, *Listener Agreement* is designed based on pragmatics studies of common ground and listener orientation (Wouk, 1998; Stalnaker, 2002); epistemic stance and speaker authority motivate the inclusion of *Epistemic Stance* (Kärkkäinen, 2003; Grzech, 2021); and work on affective meaning motivates *Emotion* as a cue to speaker attitude in interaction (Caffi and Janney, 1994; Buechel and Hahn, 2017). Table 7 summarizes the attribute settings, with full definitions in Appendix A.

Previous studies have found that LLMs struggle to interpret pragmatic functions of discourse particles (Sheffield et al., 2025; Sadlier-Brown et al., 2024; Ein-Dor et al., 2022). For example, Sheffield et al. (2025) demonstrate that LLMs fail to distinguish the overlapping senses of the English particle *just*; moreover, providing surrounding conversational context actively decreases model accuracy rather than aiding disambiguation.

However, it is worth noting that the aforementioned studies primarily used direct, coarse-grained classifications of pragmatic functions, with the hope that LLMs would learn the association between language data and annotated functions automatically. This approach, however, results in severe limitations, including low inter-rater reliability (IRR), data sparseness (De Felice et al., 2013), and models failing to grasp the underlying reasons why a particle was used. Our approach innovatively bridges the use of discourse particles and their pragmatic functions via the five attributes, by which LLMs' performance improves.

3 Methodology

This section details the construction of the MALAYPRAG dataset. We first outline the data collection and filtering process. Next, we describe the translation of linguistic theory into our annotation schema. Afterwards, we introduce how pragmatic functions were extracted from the data, and finally we elaborate the tasks that can be leveraged to understand LLMs' capabilities in handling discourse particles.

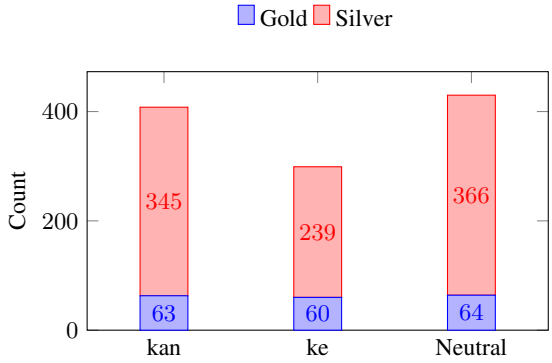

Figure 2: Distribution of Gold and Silver annotation splits across *kan*, *ke*, and neutral baseline utterances.

3.1 Data collection

Data was sourced from naturally occurring informal Malay utterances on Reddit and Twitter/X containing the particles *kan* and *ke*. To augment the dataset, we synthesised a sample of *kan* sentences to create (1) neutral sentences (by removing *kan*) and substituted utterances (replacing *kan* with *ke*). All synthesised sentences were validated by native speakers to ensure naturalness. The neutral sentences are used to test whether LLMs can distinguish the differences in pragmatic functions with/without the particle.

We used regular expressions to flag Indonesian phrasing, foreign-language interference, and prepositional uses of the homograph *ke* ('to'), followed by manual inspection for final data cleaning.

A total of 1,137 data was labeled by native Malay-speaking linguists trained in pragmatic analysis. The final dataset was bifurcated into a GOLD and SILVER split for different modelling needs. The GOLD dataset ($N = 187$) consists of utterances independently annotated by three annotators, with disagreements resolved via majority voting or lead-author adjudication. This set was used as the ground-truth for LLM evaluation. The SILVER dataset ($N = 950$) comprises utterances labeled only by a single trained annotator, and it was used for the extraction of pragmatic functions. Figure 2 illustrates the data split.

3.2 Attribute annotation

Most existing discourse annotation schemes are tailored to English and high-resource languages,

making direct porting ineffective for low-resource languages (Vargas et al., 2025). Furthermore, applying generic categories without explicit, well-documented definitions leads to subjective bias and high annotator disagreement (Crible and Zufferey, 2015). Therefore, we develop our annotation schemes based on existing linguistic research of Southeast Asian discourse particles and commonly agreed empirical findings across languages (see Section 2 and also Appendix A).

To be specific, within *Epistemic Stance*, there are three options: Certain, Uncertain, and Neutral. A *Certain* classification denotes that the speaker holds full epistemic authority, presenting the proposition as factual or unquestionable, whereas *Uncertain* reflects speculation, doubt, or an active hypothesis. For *Listener Agreement*, *Assumed Agreement* dictates that the proposition is framed as pre-existing common ground; the speaker expects the listener to simply align or concede, treating the utterance as a rhetorical check rather than a genuine inquiry. Conversely, *Confirmation Seeking* indicates a genuine request for verification where the speaker actively leaves room for the listener to contest, correct, or reject the premise. In contrast, structurally objective variables such as *Particle Position* (*Front, Middle/End, N/A*) and *Question Type* were strictly defined by their syntactic markers and require minimal interpretation.

3.3 Extracting pragmatic functions from attribute clustering

As emphasized above, we do not directly annotate pragmatic functions for LLMs to predict, which has been evidenced to be ineffective in previous studies (De Felice et al., 2013). Instead, we used the five attributes annotated to obtain pragmatic functions. Specifically, we cluster the annotated attributes using K-means on both GOLD ($N = 187$) and SILVER ($N = 950$) datasets (total $N = 1{,}137$). The reason for both datasets to be involved in this process is to take into account both collective understanding of the discourse particles – that are represented by the agreements in GOLD dataset – and individual variations represented by the SILVER dataset.

Spatially close clusters of the attributes are expected to represent the same or similar pragmatic functions of a discourse particle, while distanced clusters represent distinctive pragmatic functions. We should emphasise that the clustering results do not reveal the pragmatic functions automatically.

Prediction	Input	Prompting	+ CoT
T1 Attribute	Raw data	✓	
T2 Pragmatic Function	Raw data	✓	✓
	+ Attributes	✓	
T3 Particle	Raw data	✓	
	+ Attributes	✓	
	+ Pragmatic functions	✓	✓
	+ Attr. + prag. func.	✓	

Table 1: Evaluation tasks and conditions. We prompt ten off-the-shelf LLMs to perform three predication tasks: attribute prediction, pragmatic function prediction, and particle prediction on the MALAYPRAG benchmark.

Rather, following the K=16 clustering, human linguists conducted a qualitative review of the utterances within the 16 clusters to inductively assign discrete, overarching "pragmatic function" labels, thereby establishing a data-driven taxonomy for the subsequent evaluations of LLMs.

To illustrate the correlations between pragmatic functions and clustering of five attributes, for example, the pragmatic function "Information-seeking verification" is often represented by the cluster that contains an uncertain epistemic stance, a confirmation-seeking listener agreement, neutral emotion, and an interrogative structure. This specific constellation of attributes indicates that the discourse particle (predominantly *ke*) is functioning as a genuine request for clarification where the speaker actively leaves room for disagreement. Conversely, the "Negative rhetorical challenge" function emerges from a sharply contrasting attribute pattern: a certain epistemic stance, assumed listener agreement, a rhetorical question structure, and strong negative affect. In this context, the underlying attributes dictate that the particle functions not as an inquiry, but as a discourse device to criticize, mock, or forcefully evaluate a proposition. Appendix B presents representative examples of these derived pragmatic functions.

3.4 Evaluation Tasks

To systematically evaluate LLMs' capability to handle the pragmatics of Malay discourse particles, we conduct three evaluation tasks: attribute prediction, pragmatic function prediction, and discourse particle prediction. Table 1 outlines each task and their conditions.

Task 1: Attribute Prediction. This task evaluates whether LLMs can predict the five attributes

from raw data.

Task 2: Pragmatic Function Prediction. This task comprises three subtasks. In **(2a) Function prediction from raw data:** The model predicts predicts the overarching pragmatic function from the utterance alone. In **(2b) Function prediction via CoT:** The model predicts predicts pragmatic function given utterance and Chain-of-Thought (CoT) (Wei et al., 2022). In **(2c) Function prediction via attributes:** The model predicts pragmatic functions given an utterance and its human-annotated attributes. Comparing the three subtasks allows us to confirm the effectiveness of attributes in bridging between discourse particles and their pragmatic functions, especially compared to CoT as a strong baseline.

Task 3: Discourse Particle Prediction. This task evaluates whether LLMs can select the appropriate particle, *kan*, *ke*, or neutral (no particle), for a masked utterance. It includes five prompting conditions: **(3a) Particle prediction from raw data**, using only the masked utterance; **(3b) Attribute-provided prediction**, adding human-annotated attributes; **(3c) Function-provided prediction**, adding the target pragmatic function; **(3d) Particle prediction with attribute + function provided**, adding both; and **(3e) Particle prediction with CoT + function**, adding pragmatic functions with CoT prompts. Comparing these conditions tests whether explicit pragmatic scaffolding improves particle selection.

4 Experimental Setting and Results

This section reports the experimental settings and empirical findings for the tasks above.

Models. We evaluate ten off-the-shelf LLMs that span varying parameter scales, training paradigms, and regional specialisation. Eight are general-purpose frontier models accessed via their official APIs: GPT-5 and GPT-5.4-mini (OpenAI), Claude Sonnet 4.6 and Claude Haiku 4.5 (Anthropic), Gemini 3.1 Pro and Gemini 3.1 Flash (Google), and DeepSeek-v4-Pro and DeepSeek-v4-Flash (DeepSeek). To examine whether regional training affects pragmatic competence in Malay, we additionally include two open-weight Southeast Asia-focused models from the SEA-LION family (Ng et al., 2025; Koh et al., 2025): Llama-SEA-LION-70B and Gemma-SEA-LION-27B.

Evaluation data. All tasks are conducted on the benchmark MALAYPRAG ($N = 187$), in which every utterance has been independently annotated by three trained native Malay linguists, with disagreements resolved through majority voting or lead-author adjudication.

Prompting and inference. All tasks are run in a zero-shot setting, besides the CoT baselines in ablation experiments. To minimize sampling variance, we set temperature to 0 (greedy decoding) for all models and constrain outputs to a single label from the closed set specified by each prompt template. The full prompting templates for every task are provided in Appendix C.

Evaluation metric. Following prior work on pragmatic benchmarking (Sravanthi et al., 2024; Sheffield et al., 2025), we report classification accuracy against the benchmark MALAYPRAG. For attribute prediction Task, accuracy is computed per attribute and then averaged across the five attributes. For pragmatic function prediction tasks, accuracy is computed over the seven pragmatic-function labels; for particle prediction Tasks, accuracy is computed over the three particle choices (*kan*, *ke*, neutral). The corresponding random-chance baselines are approximately 33% or 50% for attribute prediction, 14.3% for pragmatic-function prediction, and 33.3% for particle generation.

4.1 Task 1: Attribute Prediction

We use zero-shot English and Malay prompts, separately, to test whether the selected models can accurately predict the five attributes annotated, namely, *Epistemic Stance*, *Listener Agreement*, *Emotion*, *Question Type*, and *Particle Position*. To reiterate, these five attributes underlie our identification of pragmatic functions (see Section 3.3).

As shown in Table 2, **model performance on English (EN) prompts consistently outpaces performance on native Malay (MS) prompts across all models.** The overall average for English prompts was 69.26%, compared with 65.16% for Malay prompts. The consistent deficient performance in Malay may be an indication of imbalanced data size and semantic distributions learned during model training. That is, current models may possess Malay vocabulary for processing the task, but the highly technical, meta-linguistic reasoning required to evaluate concepts such as "epistemic stance" is overwhelmingly concentrated in English data.

Model	Epistemic		Agreement		Emotion		Question		Position		Overall Avg.	
	EN	MS	EN	MS	EN	MS	EN	MS	EN	MS	EN	MS
GPT-5	0.706	0.765	0.594	0.524	0.797	0.738	0.695	0.711	0.901	0.884	0.742	0.724
GPT-5.4-mini	0.583	0.449	0.631	0.519	0.797	0.770	0.604	0.626	0.752	0.777	0.700	0.628
Claude Sonnet 4.6	0.749	0.733	0.588	0.449	0.754	0.631	0.711	0.524	0.942	0.883	0.752	0.644
Claude Haiku 4.5	0.749	0.636	0.578	0.449	0.701	0.684	0.679	0.679	0.860	0.771	0.722	0.644
Gemini 3.1 Pro	0.759	0.802	0.476	0.433	0.743	0.727	0.711	0.668	0.818	0.826	0.721	0.691
Gemini 3.1 Flash	0.759	0.802	0.476	0.449	0.743	0.727	0.711	0.684	0.818	0.843	0.721	0.701
DeepSeek-v4-Pro	0.615	0.668	0.465	0.476	0.749	0.727	0.695	0.674	0.942	0.934	0.696	0.696
DeepSeek-v4-Flash	0.706	0.722	0.476	0.438	0.754	0.754	0.679	0.690	0.942	0.884	0.713	0.698
Llama-SEA-LION-70B	0.348	0.433	0.278	0.433	0.540	0.369	0.406	0.417	0.868	0.554	0.499	0.441
Gemma-SEA-LION-27B	0.717	0.679	0.524	0.529	0.690	0.711	0.615	0.599	0.612	0.722	0.660	0.648
Average	0.669	0.669	0.509	0.470	0.727	0.684	0.651	0.627	0.845	0.808	0.693	0.652

Table 2: Attribute Prediction accuracy for English (EN) and Malay (MS) prompts. Position achieves the highest accuracy while Listener Agreement is the lowest.

Among the attributes, Particle Position achieved the highest accuracy (EN: 84.54%, MS: 80.78%), confirming that models can execute relatively objective structural and spatial parsing. Conversely, Listener Agreement yielded the poorest performance across the board (EN: 47.0%, MS: 50.85%), frequently operating at or below a random chance. Intriguingly, this finding is consistent with human annotators' perceptions of Listener Agreement: This attribute also elicited the highest degree of disagreement among our human annotators during the annotation process. Thus, we argue that **the lower accuracy in models' predication of *Listener Agreement* reflects an inherent linguistic ambiguity.** That is, in spoken Malay, assessing common knowledge and listener-oriented assumptions relies heavily on prosody, intonation, and shared conversational history (Tay et al., 2016; Wong, 2004). In the text-only vacuum of social media, where these multi-modal and multi-turn cues are stripped away, modeling perception regarding the listener becomes more challenging for both human linguists and computational models.

4.2 Task 2: Pragmatic Function Prediction

For the ease of identifying change in model performance, we first compare pragmatic functions predicted with and without attributes provided. Then, we compare the effectiveness of attributes to that of CoT in function prediction.

As Table 3 displays, **the provision of attributes significantly improves models' accuracy in predicting pragmatic functions.** Models in (2a) (predicting pragmatic functions without attributes) struggle significantly to predict pragmatic functions, averaging only 27.96% accuracy. This result is only slightly more accurate than random guessing. However, when scaffolded with the annotated attributes in (2c) (with attributes provided), model performance increased sharply: average accuracy rose to 52.46%, producing an average diagnostic delta of +24.49%.

Model	Direct Prompting	+ Attribute	Delta
GPT-5	.300	.529	.230
GPT-5.4-mini	.294	.503	.209
Claude Sonnet 4.6	.257	.524	.267
Claude Haiku 4.5	.283	.524	.241
Gemini 3.1 Pro	.305	.519	.214
Gemini 3.1 Flash	.337	.524	.187
DeepSeek-v4-Pro	.316	.588	.273
DeepSeek-v4-Flash	.353	.546	.193
Llama-SEA-LION-70B	.123	.401	.278
Gemma-SEA-LION-27B	.230	.588	.358
Average	.280	.525	.245

Table 3: Pragmatic Function prediction accuracy across models under two conditions: direct prompting (without context) and prompting with attributes. All scores are accuracy; Delta denotes the absolute improvement obtained by incorporating attribute information, computed as Attribute minus Direct Prompting. Overall, providing attributes substantially improves pragmatic function prediction across all models.

Similarly, Table 4 shows that **the provision of attributes outperforms the incorporation of CoT**

Model	CoT	Attributes	Delta (+/-)
GPT-5	.396	**.529**	.133
GPT-5.4-mini	.316	**.503**	.187
Claude Sonnet 4.6	.321	**.524**	.267
Claude Haiku 4.5	.348	**.524**	.241
Gemini 3.1 Pro	.348	**.519**	.214
Gemini 3.1 Flash	.342	**.524**	.187
DeepSeek-v4-Pro	.326	**.588**	.272
DeepSeek-v4-Flash	.364	**.546**	.193
Llama-SEA-LION-70B	.219	**.401**	.182
Gemma-SEA-LION-27B	.262	**.588**	.358
Average	.324	**.525**	.201

Table 4: Pragmatic function prediction accuracy across models under two prompting conditions: chain-of-thought prompting (CoT) and attribute-enhanced input (Attributes). Each model assigns one of seven pragmatic function labels to a sentence. Delta is computed as Attribute minus CoT. Overall, explicit attributes outperform chain-of-thought prompting, suggesting that structured pragmatic cues are more useful than elicited reasoning alone.

Model	Direct & Prompting	Attr.	Func.	Attr. & Func.
GPT-5	.513	.690	.674	.690
GPT-5.4-mini	.385	.604	.497	.690
Claude Sonnet 4.6	.471	.578	.546	.706
Claude Haiku 4.5	.471	.685	.497	.711
Gemini 3.1 Pro	.428	.631	.524	.690
Gemini 3.1 Flash	.433	.631	.519	.679
DeepSeek-v4-Pro	.476	.685	.562	.754
DeepSeek-v4-Flash	.449	.668	.588	.717
Llama-SEA-LION-70B	.332	.471	.417	.711
Gemma-SEA-LION-27B	.337	.503	.487	.562
Average	.430	.615	.531	.691

Table 5: Masked-slot particle prediction accuracy (ke, kan, or neutral) under four prompting conditions: direct prompting without context, attribute-only, pragmatic function-only, and combined attribute and pragmatic function context. Overall, the strongest performance comes from combining attribute and pragmatic function information.

in pragmatic function prediction. Applying CoT yields an average accuracy of 32.4%, which represents only a marginal improvement over the baseline prediction from raw data (27.9%). The results remains drastically inferior to Attribute-provided Function Prediction performance (52.5%), with our attribute framework presenting a significant delta of 20.1%.

Interestingly, while global models such as GPT-5 and Claude Sonnet 4.6 saw reliable gains, Gemma-SEA-LION-27B exhibited the most extreme improvement, with delta over 30%. With the provision of the attributes, Gemma-SEA-LION-27B even outperformed GPT-5 (52.94%) and recorded the highest pragmatic function prediction score alongside DeepSeek-v4-Pro.

4.3 Task 3: Discourse Particle Prediction

Task 3 contains five sub-tasks. For the ease of comparison, we again divide them into two tables: Table 5 compares which of attributes or pragmatic functions is more effective in predicting discourse particles, while Table 6 compares attributes to CoT.

As shown in Table 5, when perdicting from raw data, models average 43.0% accuracy, which is the lowest across the three tasks. Providing pragmatic functions (3c) raises average accuracy to 53.1%. However, providing models with the attributes (3b), instead of pragmatic functions, yields an even higher accuracy, averaging 61.5%. Providing both attributes and functions (2d) results in the highest average accuracy of 69.1%. The findings corroborate our argument above that pragmatic functions alone are insufficient for LLMs to interpret discourse particles in low-resource languages.

Interestingly, **CoT becomes more effective in predicting discourse particles than in predicting pragmatic functions in Task 2.** As shown in Table 6, CoT is combined with pragmatic functions and compared to the combination of pragmatic function + attributes. It achieves similar accuracy, although attributes + functions still outperform by a marginal delta.

This finding aligns with previous studies that find CoT is less effective in pragmatic reasoning tasks, but better at capturing semantic connections (Chen and Wang, 2025; Liu et al., 2025; Chen et al., 2026). In other words, CoT performing strongly in predicting discourse particles may be because the discourse particles fall in the semantic distribution of CoT steps. Pragmatic functions, on the other hand, are usually the "unsaid" effects created by utterances, which CoT can hardly reason. In both Tasks 2 and 3, however, our design of attributes appears to be the strongest scaffolding for LLMs to achieve better prediction accuracy.

5 Discussion

The Superiority of Attribute-Based Understanding of Discourse Particles. Across different prediction tasks, attributes have consistently been ef-

Model	CoT & Function	Attribute & Function	Delta (+/-)
GPT-5	.679	**.690**	.011
GPT-5.4-mini	.679	**.690**	.011
Claude Sonnet 4.6	**.743**	.706	-.037
Claude Haiku 4.5	.679	**.711**	.032
Gemini 3.1 Pro	.674	**.690**	.016
Gemini 3.1 Flash	.674	**.679**	.005
DeepSeek-v4-Pro	.738	**.754**	.016
DeepSeek-v4-Flash	.706	**.717**	.011
Llama-SEA-LION-70B	.658	**.711**	.053
Gemma-SEA-LION-27B	**.658**	.562	-.096
Average	.689	**.691**	.002

Table 6: Masked-slot particle prediction accuracy under two conditions: chain-of-thought with pragmatic function context (CoT & Function) and joint attribute plus pragmatic function context (Attribute & Function). Models predict ke, kan, or neutral. Delta is computed as Attribute & Function minus CoT Function. Overall, adding attribute context to pragmatic function context yields a small average improvement over CoT with pragmatic function context.

fective in improving model performance on discourse particles. The findings corroborated previous linguistic insights into how they underlie the great variety of pragmatic functions (Wouk, 1998; Stalnaker, 2002; Kärkkäinen, 2003; Caffi and Janney, 1994). Considering that LLMs struggle to learn pragmatic functions directly from language data and the function annotations alone, the current five attributes as a framework have the potential to become the common "bridge" that connects LLMs' generation of discourse particles to recognition of their pragmatic functions, especially in low-resource, Southeast Asian languages.

Regional LLMs' paradox Compared to other general-purpose LLMs, the SEA-LION models, which are trained specifically for Southeast Asian languages, are rather inconsistent in performance. Recall that Gemma-SEA-LION-27B improved the most in (2c) pragmatic function prediction with attributes provided. In Task 3, both Llama- and Gemma-based SEA-LION models are much less accurate than other models. We argue that the gap may lie in the pre- and post-training of the SEA-LION models. As Yu et al. (2026) finds, models' sensitivity to pragmatic cues increases consistently with model and data scale, and post-training further consolidates the gains of pragmatic knowledge. Global models receive more training in both stages than SEA-LION models.

However, SEA-LION models utilize specialized tokenizers relevant to Southeast Asian languages and are pre-trained on billions of such tokens (Ng et al., 2025; Koh et al., 2025). These efforts seem to have paid off in the success of the smaller SEA-LION models in leveraging the five attributes and connecting them with pragmatic functions in Malay. In other words, regional LLMs, albeit falling short in model scale and post-training, may be equipped with latent cultural and pragmatic knowledge. By supplying explicit pragmatic scaffolding like our five attributes, the regional datasets can effectively transform raw lexical exposure into actionable, culturally accurate reasoning.

Comparing with other Evaluation Benchmarks. When comparing our findings to existing pragmatic evaluations conducted in English, a stark disparity emerges regarding the impact of model scale. Recent studies have demonstrated that even smaller LLMs can achieve moderate to high baseline accuracy in English discourse particle prediction tasks (Sheffield et al., 2025; Ruis et al., 2023). In contrast, the state-of-the-art frontier models used in the current study still exhibited severe performance degradation in Colloquial Malay. The findings further emphasise the need for theoretically sound and computationally efficient methods, like our attribute design, to overcome the imbalance in the accessibility of language data.

6 Conclusion and Future Work

In this paper, we introduced a new Colloquial Malay dataset, with five attributes designed as its evaluation metrics and for predicting discourse particles. Our findings demonstrate that, without the attributes, even the state-of-the-art models fall seriously short in accurately predicting pragmatic functions of discourse particles. Providing the structured, attribute-level grounding drastically improves model performance, outperforming both zero-shot baselines and CoT as well as other conditions.

Although our work only showcased the design on *kan* and *ke* in Colloquial Malay, the Southeast Asian linguistic landscape as a whole is rich with highly polysemous particles. We encourage future work to make further expansion to encompass a wider array of particles and investigate multi-turn conversational contexts, using the attribute-based framework.

Limitations

This study concerns primarily textual data, with the absence of prosodic cues, which may play an important role in annotating the attributes. As much as discourse particles are frequently used in colloquial language, incorporating prosodic features may further enhance the consistency of attribute annotations and, more importantly, enable tests on multimodal LLMs. At the moment, the current study relied on prompting to evaluate LLMs' capability to interpret discourse particles. Whether other methods, such as fine-tuning, may further facilitate LLMs' acquisition of pragmatic functions over attributes and enhance their human-like discourse particle usage is yet to be explored.

References

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Claudia Caffi and Richard W. Janney. 1994. Towards a pragmatics of emotive communication. *Journal of Pragmatics*, 22(3–4):325–373.

Bocheng Chen, Han Zi, Xi Chen, Xitong Zhang, Kristen Johnson, and Guangliang Liu. 2026. Learning to diagnose and correct moral errors: Towards enhancing moral sensitivity in large language models. *arXiv preprint arXiv:2601.03079*.

Xi Chen and Wei Ren. 2023. Functions, sociocultural explanations and conversational influence of discourse markers: focus on zenme shuo ne in L2 Chinese. *International Review of Applied Linguistics in Language Teaching*. Publisher: De Gruyter Mouton.

Xi Chen and Shuo Wang. 2025. Pragmatic inference chain (PIC) improving LLMs' reasoning of authentic implicit toxic language. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5826–5841, Suzhou, China. Association for Computational Linguistics.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do llms understand social knowledge? evaluating the sociability of large language models with socket benchmark. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 11370–11403.

Yan Cong. 2024. Manner implicatures in large language models. *Scientific Reports*, 14(1):29113.

Ludivine Crible and Sandrine Zufferey. 2015. Using a unified taxonomy to annotate discourse markers in speech and writing. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, London, UK. Association for Computational Linguistics.

Rachele De Felice, Jeannique Darby, Anthony Fisher, and David Peplow. 2013. A classification scheme for annotating speech acts in a business email corpus. *ICAME Journal*, 37:71–105.

L. Ein-Dor, I. Shnayderman, A. Spector, L. Dankin, R. Aharonov, and N. Slonim. 2022. Fortunately, discourse markers can enhance language models for sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10608–10617.

Bruce Fraser. 1999. What are discourse markers? *Journal of pragmatics*, 31(7):931–952.

Karolina Grzech. 2021. Using discourse markers to negotiate epistemic stance: A view from situated language use. *Journal of Pragmatics*, 177:208–223.

Tom G. Hoogervorst. 2018. Utterance-final particles in klang valley malay. *Wacana, Journal of the Humanities of Indonesia*, 19(2):292–325.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 588–602.

Lionel Koh, Jirat Boomuang, Pyn Lim, and Francis Wang. 2025. Mitigating bias, ensuring fairness: Sea-lion's strategies for inclusive llm in a diverse region. *Ensuring Fairness: SEA-LION's Strategies for Inclusive LLM in a Diverse Region (February 07, 2025)*.

Elise Kärkkäinen. 2003. *Epistemic stance in English conversation: A description of its interactional functions, with a focus on I think*. John Benjamins Publishing Company.

Guangliang Liu, Xi Chen, Bocheng Chen, Xitong Zhang, and Kristen Johnson. 2025. Pragmatic Inference for Moral Reasoning Acquisition: Generalization via Distributional Semantics. *arXiv preprint*. ArXiv:2509.24102 [cs].

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. *arXiv preprint arXiv:2502.12378*.

Raymond Ng, Thanh Ngan Nguyen, Huang Yuli, Tai Ngee Chia, Leong Wai Yi, Wei Qi Leong, Xianbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng, and 1 others. 2025. Sea-lion: Southeast asian languages in one network. In *Proceedings*

of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 512–526.

G. Rocha, H. Lopes Cardoso, J. Belouadi, and S. Eger. 2025. Cross-genre argument mining: Can language models automatically fill in missing discourse markers? *Argument & Computation*, 16(1):3–35.

Laura Eline Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. Large language models are not zero-shot communicators.

Emily Sadlier-Brown, Millie Lou, Miikka Silfverberg, and Carla Kam. 2024. How useful is context, actually? comparing llms and humans on discourse marker prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 231–241, Bangkok, Thailand. Association for Computational Linguistics.

Deborah Schiffrin. 1987. *Discourse markers*. 5. Cambridge University Press.

William Sheffield, Kanishka Misra, Valentina Pyatkin, Ashwini Deo, Kyle Mahowald, and Junyi Jessy Li. 2025. Is it just semantics? a case study of discourse particle understanding in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21704–21715, Vienna, Austria. Association for Computational Linguistics.

Settaluri Lakshmi Sravanthi, Meet Doshi, Tankala Pavan Kalyan, Pushpak Bhattacharyya, Rudra Murthy, and Raj Dabre. 2024. Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities. *arXiv preprint arXiv:2401.07078*.

Robert Stalnaker. 2002. Common ground. *Linguistics and Philosophy*, 25:701–721.

Li Chia Tay, Mei Yuit Chan, Ngee Thai Yap, and Bee Eng Wong. 2016. Discourse particles in malaysian english: What do they mean? *Bijdragen tot de Taal-, Land- en Volkenkunde*, 172(4):479–509.

Francielle Vargas, Wolfgang Schmeisser-Nieto, Zohar Rabinovich, Thiago A. S. Pardo, and Fabrício Benevenuto. 2025. Discourse annotation guideline for low-resource languages. *Natural Language Processing*, 31:700–743.

Sheng-Fu Wang, Ri-Sheng Huang, Shu-Kai Hsieh, and Laurent Prévot. 2025. Zero-shot evaluation of conversational language competence in data-efficient llms across english, mandarin, and french. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 32–47, Avignon, France. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jock Wong. 2004. The particles of singapore english: a semantic and cultural interpretation. *Journal of Pragmatics*, 36:739–793.

Fay Wouk. 1998. Solidarity in indonesian conversation: The discourse marker kan. *Multilingua*, 17(4):379–406.

Kefan Yu, Qingcheng Zeng, Weihao Xuan, Wanxin Li, Jingyi Wu, and Rob Voigt. 2026. The pragmatic mind of machines: Tracing the emergence of pragmatic competence in large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 192–213, Rabat, Morocco. Association for Computational Linguistics.

A Flattened Codebook Definitions

This appendix provides the annotation codebook used by human annotators. Each attribute was annotated using a guiding question and a set of discrete tag options.

A.1 Epistemic Stance

Guiding question: How sure does the speaker sound about the information?

Tag	One-sentence rule
Certain	Pick this if the speaker treats the statement as already true.
Uncertain	Pick this if the speaker sounds unsure, guessing, or checking.
Neutral / Unclear	Pick this if you cannot tell whether the speaker is certain or not.

A.2 Listener Agreement

Guiding question: What kind of response does the speaker seem to want?

Tag	One-sentence rule
Confirmation Seeking	The speaker genuinely expects a direct "Yes" or "No" answer.
Assumed Agreement	Pick this if the speaker expects or leans toward agreement.
Neutral / Unclear	The speaker is not looking for any agreement or response from the listener.

Attribute	Classes	Theoretical Basis
Epistemic Stance	Certain; Uncertain; Neutral	Kärkkäinen (2003); Grzech (2021)
Listener Agreement	Confirmation Seeking; Assumed Agreement; Neutral	Wouk (1998); Stalnaker (2002)
Emotion	Positive; Negative; Neutral	Caffi and Janney (1994); Buechel and Hahn (2017)
Question Type	Declarative; Yes/No Interrogative; Rhetorical	Hoogervorst (2018)
Particle Position	Front; Middle/End; N/A	Tay et al. (2016)

Table 7: Five attributes for annotating our Malay discourse particle utterances.

A.3 Emotion / Affect

Guiding question: What is the underlying "vibe" or emotional payload? Choose the single most dominant emotion.

Tag	One-sentence rule
Positive	Pick this if the tone feels positive or approving.
Negative	Pick this if the tone feels critical, annoyed, or negative.
Neutral / Unclear	Pick this if there is little or no emotional signal.

A.4 Question Type

Guiding question: Is this a real question, a rhetorical one, or a statement?

Tag	One-sentence rule
Yes/No Interrogative	Pick this if the speaker is genuinely asking for an answer.
Rhetorical Interrogative	Pick this if it looks like a question but is making a point.
Declarative / Statement	Pick this if it is mainly stating something, not asking.
Unclear	Pick this if the function is ambiguous.

A.5 Particle Position

Guiding question: Where is the particle located?

Tag	One-sentence rule
Front	Pick this if the particle appears at the start of the sentence.
Middle/End	Pick this if the particle appears anywhere else.
N/A	Pick this if no particle is present.

B Pragmatic Function Labels

This appendix lists the seven pragmatic function labels derived from the clustering analysis.

Assumed-Agreement Rhetorical Stance
Speaker presents the proposition as already obvious or shared knowledge; the listener is expected to align rather than genuinely answer.

Neutral Declarative
Plain informational statements with minimal discourse pressure or stance marking.

Information-Seeking Verification
Genuine request for verification or clarification; the speaker leaves room for disagreement.

Affective Confirmation-Seeking Question
Speaker seeks confirmation while simultaneously expressing affect, such as surprise, irritation, humor, excitement, or disbelief.

Emphatic / Discourse-Marking
Particle functions less as a literal confirmation marker and more as a discourse-management or emphasis device.

Null Form Retaining Particle-Like Pragmatic Meaning
Pragmatic meaning associated with particles remains inferable even after overt particle removal.

Negative Rhetorical Challenge / Evaluation
Speaker uses rhetorical questioning to criticize, challenge, mock, or negatively evaluate a proposition rather than genuinely seek information.

C Evaluation Prompt Templates

C.1 Benchmark Prompt Examples

C.1.1 Task 1a: Attribute Prediction

Attribute 1: Epistemic Stance

```
You are a linguist specializing in colloquial
Malay.  Your task is to read the Malay
sentence below and decide how certain the
speaker sounds about the information they are
conveying.
Referring to the following three labels and
their definitions to make your decision:
Certain:  The speaker treats the statement
as already true or established.  There is
no hedging, no doubt, and no checking.  The
speaker is asserting the information with full
confidence.
Uncertain:  The speaker sounds unsure, is
```

making a guess, is estimating, or is checking whether something is the case. Words like "agaknya" (I think/probably), "kot" (maybe), or question particles that probe for confirmation are typical signals.
Neutral/NA: The sentence does not carry any detectable certainty signal in either direction. This applies to neutral descriptions, commands, or sentences where certainty is simply not relevant.
Speaker: "{TEXT}"
Given the three labels "Certain, Uncertain, Neutral/NA", the most likely, single label of the speaker's utterance is:

Attribute 2: Particle Position

You are a linguist specializing in colloquial Malay. Your task is to read the Malay sentence below and decide where the discourse particle appears in it.
Referring to the following three labels and their definitions to make your decision:
Front: The particle appears at the very start of the sentence, before any other content words.
Middle/End: The particle appears anywhere other than the front – mid-sentence, before the final word, or at the end.
N/A: No discourse particle is present in the sentence (e.g. the particle slot is shown as "[___]" or the sentence simply contains no particle).
Speaker: "TEXT"
Given the three labels "Front, Middle/End, N/A", the most likely, single label of the speaker's utterance is:

Attribute 3: Listener Agreement

You are a linguist specializing in colloquial Malay. Your task is to read the Malay sentence below and decide how the speaker is orienting toward the listener in terms of shared knowledge or agreement.
Referring to the following three labels and their definitions to make your decision:
Assumed Agreement: The speaker treats the information as already shared or obvious to the listener. The sentence is presented as common ground – the underlying tone is "you already know this" or "of course this is true". No explicit confirmation is being requested.
Confirmation Seeking: The speaker is actively checking whether the listener agrees, knows, or can confirm the information. The sentence invites or requests the listener's validation before the speaker can proceed with confidence.
Neutral/Unclear: The sentence does not show any clear orientation toward listener agreement. This applies to plain statements, commands, or cases where the interpersonal stance toward agreement is ambiguous or absent.
Speaker: "TEXT"
Given the three labels "Assumed Agreement, Confirmation Seeking, Neutral/Unclear", the most likely, single label of the speaker's utterance is:

Attribute 4: Emotion

You are a linguist specializing in colloquial Malay. Your task is to read the Malay sentence below and decide the emotional tone of the speaker.
Referring to the following three labels and their definitions to make your decision:
Positive: The speaker expresses happiness, excitement, enthusiasm, satisfaction, humor, affection, relief, or any other clearly positive feeling. This includes light-hearted teasing or playful sarcasm that is warm in tone.
Negative: The speaker expresses frustration, annoyance, disappointment, sadness, anger, bitterness, or any other clearly negative feeling. This includes hostile or bitter sarcasm.
Neutral/Unclear: The sentence carries no detectable emotional charge in either direction, or the emotion is genuinely ambiguous and cannot be reliably classified as positive or negative.
Speaker: "TEXT"
Given the three labels "Positive, Negative, Neutral/Unclear", the most likely, single label of the speaker's utterance is:

Attribute 5: Question Type

You are a linguist specializing in colloquial Malay. Your task is to read the Malay sentence below and decide its primary sentence function.
Referring to the following three labels and their definitions to make your decision:
Declarative/Statement: The sentence makes an assertion or conveys information. It describes a situation, states a fact, or expresses a view. It is not structured as a question, even if it ends with a particle.
Rhetorical Interrogative: The sentence is phrased as a question but does not expect a genuine answer from the listener. It is used to make a point, express emotion, or emphasize something – the speaker already implies the answer through the question itself.
Yes/No Interrogative: The sentence is a genuine question that invites the listener to confirm or deny something. The speaker does not already know the answer and is seeking a real yes-or-no response.
Speaker: "TEXT"
Given the three labels "Declarative/Statement, Rhetorical Interrogative, Yes/No Interrogative", the most likely, single label of the speaker's utterance is:

C.1.2 Task 1b: Attribute-unprovided Function Prediction

System message

```
You are a linguist specializing in colloquial
Malay discourse pragmatics. You must output
exactly one label from the provided list and
nothing else.
```

User message

```
You are a linguist specializing in colloquial
Malay discourse pragmatics. Your task is to
read the Malay sentence below and identify
the primary communicative role the utterance
plays in interaction, beyond its literal
propositional content.
Referring to the following seven labels and
their definitions to make your decision:
Assumed-Agreement Rhetorical Stance:
Speaker presents proposition as already
obvious/shared knowledge; listener is
expected to align rather than genuinely
answer.
Neutral Declarative: Plain informational
statements with minimal discourse pressure or
stance marking.
Information-Seeking Verification: Genuine
request for verification or clarification;
speaker leaves room for disagreement.
Affective Confirmation-Seeking Question:
Speaker seeks confirmation while
simultaneously expressing affect (surprise,
irritation, humor, excitement, disbelief,
etc.)
Emphatic / Discourse-Marking: Particle
functions less as a literal confirmation
marker and more as a discourse-management or
emphasis device.
Null Form Retaining Particle-Like Pragmatic
Meaning: Pragmatic meaning associated with
particles remains inferable even after overt
particle removal.
Negative Rhetorical Challenge / Evaluation:
Speaker uses rhetorical questioning to
criticise, challenge, mock, or negatively
evaluate a proposition rather than genuinely
seek information.
Speaker: "Menggatal/Tuaran no electric since
8:50pm. Ada ke notis dari tadi sy check fb
sesb tda notis. Tu karen kan cukup2 utk buka
1 lampu ja"
Considering what the speaker is
communicatively doing with this utterance,
their stance, their orientation toward the
listener, and the function the sentence
serves in interaction, which of the seven
labels best captures its primary discourse
role?
The most likely, single label is:
```

C.1.3 Task 1c: Attribute-provided Function Prediction

System message

```
You are a linguist specializing in colloquial
Malay discourse pragmatics. You must output
exactly one label from the provided list and
nothing else.
```

User message

```
You are a linguist specializing in colloquial
Malay discourse pragmatics. Your task is to
read the Malay sentence below and identify
the primary communicative role the utterance
plays in interaction, beyond its literal
propositional content.
You are provided with the following
human-annotated attribute labels for this
sentence as additional context:
Certain: The speaker treats the statement as
already true or established — no hedging, no
doubt, full confidence.
Middle/End: The particle appears anywhere
other than the front — mid-sentence or at the
end.
Neutral/Unclear: No clear orientation toward
listener agreement; a plain statement or
ambiguous stance.
Neutral/Unclear: No detectable emotional
charge, or the emotion is genuinely ambiguous.
Rhetorical Interrogative: Phrased as a
question but does not expect a genuine answer;
used to make a point or emphasize something.
Use these attributes to inform your decision,
but base your final label on the overall
communicative function of the utterance.
Referring to the following seven labels and
their definitions to make your decision:
Assumed-Agreement Rhetorical Stance:
Speaker presents proposition as already
obvious/shared knowledge; listener is
expected to align rather than genuinely
answer.
Neutral Declarative: Plain informational
statements with minimal discourse pressure or
stance marking.
Information-Seeking Verification: Genuine
request for verification or clarification;
speaker leaves room for disagreement.
Affective Confirmation-Seeking Question:
Speaker seeks confirmation while
simultaneously expressing affect (surprise,
irritation, humor, excitement, disbelief,
etc.)
Emphatic / Discourse-Marking: Particle
functions less as a literal confirmation
marker and more as a discourse-management or
emphasis device.
Null Form Retaining Particle-Like Pragmatic
Meaning: Pragmatic meaning associated with
particles remains inferable even after overt
particle removal.
Negative Rhetorical Challenge / Evaluation:
Speaker uses rhetorical questioning to
criticise, challenge, mock, or negatively
evaluate a proposition rather than genuinely
seek information.
Speaker: "Menggatal/Tuaran no electric since
8:50pm. Ada ke notis dari tadi sy check fb
sesb tda notis. Tu karen kan cukup2 utk buka
1 lampu ja"
Considering what the speaker is
communicatively doing with this utterance,
their stance, their orientation toward the
listener, and the function the sentence
serves in interaction, which of the seven
labels best captures its primary discourse
role?
```

```
The most likely, single label is:
```

C.1.4 Task 2a: Unprovided Particle Generation

```
You are a linguist specializing in colloquial
Malay discourse particles.  A discourse
particle has been removed from the Malay
sentence below and replaced with [___]. Your
task is to predict which particle, either "ke"
or "kan" or "neutral", belongs in that slot.
Particle meanings:
ke : signals genuine uncertainty or invites
the listener to confirm something the speaker
is unsure about.
kan : signals assumed shared knowledge and
seeks light confirmation of something the
speaker already believes ("right?").
neutral : indicates no particle.
Speaker: "TEXT"
Given the three candidate particles "kan" and
"ke" and "neutral", the single most likely
particle to fill [___] is:
```

C.1.5 Task 2b: Attribute-provided Particle Generation

System message

```
You are a linguist specializing in colloquial
Malay discourse particles.  You must output
exactly one word - either "ke" or "kan" or
"neutral" - and nothing else.
```

User message Example attributes: Epistemic Stance = Certain, Particle Position = Middle/End, Listener Agreement = Assumed Agreement, Emotion = Negative, Question Type = Rhetorical Interrogative.

```
You are given a Malay sentence in which one
discourse particle has been replaced with
[___].
Your task is to predict which particle, either
"ke," "kan," or "neutral," belongs in the
[___] slot, such that the discourse-context
attributes for this sentence are
The speaker treats the statement as already
true or established - no hedging, no doubt,
full confidence.
The particle appears anywhere other than the
front - mid-sentence or at the end.
The speaker treats the information as shared
or obvious - the underlying tone is 'you
already know this'.
The speaker expresses frustration, annoyance,
disappointment, sadness, anger, or
bitterness.
Phrased as a question but does not expect
a genuine answer; used to make a point or
emphasise something.
Particle meanings:
ke : signals genuine uncertainty or invites
the listener to confirm something the speaker
is unsure about.
kan : signals assumed shared knowledge and
seeks light confirmation of something the
speaker already believes ("right?").
neutral : indicates no particle.
Speaker: "TEXT"
Using the sentence context and the attributes
above, which single particle - "ke" or "kan"
or "neutral" - best fills [___]?
Return exactly one word from this set and
nothing else: ke, kan, neutral
```

C.1.6 Task 2c: Function-provided Particle Generation

System message

```
You are a linguist specializing in colloquial
Malay discourse particles.  You must output
exactly one word - either "ke" or "kan" or
"neutral" - and nothing else.
```

User message

```
You are given a Malay sentence in which
one discourse particle has been replaced
with [___].  Your task is to predict which
particle - "ke", "kan", or "neutral" -
belongs in the [___] slot, such that the
sentence is consistent with the primary
communicative role the utterance plays in
interaction, beyond its literal propositional
content: Assumed-Agreement Rhetorical Stance:
Speaker presents proposition as already
obvious/shared knowledge; listener is
expected to align rather than genuinely
answer.
Particle meanings:
ke: signals genuine uncertainty or invites the
listener to confirm something the speaker is
unsure about.
kan: signals assumed shared knowledge and
seeks light confirmation of something the
speaker already believes ("right?").
neutral: indicates no particle.
Speaker: "Menggatal/Tuaran no electric since
8:50pm. Ada [___] notis dari tadi sy check fb
sesb tda notis. Tu karen kan cukup2 utk buka
1 lampu ja"
Using the sentence context and the
macro-function above, which single particle
- "ke" or "kan" or "neutral" - best fills
[___]?
Return exactly one word from this set and
nothing else: ke, kan, neutral
```

C.1.7 Task 2d: Attribute & Function-provided Particle Generation

System message

```
You are a linguist specialising in colloquial
Malay discourse particles.  You must output
exactly one word - either "ke", "kan", or
"neutral" - and nothing else.
```

User prompt

```
You are given a Malay sentence in which one discourse particle has been replaced with [___].
Discourse-context attributes for this sentence: - Epistemic Stance: {ES_VAL} — {ES_DESC} - Particle Position: {PP_VAL} — {PP_DESC} - Listener Agreement: {LA_VAL} — {LA_DESC} - Emotion: {EM_VAL} — {EM_DESC} - Question Type: {QT_VAL} — {QT_DESC}
Macro-pragmatic function of this utterance: {MACRO_FUNCTION}: {MACRO_FUNCTION_DEFINITION}
Particle meanings: ke : signals genuine uncertainty or invites the listener to confirm something the speaker is unsure about. kan : signals assumed shared knowledge and seeks light confirmation of something the speaker already believes ("right?"). neutral : indicates no particle.
Speaker: "{TEXT_MASKED}"
Using both the attribute breakdown and the macro-function above, which single particle — "ke", "kan", or "neutral" — best fills [___]? Return exactly one word from this set and nothing else: ke, kan, neutral
```

C.1.8 Task 3a: CoT Attribute-unprovided Function Prediction

System message

```
You are a linguist specialising in colloquial Malay discourse pragmatics. Reason step-by-step about the speaker's communicative intent, then output exactly one macro-function label from the provided list.
```

User prompt

```
You are given a Malay sentence. Your task is to classify its primary macro-pragmatic function — the communicative role the utterance plays in interaction, beyond its literal propositional content.
Referring to the following seven labels and their definitions:
Assumed-Agreement Rhetorical Stance: Speaker presents proposition as already obvious/shared knowledge; listener is expected to align rather than genuinely answer.
Neutral Declarative: Plain informational statements with minimal discourse pressure or stance marking.
Information-Seeking Verification: Genuine request for verification or clarification; speaker leaves room for disagreement.
Affective Confirmation-Seeking Question: Speaker seeks confirmation while simultaneously expressing affect (surprise, irritation, humour, excitement, disbelief, etc.).
Emphatic / Discourse-Marking: Particle functions less as a literal confirmation marker and more as a discourse-management or emphasis device.
Null Form Retaining Particle-Like Pragmatic Meaning: Pragmatic meaning associated with particles remains inferable even after overt particle removal.
Negative Rhetorical Challenge / Evaluation: Speaker uses rhetorical questioning to criticise, challenge, mock, or negatively evaluate a proposition rather than genuinely seek information.
Speaker: "{TEXT}"
Think step-by-step: 1. What is the speaker doing in this utterance (asserting, asking, hedging, emphasising, challenging)? 2. What is the speaker's stance toward the listener (shared knowledge, seeking confirmation, neutral information, etc.)? 3. Which of the seven labels best captures this combination?
Format your response EXACTLY as follows, with no extra text before or after:
Reasoning: <step-by-step analysis> Final answer: <one label from the list above>
```

C.1.9 Task 3b: CoT Function-Constrained Particle Generation

System message

```
You are a linguist specialising in colloquial Malay discourse particles. Reason step-by-step, then output exactly one particle — "ke", "kan", or "neutral".
```

User prompt

```
You are given a Malay sentence in which one discourse particle has been replaced with [___].
The macro-pragmatic function of this utterance is: {MACRO_FUNCTION}: {MACRO_FUNCTION_DEFINITION}
Particle meanings: ke : signals genuine uncertainty or invites the listener to confirm something the speaker is unsure about. kan : signals assumed shared knowledge and seeks light confirmation of something the speaker already believes ("right?"). neutral : indicates no particle.
Speaker: "{TEXT_MASKED}"
Think step-by-step: 1. What is the speaker communicatively doing, given the macro-function above? 2. Does the masked slot call for genuine uncertainty (ke), assumed shared knowledge (kan), or no particle (neutral)? 3. Which choice best fits both the sentence flow and the macro-function?
Format your response EXACTLY as follows, with no extra text before or after:
Reasoning: <step-by-step analysis> Final answer: <ke | kan | neutral>
```

C.1.10 Task 3c: Attribute- and Function-Constrained Particle Generation

System message

```
You are a linguist specialising in colloquial Malay discourse particles. You must output exactly one word — either "ke", "kan", or "neutral" — and nothing else.
```

User prompt

```
You are given a Malay sentence in which one discourse particle has been replaced with [___].
Discourse-context attributes for this sentence: - Epistemic Stance: {ES_VAL} — {ES_DESC} - Particle Position: {PP_VAL} — {PP_DESC} - Listener Agreement: {LA_VAL} — {LA_DESC} - Emotion: {EM_VAL} — {EM_DESC} - Question Type: {QT_VAL} — {QT_DESC}
Macro-pragmatic function of this utterance: {MACRO_FUNCTION}: {MACRO_FUNCTION_DEFINITION}
Particle meanings: ke : signals genuine uncertainty or invites the listener to confirm something the speaker is unsure about. kan : signals assumed shared knowledge and seeks light confirmation of something the speaker already believes ("right?"). neutral : indicates no particle.
Speaker: "{TEXT_MASKED}"
Using both the attribute breakdown and the macro-function above, which single particle — "ke", "kan", or "neutral" — best fills [___]? Return exactly one word from this set and nothing else: ke, kan, neutral
```

C.1.11 Task 3b: CoT Function-provided Particle Generation

System message

```
You are a linguist specialising in colloquial Malay discourse particles. Reason step-by-step, then output exactly one particle — "ke", "kan", or "neutral".
```

User prompt

```
You are given a Malay sentence in which one discourse particle has been replaced with [___].
The macro-pragmatic function of this utterance is: {MACRO_FUNCTION}: {MACRO_FUNCTION_DEFINITION}
Particle meanings: ke : signals genuine uncertainty or invites the listener to confirm something the speaker is unsure about. kan : signals assumed shared knowledge and seeks light confirmation of something the speaker already believes ("right?"). neutral : indicates no particle.
Speaker: "{TEXT_MASKED}"
Think step-by-step: 1. What is the speaker communicatively doing, given the macro-function above? 2. Does the masked slot call for genuine uncertainty (ke), assumed shared knowledge (kan), or no particle (neutral)? 3. Which choice best fits both the sentence flow and the macro-function?
Format your response EXACTLY as follows, with no extra text before or after:
Reasoning: <step-by-step analysis> Final answer: <ke | kan | neutral>
```

C.2 Translated Malay Prompts Examples

The prompt for Task (1a) Attribute Prediction was translated by the principal author (a native Malay speaker) to test the same models on whether their performance would change with the prompt language.

C.2.1 Tugas 1a: Ramalan Atribut

Atribut 1: Pendirian Epistemik

```
Anda seorang ahli bahasa yang pakar dalam bahasa Melayu kolokial. Tugas anda adalah untuk membaca ayat Bahasa Melayu di bawah dan menentukan sejauh mana kepastian penutur tentang maklumat yang mereka sampaikan.
Merujuk kepada tiga label berikut dan definisinya untuk membuat keputusan anda:
Pasti: Penutur menganggap pernyataan itu sebagai benar atau sah. Tiada lindung nilai, tiada keraguan, dan tiada semakan. Penutur menegaskan maklumat tersebut dengan penuh keyakinan.
Tidak Pasti: Penutur kedengaran tidak pasti, sedang meneka, sedang menganggarkan atau sedang menyemak sama ada sesuatu itu benar. Perkataan seperti "agaknya", "kot", atau partikel soalan yang menyiasat untuk pengesahan adalah isyarat tipikal.
Neutral/Tidak Jelas: Ayat ini tidak membawa sebarang isyarat kepastian yang boleh dikesan sama ada dalam arah. Ini terpakai kepada penerangan, arahan atau ayat neutral di mana kepastian tidak relevan sama sekali.
Penutur: "{TEKS}"
Memandangkan tiga label "Pasti, Tidak Pasti, Neutral/Tidak Jelas", label tunggal yang paling mungkin untuk ujaran penutur ialah:
```

Atribut 2: Kedudukan Partikel

```
Anda seorang ahli bahasa yang pakar dalam bahasa Melayu kolokial. Tugas anda adalah untuk membaca ayat Bahasa Melayu di bawah dan menentukan di mana partikel wacana muncul di dalamnya.
Merujuk kepada tiga label berikut dan definisinya untuk membuat keputusan anda:
Depan: Partikel itu muncul pada permulaan ayat, sebelum sebarang perkataan kandungan yang lain.
Tengah/Akhir: Partikel tersebut muncul di mana-mana sahaja selain bahagian hadapan — pertengahan ayat, sebelum perkataan terakhir, atau di hujungnya.
Tiada Partikel: Tiada partikel wacana yang terdapat dalam ayat.
Penutur: "{TEKS}"
Memandangkan tiga label "Depan, Tengah/Akhir, Tiada Partikel", label tunggal yang paling mungkin untuk ujaran penutur ialah:
```

Atribut 3: Persetujuan Pendengar

Anda seorang ahli bahasa yang pakar dalam bahasa Melayu kolokial. Tugas anda adalah untuk membaca ayat Bahasa Melayu di bawah dan memutuskan bagaimana penutur itu memberi orientasi kepada pendengar dari segi pengetahuan atau persetujuan bersama.
Merujuk kepada tiga label berikut dan definisinya untuk membuat keputusan anda:
Perjanjian Diandaikan: Penutur menganggap maklumat tersebut sebagai telah dikongsi atau jelas kepada pendengar. Ayat tersebut dibentangkan sebagai titik persamaan — nada yang mendasarinya ialah "anda sudah tahu ini" atau "sudah tentu ini benar". Tiada pengesahan eksplisit diminta.
Mencari Pengesahan: Penutur sedang giat menyemak sama ada pendengar bersetuju, tahu atau boleh mengesahkan maklumat tersebut. Ayat tersebut menjemput atau meminta pengesahan pendengar sebelum penutur boleh meneruskan dengan yakin.
Neutral/Tidak Jelas: Ayat ini tidak menunjukkan sebarang orientasi yang jelas ke arah persetujuan pendengar. Ini terpakai kepada pernyataan, arahan atau kes yang jelas di mana pendirian interpersonal terhadap persetujuan adalah samar-samar atau tiada.
Penutur: "{TEKS}"
Memandangkan tiga label "Perjanjian Diandaikan, Mencari Pengesahan, Neutral/Tidak Jelas", label tunggal yang paling mungkin untuk ujaran penutur ialah:

Atribut 4: Emosi / Kesan

Anda seorang ahli bahasa yang pakar dalam bahasa Melayu kolokial. Tugas anda adalah untuk membaca ayat Bahasa Melayu di bawah dan menentukan nada emosi penuturnya.
Merujuk kepada tiga label berikut dan definisinya untuk membuat keputusan anda:
Positif: Penutur meluahkan kegembiraan, keterujaan, semangat, kepuasan, humor, kasih sayang, kelegaan, atau sebarang perasaan positif yang jelas. Ini termasuk usikan ringan atau sindiran suka bermain yang bernada hangat.
Negatif: Penutur meluahkan rasa kecewa, kegusaran, kekecewaan, kesedihan, kemarahan, kepahitan atau sebarang perasaan negatif yang lain. Ini termasuk sindiran bermusuhan atau pahit.
Neutral/Tidak Jelas: Ayat tersebut tidak membawa sebarang cas emosi yang boleh dikesan dalam kedua-dua arah, atau emosi tersebut benar-benar samar-samar dan tidak boleh diklasifikasikan dengan andal sebagai positif atau negatif.
Penutur: "{TEKS}"
Memandangkan tiga label "Positif, Negatif, Neutral/Tidak Jelas", label tunggal yang paling mungkin untuk ujaran penutur ialah:

Atribut 5: Jenis Soalan

Anda seorang ahli bahasa yang pakar dalam bahasa Melayu kolokial. Tugas anda adalah untuk membaca ayat Bahasa Melayu di bawah dan menentukan fungsi ayat utamanya.
Merujuk kepada tiga label berikut dan definisinya untuk membuat keputusan anda:
Deklaratif/Pernyataan: Ayat ini membuat penegasan atau menyampaikan maklumat. Ia menerangkan situasi, menyatakan fakta atau menyatakan pandangan. Ia tidak berstruktur sebagai soalan, walaupun ia berakhir dengan partikel.
Tanya Jawab Retorik: Ayat ini diungkapkan sebagai soalan tetapi tidak mengharapkan jawapan yang tulen daripada pendengar. Ia digunakan untuk mengemukakan sesuatu perkara, meluahkan emosi atau menekankan sesuatu — penutur sudah menyiratkan jawapan tersebut melalui soalan itu sendiri.
Tanya Jawab Ya/Tidak: Ayat ini merupakan soalan tulen yang mengajak pendengar untuk mengesahkan atau menafikan sesuatu. Penutur belum mengetahui jawapannya dan sedang mencari jawapan ya atau tidak yang sebenar.
Penutur: "{TEKS}"
Memandangkan tiga label "Deklaratif/Pernyataan, Tanya Jawab Retorik, Tanya Jawab Ya/Tidak", label tunggal yang paling mungkin untuk ujaran penutur ialah: