YUE-PUB: A Cantonese Pragmatic Understanding Benchmark

Chengyan Wu[*1] **Ziwei Gong**[*2] **Jinzhi Chen**[*3] **Yajie Wen**[3] **Mengmeng Su**[3]
Yun Xue[3] **Bolei Ma**[4]

[1]Sun Yat-sen University [2]Columbia University
[3]South China Normal University [4]LMU Munich & MCML

Correspondence: xueyun [at] m.scnu.edu.cn; bolei.ma [at] lmu.de

Abstract

Pragmatic understanding remains a fundamental challenge in natural language processing, particularly for underrepresented languages such as Cantonese. To alleviate the scarcity of resources in this area, we introduce **YUE-PUB**[1], the first large-scale pragmatic understanding benchmark for Cantonese. Constructed from the PUB dataset through automatic translation with human verification for quality assurance, YUE-PUB covers three core pragmatic phenomena: *Implicature*, *Presupposition*, and *Reference*. We conduct systematic experiments across multiple baseline models to evaluate their performance under several cross-lingual transfer settings, compare models with and without pragmatic topic awareness, and further analyze their sensitivity to different pragmatic phenomena as well as their Chain-of-Thought (CoT) reasoning behavior. Experimental results show that current large language models still exhibit substantial limitations in Cantonese pragmatic understanding, underscoring the value of YUE-PUB as an important resource for advancing pragmatic reasoning research in low-resource languages.

1 Introduction

Pragmatics investigates how context shapes meaning and how individuals convey implied intentions beyond literal text (Huang, 2017). With the proliferation of Large Language Models (LLMs), assessing their ability to generate pragmatically appropriate responses has become critical for building trustworthy human–machine interaction systems (Chkirbene et al., 2024). While recent LLMs demonstrate impressive capabilities in text generation and cross-lingual understanding, their pragmatic comprehension remains under-explored, especially beyond high-resource languages like English (Ma et al., 2025).

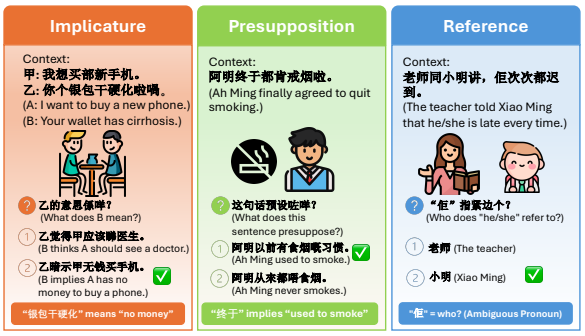


Figure 1: Overview of example task types in YUE-PUB. We evaluate LLMs on three pragmatic phenomena: (A) Implicature, where meaning is implied (e.g., metaphors for poverty); (B) Presupposition, dealing with background assumptions; and (C) Reference, resolving ambiguous pronouns in context. The benchmark challenges models to understand colloquial Cantonese nuances.

Current pragmatic evaluations are predominantly centered on English, benefiting from abundant annotated corpora (Hu et al., 2023; Sravanthi et al., 2024). In sharp contrast, pragmatic resources for low-resource languages are exceedingly scarce (Aji et al., 2022). Existing non-English studies often focus on isolated phenomena, such as idioms or figurative language, rather than offering a comprehensive evaluation of communicative competence (Gallipoli and Cagliero, 2025; Wang et al., 2025). This data scarcity creates a significant bottleneck for understanding how LLMs handle cultural nuance and implicit reasoning in diverse linguistic landscapes.

The gap is particularly prominent in Cantonese. As a major variety of Chinese with over 85 million speakers and a vital role in global communities, Cantonese is linguistically rich and colloquially dynamic (Jiang et al., 2025). However, it remains a low-resource language in NLP due to the lack of standardized written norms and high-quality annotated datasets (Cheng et al., 2025a). The absence of systematic benchmarks hinders the development of

[*] Equal contributions.

[1]Data available at: https://huggingface.co/datasets/Multilingual-NLP/YUE-PUB.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 1–22
November 6-10, 2026 ©2026 Association for Computational Linguistics

Cantonese language technologies capable of deep pragmatic reasoning.

To bridge this gap, we introduce **YUE-PUB**, the first large-scale pragmatic understanding benchmark for Cantonese. YUE-PUB comprises 26,743 high-quality examples across 14 tasks, systematically covering three core pragmatic categories: *implicature*, *presupposition*, and *reference*. We formulate the evaluation as a multiple-choice question answering (MCQA) task to align with real-world interactive scenarios. Our construction pipeline combines automated translation from curated English corpora with rigorous dual-verification by native speakers, ensuring both linguistic naturalness and pragmatic soundness.

We conduct extensive experiments on YUE-PUB using mainstream Cantonese-specific and general-purpose LLMs under zero-shot, few-shot, and Chain-of-Thought (CoT) settings. Our study provides the first comprehensive assessment of LLMs' native pragmatic reasoning capabilities in Cantonese.

In summary, our contributions are:

- We present YUE-PUB, the first comprehensive benchmark dedicated to Cantonese pragmatic understanding, filling a critical gap in low-resource language evaluation.
- We develop a robust data construction pipeline that integrates LLM-assisted translation with native human verification to ensure high-quality, culturally aligned data.
- We provide an extensive evaluation of state-of-the-art LLMs, offering insights into the cross-lingual and native pragmatic reasoning abilities of models in Cantonese.

2 Related Work

Pragmatic Evaluation in LLMs. Early research in computational pragmatics relied on rule-based systems or limited deep learning models, which struggled with complex implicit reasoning (Ma et al., 2025). With the advent of LLMs, the focus has shifted to evaluating how well these models capture context and intent. Recent benchmarks typically assess pragmatic phenomena through tasks like implicature recovery and reference resolution (Hu et al., 2023). For instance, Sravanthi et al. (2024) introduced frameworks to test pragmatic understanding in varied dialogue settings. However, these initiatives are overwhelmingly concentrated on English, leaving a gap in multilingual applicability.

Multilingual and Low-Resource Pragmatics. Research on non-English pragmatic evaluation is emerging but remains fragmented. Some studies have explored specific languages, such as Korean pragmatic inference (Kim et al., 2024; Park et al., 2024a) or Arabic idiom comprehension (Attia et al., 2026). Others have targeted specific phenomena like figurative language or proverbs in multilingual settings (Wu et al., 2024; Gallipoli and Cagliero, 2025). Despite these efforts, most low-resource languages lack comprehensive benchmarks that cover a wide spectrum of pragmatic phenomena (e.g., combining presupposition, reference, and implicature). Park et al. (2024b) attempted to extend evaluations multilingually, yet the coverage of truly low-resource languages remains limited.

Cantonese Natural Language Processing. Cantonese serves as an important lingua franca but faces significant challenges in NLP due to the scarcity of training data and the divergence between spoken and written forms (Cheng et al., 2025a). While there have been recent efforts to build Cantonese corpora and foundational models (Jiang et al., 2025; Min et al., 2025), research specifically targeting higher-level semantic and pragmatic understanding is scarce. Our work addresses this deficiency by providing a structured benchmark for assessing pragmatic competence in Cantonese. We recently also released a speech-based version of the benchmark (Wen et al., 2026).

3 Data Construction

To build a robust benchmark for evaluating Cantonese pragmatic understanding, we introduce **YUE-PUB**. Our construction pipeline translates and adapts high-quality English pragmatic resources into Cantonese, ensuring both linguistic naturalness and the preservation of original pragmatic intents. The overall workflow is illustrated in Figure 2.

3.1 Source Data Selection

We construct YUE-PUB based on the publicly available PUB benchmark (Sravanthi et al., 2024), which aggregates diverse expert-annotated datasets including CIRCA (Louis et al., 2020), GRICE (Zheng et al., 2021), FigQA (Liu et al., 2022), FLUTE (Chakrabarty et al., 2022), IMPPRES (Jeretic et al.,

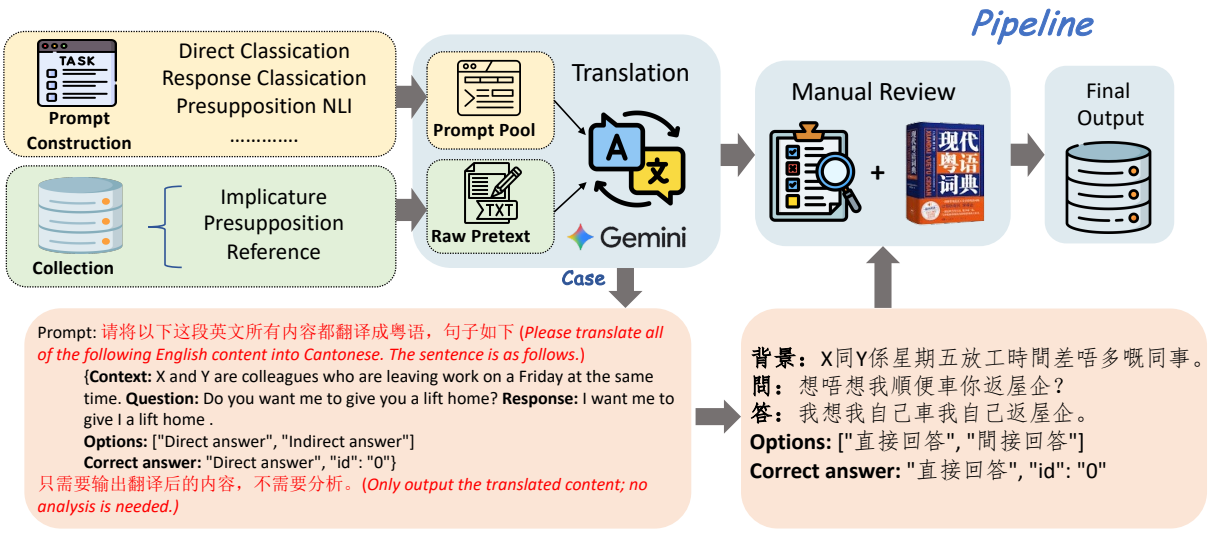


Figure 2: The construction process of the YUE-PUB dataset.

2020), and others. These sources cover a wide spectrum of pragmatic phenomena, ranging from conversational implicature and presupposition to figurative language. For consistency, all datasets are adapted into a unified multiple-choice question–answering format. This unification allows for consistent assessment across different pragmatic categories while mimicking real-world discriminative tasks in dialogue systems.

3.2 Construction Pipeline

Our data construction consists of three main stages: Prompt Construction, Automated Translation, and Manual Verification.

Stage 1: Prompt Construction. Since the original datasets vary in format (e.g., NLI, classification), we first design task-specific templates to structure the raw instances. As shown in the "Prompt Pool" of Figure 2, these templates contextualize the input text and candidate options, preparing them for the subsequent translation and QA formatting steps.

Stage 2: Automated Translation. We employ an advanced LLM Gemini (Team et al., 2025) to translate the English instances into Cantonese. The model is prompted to translate the background context, questions, and options while maintaining the subtle pragmatic nuances of the original text. This automated step ensures scalability and provides a strong initial baseline for the dataset.

Stage 3: Dual-Stage Manual Verification. To address the low-resource challenges and lack of standardized norms in Cantonese, we implement a rigorous human-in-the-loop verification process. Two native Cantonese speakers review the translated data in two phases:

1) **Naturalness & Accuracy:** Annotators correct literal translation errors and ensure the text aligns with Cantonese idiomatic usage and grammar.

2) **Consistency & Terminology:** A second round of review is conducted to unify lexical choices across the dataset. Annotators consult the *Modern Cantonese Dictionary* (Man et al., 2025) to resolve ambiguities and standardize terminology.

This dual-verification mechanism ensures that YUE-PUB achieves native-level fluency while rigorously preserving the logic required for pragmatic reasoning.

3.3 Dataset Statistics

The final YUE-PUB benchmark comprises **26,743** examples across 14 distinct tasks, categorized into three core pragmatic phenomena: *Implicature*, *Presupposition*, and *Reference*. The detailed statistics and source distributions are presented in Table 1.

4 Data Quality Evaluation

To ensure the reliability of YUE-PUB as a benchmark, we conduct a comprehensive quality assessment of the translated Cantonese data using both automated metrics and rigorous human evaluation.

Task ID	Size	Phenomena	Source
Task 1–3	7,512	Implicature	CIRCA
Task 4	2,000	Implicature	GRICE
Task 5–6	4,000	Implicature	FigQA
Task 7–9	5,310	Implicature	FLUTE
Task 10	2,100	Implicature	IMPPRES
Task 11	1,800	Presupposition	IMPPRES
Task 12	2,563	Presupposition	DailyDialog
Task 13	1,000	Reference	GRICE
Task 14	458	Reference	Metonymy
Total	**26,743**		

Table 1: Statistics of YUE-PUB. The benchmark covers three major pragmatic categories derived from seven high-quality source datasets.

Auto Evaluation				
Task	BLEU	ChrF++	BERTScore	SBERT
Implicature	33.03	75.95	94.51	81.51
Presupposition	36.85	78.35	94.70	87.35
Reference	38.20	77.55	93.95	81.10
Human Evaluation (1–5 Scale)				
Task	Grammar	Fluency	Adequacy	Idiomatic
Implicature	3.98	4.03	3.85	3.95
Presupposition	4.10	4.10	4.10	3.97
Reference	4.04	4.10	3.92	4.04
Average	**4.01**	**4.05**	**3.90**	**3.96**

Table 2: Data quality evaluation results. The upper section reports automated metrics comparing model outputs against human-verified references. The lower section details human ratings averaged over 14 tasks.

4.1 Automated Evaluation

We employ a diverse set of metrics to assess translation quality from lexical, character-based, and semantic perspectives. Specifically, we report:

- **Lexical Overlap: BLEU-4** (Papineni et al., 2002) and **chrF++** (Popović, 2017) are used to measure n-gram and character-level correspondence between the model outputs and human-verified references.
- **Semantic Similarity:** We utilize **BERTScore** (Zhang et al., 2020) and **SBERT** (Reimers and Gurevych, 2019) to evaluate how well the translations preserve the semantic meaning of the source text beyond surface-level matching.

As shown in the upper section of Table 2, the dataset achieves high scores across all metrics. Notably, the high BERTScore (> 94.0 on average) and SBERT scores confirm that the semantic integrity of the pragmatic phenomena is well-preserved during the translation process.

4.2 Human Evaluation

Given the complexity of pragmatic nuances, automated metrics alone are insufficient. We therefore conduct a fine-grained human evaluation involving 420 randomly sampled instances (30 from each of the 14 tasks). Two native Cantonese speakers annotated these samples. Detailed description is provided in Appendix B.

Evaluation Criteria. Following established machine translation protocols (Chen et al., 2022; Wu et al., 2025), we assess the data on a 5-point Likert scale (1=Poor, 5=Excellent) across four dimensions:

1) **Grammar:** Syntactic correctness of the Cantonese text.
2) **Fluency:** Naturalness and flow of the translation.
3) **Adequacy:** The extent to which the original meaning is preserved.
4) **Idiomaticity:** A specific metric introduced to measure the alignment with authentic, contemporary Cantonese usage, distinguishing it from rigid machine-translated text.

Results. The lower section of Table 2 presents the human evaluation results. The translations receive consistently high ratings (> 3.9 average) across all categories. The scores for *Implicature*, *Presupposition*, and *Reference* are balanced, indicating consistent quality across diverse pragmatic tasks. The high **Idiomaticity** score (3.96) particularly highlights the effectiveness of our pipeline in generating natural, culturally aligned Cantonese, validating YUE-PUB as a high-quality resource for pragmatic research.

5 Task and Experimental Settings

5.1 Task Definitions

Our benchmark evaluates pragmatic understanding through a unified multiple-choice question framework. We adapt methodologies from prior work on figurative language and pragmatic reasoning (Kabra

et al., 2023; Sravanthi et al., 2024, e.g.,) to the context of Cantonese.

5.2 Prompt Strategies

In this work, we adopt a prompt-based evaluation framework and design both monolingual (yue-yue) and cross-lingual (en-yue, zh-yue) experimental settings. Our experiments cover three major paradigms: *zero-shot*, *chain-of-thought* (CoT), and *few-shot* prompting. The detailed prompts used are presented in Appendix A.

(1) Zero-shot Setting. In the zero-shot scenario, the model is directly provided with a task prompt and a question, without any additional labeled examples. To systematically examine the role of topic-aware prompting, we consider the following variants:

- **With-Topic-Aware prompts:** Prompts explicitly include the topic. (i.e., You are now working on a task about pragmatic phenomena, with the topic being 'Implicature'.)
- **Without-Topic-Aware prompts:** Prompts do not include any topic information.
- **Model-size comparison:** To investigate the effect of model scaling, we conduct controlled experiments across multiple sizes of the Qwen family of models, including 0.5B, 1.5B, 3B, 7B, 32B, and 72B variants.

(2) Chain-of-Thought (CoT). To evaluate whether explicit multi-step reasoning improves performance on multiple-choice QA tasks, we adopt CoT prompting (Wei et al., 2022; Wang et al., 2023). The model is instructed to first produce a structured reasoning process and then output the final answer, enabling us to assess the contribution of intermediate reasoning steps.

(3) Few-shot Setting. We further explore the impact of demonstration quantity by adopting a few-shot prompting setup with varying numbers of in-context examples. Specifically, we experiment with $K \in \{2, 4, 8, 16\}$, allowing us to systematically analyze how model performance scales with additional task demonstrations and to assess the model's generalization ability under limited supervision.

5.3 Models and Settings

We evaluate 10 models, including eight open-source models, spanning multilingual, Chinese, and Cantonese-specialized categories, as well as two commercial models. The LLMs evaluated are as follows:

(1) Multilingual model: Aya 23 8B (Aryabumi et al., 2024), Gemma-2-9B (Team et al., 2024), Llama-3.1-8B (AI@Meta, 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023).

(2) Chinese model: Yi-1.5-9B (Young et al., 2024) and Qwen 2.5 7B (Team, 2024), Deepseek-llm-7b-chat (DeepSeek-AI et al., 2024).

(3) Specific-Model: CantoneseLLM (CLLM) (Cheng et al., 2025b).

(4) Commercial models: Gemini 2.5 (Team et al., 2025), GPT-4o mini (OpenAI et al., 2024).

Evaluation. All experiments across different LLMs were conducted on 2 $\times$ NVIDIA A6000 GPUs with 48GB memory each. For generation, we used a sampling configuration with a top-p value of 1.0 and a temperature of 0.2, together with the task-specific prompts provided in Appendix A. Besides, we adopt accuracy as the main evaluation metric for all tasks. A prediction is considered correct if and only if the model selects the correct option for the given question.

6 Results and Analysis

6.1 Mono- and Cross-Lingual Pragmatics Transfer

We present the evaluation results of 10 LLMs across three pragmatic categories under varying experimental settings. Table 2 summarizes the performance for Cantonese, English, and Mandarin source prompts, respectively.

Performance Disparities across Pragmatic Phenomena. Under the monolingual setting, Presupposition accuracy lags by over 9%, a disparity we attribute to the interplay between translation artifacts and cognitive complexity. Specifically, the loss of subtle triggers during translation weakens input cues, while the intrinsic need to maintain global shared background assumptions for Presupposition demands higher pragmatic modeling capabilities than the local context inference sufficient for Implicature. Unlike Implicature which often relies on local context (Grice, 1975), Presupposition requires the model to maintain global shared background assumptions (Bach, 2006; Jeretic et al., 2020). This demands higher-order pragmatic modeling capabilities that current LLMs struggle to generalize in low-resource settings.

Source	Task	Aya 23	Gemma 2	Llama 3.1	Mistral	Yi 1.5	Qwen 2.5	Deepseek	CantoneseLLM	Gemini 2.5	GPT-4o mini
Cantonese (Yue → Yue)	***Zero-Shot (with topic)***										
	Implicature	63.46	68.65	65.40	58.58	61.53	65.23	58.22	64.93	69.96	68.53
	Presupposition	53.41	53.49	54.49	40.82	66.25	52.94	44.85	62.81	57.84	56.78
	Reference	60.59	62.38	63.99	47.12	56.44	58.55	53.75	65.66	63.44	63.54
	Avg.	59.15	61.51	61.29	48.84	61.41	58.91	52.27	64.47	63.75	62.95
	Zero-Shot (without topic)										
	Implicature	62.46	66.99	64.82	58.66	59.28	66.86	57.72	68.92	68.75	68.77
	Presupposition	51.49	55.10	53.70	39.52	64.45	53.93	44.34	59.51	58.35	57.76
	Reference	60.56	63.09	63.47	46.52	54.80	57.25	51.76	65.01	64.18	64.12
	Avg.	58.17	61.73	60.66	48.23	59.51	59.35	51.27	64.48	63.76	63.55
	Chain-of-Thought (CoT)										
	Implicature	61.24	68.97	64.02	58.79	65.59	66.10	62.89	65.34	70.48	69.95
	Presupposition	54.64	57.83	56.40	41.32	69.15	57.85	51.55	66.25	59.13	58.69
	Reference	70.18	65.54	64.72	47.23	60.00	59.67	57.22	66.65	65.47	66.11
	Avg.	62.02	64.11	61.71	49.11	64.91	61.21	57.22	66.08	65.03	64.92
English (En → Yue)	***Zero-Shot (with topic)***										
	Implicature	64.46	68.96	57.75	55.96	63.83	65.79	60.85	62.02	70.26	69.79
	Presupposition	60.35	56.47	59.21	35.85	62.08	56.42	49.60	58.32	59.34	58.45
	Reference	60.32	66.26	63.29	42.69	57.98	61.50	53.15	64.60	65.27	65.12
	Avg.	61.71	63.90	60.08	44.83	61.30	61.24	54.53	61.65	64.96	64.45
	Zero-Shot (without topic)										
	Implicature	65.82	65.21	60.63	56.94	63.61	67.16	60.80	64.54	71.23	70.13
	Presupposition	58.62	55.61	56.89	35.01	61.45	59.18	47.12	60.44	60.24	58.89
	Reference	60.69	62.75	63.33	42.74	57.30	60.98	49.28	61.71	65.33	65.01
	Avg.	61.71	61.19	60.28	44.90	60.79	62.44	52.40	62.23	65.60	64.68
	Chain-of-Thought (CoT)										
	Implicature	59.14	69.51	58.70	59.88	66.38	64.81	61.36	66.15	72.34	72.41
	Presupposition	68.96	56.47	61.35	50.67	63.80	63.96	53.00	63.80	61.23	61.56
	Reference	68.03	69.66	63.35	51.67	59.45	64.59	57.18	65.85	66.37	66.23
	Avg.	65.38	65.21	61.13	54.07	63.21	64.45	57.18	65.27	66.65	66.73
Mandarin (Zh → Yue)	***Zero-Shot (with topic)***										
	Implicature	65.88	68.05	63.08	57.15	65.79	65.13	59.07	64.60	70.43	69.21
	Presupposition	51.12	54.48	59.85	43.97	65.07	50.74	41.60	61.35	58.68	57.79
	Reference	61.49	63.76	64.11	48.82	61.92	58.73	51.05	67.57	62.43	64.53
	Avg.	59.50	62.10	62.35	49.98	64.26	58.20	50.57	64.51	63.85	63.84
	Zero-Shot (without topic)										
	Implicature	64.36	67.20	62.71	57.41	61.81	65.13	56.08	65.10	70.95	69.93
	Presupposition	46.90	55.10	57.02	34.21	63.45	52.74	38.56	63.26	57.97	58.11
	Reference	59.64	62.57	63.76	46.71	61.10	59.63	49.40	65.55	64.35	64.79
	Avg.	56.97	61.62	61.16	46.11	62.12	59.17	48.01	64.64	64.42	64.28
	Chain-of-Thought (CoT)										
	Implicature	55.92	59.32	50.42	54.21	68.39	66.24	61.55	65.97	71.75	70.54
	Presupposition	58.75	58.41	61.40	31.17	66.85	56.66	40.50	61.85	59.99	59.38
	Reference	61.72	64.33	60.45	48.59	63.65	59.95	51.03	69.15	66.21	66.64
	Avg.	58.80	60.69	57.42	44.66	66.30	60.95	51.03	65.66	65.98	65.52

Table 3: Consolidated model performance across different instruction sources: **Cantonese** (Yue → Yue), **English** (En → Yue), and **Mandarin** (Zh → Yue).

Cross-Lingual Transfer and Linguistic Proximity. We observe that general-purpose multilingual models demonstrate superior performance in the Mandarin → Cantonese setting compared to English → Cantonese. This advantage stems from the syntactic and lexical overlap within the Sino-Tibetan family (Thurgood et al., 1985), which preserves pragmatic cues more effectively than transfer from distinct language families like Indo-European. Notably, general multilingual models (e.g., Qwen 2.5) often surpass Chinese-centric models in cross-lingual transfer due to robust alignment capabilities (Conneau et al., 2020), though the domain-specialized **CantoneseLLM** dominates the native Cantonese setting by leveraging culture-specific pre-training data.

Prompting strategies yield mixed results; explicit topic tags offer only marginal gains, while CoT benefits are scale-dependent. Topic-aware prompting acts merely as a "weak constraint" that fails to bridge gaps in background knowledge. Similarly, while CoT improves complex Implicature tasks in larger models, it occasionally degrades performance in simpler tasks or smaller models, suggesting that forcing verbose reasoning for intuitive pragmatic judgments may introduce hallucinations (Turpin et al., 2023).

Commercial models perform slightly better than open models. We observe a consistent trend where the two commercial models (Gemini 2.5 and

Setting	Imp.	Pre.	Ref.	Avg.
En → Yue	65.79	56.42	61.50	61.24
Zh → Yue	65.13	50.74	58.73	58.20
Yue → Yue	65.23	52.94	58.55	58.91
Reference Baseline (English Source Data)				
En → En	70.92	57.68	63.30	63.97

Table 4: Cross-lingual transfer performance comparison using Qwen 2.5 7B. Detailed results on other models can be found in Appendix D.

GPT-4o mini) achieve higher overall average scores compared to the open-weight models across the evaluation sets. However, the performance gap remains quite narrow. Several open models, particularly CantoneseLLM 7B, Gemma 2 9B, and Yi 1.5 9B, demonstrate highly competitive performance, occasionally matching or slightly exceeding the commercial models on specific sub-tasks.

Instruction Following and Language Asymmetry. A notable asymmetry exists in instruction following, where models favor prompts in their dominant pre-training languages over Cantonese. As indicated in Table 4, even when the task content is in Cantonese, utilizing mixed-code prompts (e.g., English instructions) often yields better results. This underscores the necessity for improved alignment strategies to ensure that pragmatic competence in low-resource languages is not bounded by instruction-following limitations. Furthermore, performance on YUE-PUB remains significantly lower than on the original English PUB benchmark. This substantial gap highlights that despite cross-lingual transfer capabilities, the fundamental deficiency in low-resource pragmatic reasoning remains a critical bottleneck requiring dedicated alignment strategies.

6.2 Scaling Analysis of Qwen Models

To investigate the scaling laws of pragmatic competence in a low-resource setting, we conduct a systematic study using the Qwen model family, spanning parameter sizes from 0.5B to 72B. The performance trajectories under the zero-shot (topic-aware) setting are illustrated in Figure 3.

Positive Scaling. We observe a robust and statistically significant positive correlation between model size and pragmatic accuracy across all linguistic settings. Pearson correlation analysis confirms this strong scaling behavior: EN-YUE ($r = 0.75, p = 0.05$), ZH-YUE ($r = 0.75, p = 0.05$), and YUE-YUE ($r = 0.76, p = 0.05$). These results indicate that increasing parameter capacity effectively enhances the model's ability to handle cross-lingual mapping and contextual reasoning in Cantonese, validating that the "scaling law" holds true for pragmatic understanding even in lower-resource scenarios.

Persistent Structural Bottlenecks. However, despite the gains from scaling, the relative performance stratification among pragmatic phenomena remains remarkably consistent. As visualized in Figure 3, while the performance curves trend upward, the significant gap between the stronger tasks (Implicature/Reference) and the weaker task (Presupposition) is preserved from 0.5B up to 72B. This observation supports a critical nuance: while scaling improves general capability, current pragmatic reasoning deficits are *structural* rather than merely *capacity-based*. The inability of even the largest models to close the "Presupposition Gap" suggests that simply scaling parameters under the standard next-token prediction paradigm is insufficient to resolve high-order pragmatic failures (such as maintaining global shared assumptions). Consequently, resolving these specific bottlenecks likely requires paradigm shifts in training objectives or data composition rather than brute-force scaling alone.

6.3 Few-shot Analysis

We further explore the in-context learning capabilities of LLMs by varying the number of demonstration examples $k \in \{2, 4, 8, 16\}$. Figure 4 illustrates the performance scaling of Qwen2.5-7B across different pragmatic tasks and transfer settings. Due to space constraints, full few-shot results for other models are provided in Appendix E.

Scaling with Demonstrations. Consistent with general in-context learning observations, we find a positive correlation between the number of shots and model performance. As the number of examples increases from 2 to 16, accuracy generally improves across all three pragmatic categories (Implicature, Presupposition, and Reference). This trend holds true for both cross-lingual (English/Mandarin → Cantonese) and monolingual (Cantonese → Cantonese) settings, validating that providing explicit pragmatic demonstrations effectively guides the model to align with the target language's reasoning patterns.

Mitigating the Presupposition Bottleneck. Notably, the gains are most pronounced in the Pre-

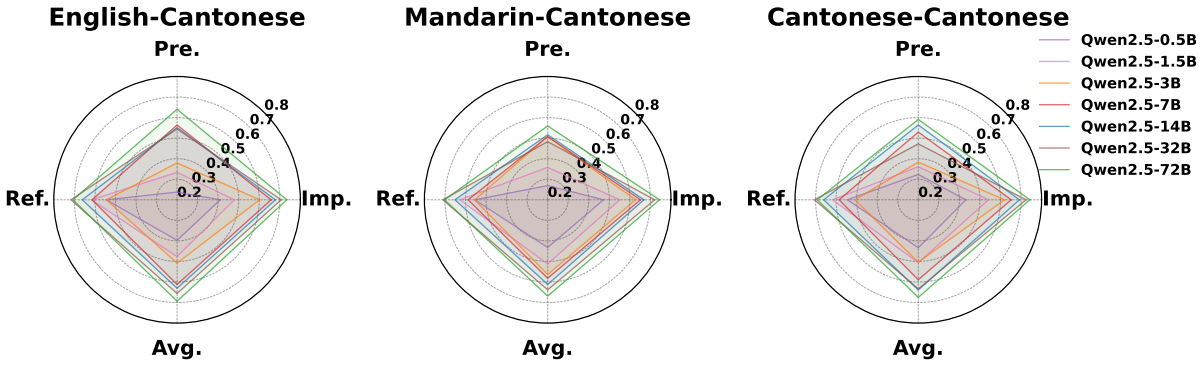


Figure 3: Scaling behavior of Qwen models across varying parameter sizes (0.5B–72B) under the zero-shot topic-aware setting. The trends demonstrate a robust positive correlation between model size and performance across all transfer settings.

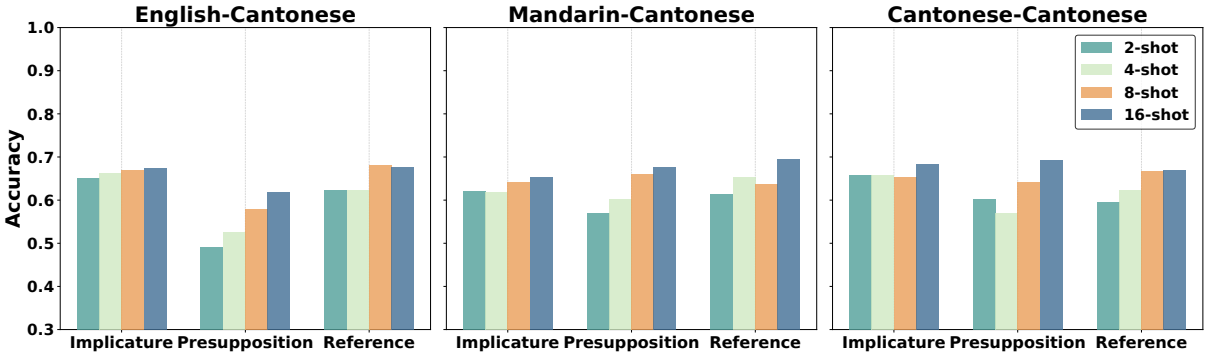


Figure 4: Performance scaling of Qwen2.5-7B under varying numbers of few-shot demonstrations ($k \in \{2, 4, 8, 16\}$). The results demonstrate a overall consistent performance improvement as context increases (EN-YUE: $r = 0.92, p = 0.07$, ZH-YUE: $r = 0.97, p = 0.03$, YUE-YUE: $r = 0.97, p = 0.03$).

supposition task, particularly in cross-lingual settings. As shown in the Mandarin-Cantonese and English-Cantonese panels of Figure 4, Presupposition accuracy sees a substantial boost at 16-shot compared to 2-shot. This suggests that while zero-shot transfer struggles with the "shared background assumptions" required for presupposition (as discussed in Section 6.1), providing ample in-context examples helps bridge this gap, allowing the model to better infer the implicit preconditions necessary for correct pragmatic judgment.

6.4 Case Studies

To better understand the failure modes of LLMs in Cantonese pragmatic reasoning, we conduct a qualitative analysis of error patterns across different instruction languages. We categorize the observed errors into three primary types: (1) Reasoning-Answer Inconsistency, (2) Cultural and Metaphorical Misinterpretation, and (3) Pragmatic Nuance Failure. Figure 5 visualizes the examples.

Reasoning-Answer Inconsistency. As shown in Case 1 (Task 1), we frequently observe a disconnect between the CoT reasoning and the final selection. In the English-to-Cantonese setting, the model correctly analyzes the response as a "direct statement" but inexplicably selects "Indirect Answer" as the final prediction. This phenomenon suggests that while the model possesses the linguistic capability to analyze the text, the instruction-following alignment, specifically mapping the conclusion to a specific label symbol, remains brittle in cross-lingual low-resource settings.

Cultural Misinterpretation in Cross-Lingual Prompts. Case 2 (Task 6) illustrates the risks of transferring pragmatic norms from Mandarin to Cantonese. The phrase "useful than a bank account" is a specific Cantonese metaphor expressing strong agreement or high value. However, the model prompted in Mandarin (Zh $\rightarrow$ Yue) incorrectly hallucinates sarcasm, reasoning that the comparison is absurd or mocking. In contrast, models prompted in Cantonese (Yue $\rightarrow$ Yue) or English correctly identified the positive sentiment. This underscores that *shared script* (Chinese characters)

Case 1: Reasoning-Answer Inconsistency (Task 1: Direct Classification)

Context: X 想知道 Y 喺周末钟意做咩活动。 *(X wants to know Y's weekend activities.)*
Q: 你上个周末有去边度玩呀？ *(Did you go anywhere fun last weekend?)*
A: 我上个周末无去边度玩。 *(I didn't go anywhere last weekend.)*
Gold Label: A
Model Input (En → Yue): ... Select the most appropriate answer...選項:A.直接回答 B.間接回答 *(Options: A. Direct Answer, B. Indirect Answer)*
Model Output:
- *Reasoning:* The question asks about activity... the answer is a **direct statement** about not going anywhere. ✓(Correct Reasoning)
- *Prediction:* **Option B: 間接回答 (Indirect Answer)** ✗(Wrong Selection)

Analysis: The model correctly analyzes the linguistic structure in the CoT but fails to map the conclusion to the correct option symbol, indicating a misalignment in instruction following.

Case 2: Cultural Misinterpretation (Task 6: Sarcasm Detection)

講者一：佢好有用㗎。 *(S1: He is very useful.)*
講者二: 梗係啦，有用過銀行戶口啦。 *(S2: Of course, more useful than a bank account.)*
Gold Label: 同意 (Agreement)
Model Input (Zh → Yue): … 分析句子中的含义 … A.同意 B.諷刺 *(Analyze the implicature... A. Agreement B. Sarcasm)*
Model Output:
- *Reasoning:* 讲者二将“有用”与“银行户口”进行比较，暗示…毫无意义。 (Comparing ”useful” to ”bank account” implies... meaningless.) ✗(Hallucinated Sarcasm)
- *Prediction:* **Option B: 諷刺 (Sarcasm)**

Analysis: The model (prompted in Mandarin) fails to recognize the Cantonese colloquial metaphor where ”useful than a bank account” is a positive intensifier, instead interpreting it as a sarcastic comparison.

Case 3: Loss of Pragmatic Nuance (Task 3: Implicature)

Context: Y啱啱從另一個城市過嚟見X *(Y has just come from another city to see X.)*
X: 你想唔想摸吓我隻狗？ *(X: Want to pet my dog?)*
Y: 我有敏感。 *(Y: I have allergies.)*
Claimed Implicature: 佢唔鍾意摸狗，因為佢有敏感。 *(He dislikes petting dogs because he has allergies.)*
Gold Label: 唔係 (No) *[Implicature is false; allergy ≠ dislike]*
Model Input (Yue → Yue): … 選項:A.係 B.唔係 C.係，但要視乎情況 D.中立… *(... Options: A. Yes B. No C. Yes, conditional D. Neutral...)*
Model Output:
- *Reasoning:* 言外之意是Y不喜欢摸狗… (The implied meaning is Y doesn't like dogs...) ✗(Literal Association)
- *Prediction:* **Option A: 係 (Yes)**

Contrast (Zh → Yue): The Mandarin-prompted model correctly reasoned: ”Y answered having allergies... does not necessarily mean Y dislikes dogs.” → Option B (Correct).

Figure 5: Common error cases in YUE-PUB.

does not guarantee *shared pragmatics*; distinct cultural metaphors require specific native alignment.

Surface-Level vs. Deep Pragmatic Reasoning. Case 3 (Task 3) reveals the challenge of distinguishing between literal constraints and implied preferences. The native Cantonese model (Yue → Yue) fails to separate “having allergies” (a physical fact) from “disliking dogs” (a psychological stance), accepting the flawed implicature as true. Interestingly, the model prompted in Mandarin (Zh → Yue) successfully disentangled these concepts, reasoning that an allergy does not equate to dislike. This suggests that general-purpose multilingual models (often stronger in Mandarin) may sometimes possess stronger reasoning backbones that can override native surface-level associations, provided the cultural gap (as seen in Case 2) does not interfere.

7 Conclusion

We introduce **YUE-PUB**, the first comprehensive benchmark for Cantonese pragmatic understanding, filling a critical gap in low-resource evaluation. Our experiments reveal that while linguistic proximity significantly facilitates cross-lingual transfer (e.g., Zh-to-Yue), current models still face structural limitations in handling complex pragmatic phenomena.We hope YUE-PUB serves as resource to advance Cantonese pragmatic research in NLP.

Limitations

We acknowledge that this work still has the following limitations:

Translation-Induced Bias in Dataset Construction. Our Cantonese pragmatics evaluation data are derived from English sources via machine translation, which may introduce systematic biases. Despite careful post-processing, translation artifacts such as altered implicatures, weakened pragmatic cues, or unnatural discourse patterns may affect the fidelity of pragmatic phenomena in Cantonese. This limitation highlights the need for future benchmarks constructed natively in Cantonese to better reflect authentic pragmatic usage.

Restricted Evaluation Settings and Prompting Strategies. The current evaluation primarily focuses on zero-shot and few-shot settings, including topic-aware prompting and chain-of-thought reasoning. While these settings are representative of common LLM usage scenarios, they do not exhaust the space of possible inference paradigms. Future work could explore additional prompting strategies such as self-consistency, multi-agent reasoning, or iterative refinement to provide a more comprehensive assessment of pragmatic understanding.

Limited Model Coverage Due to Computational Cost. Due to computational and cost constraints, our study primarily evaluates open-source large language models. While this design choice facilitates reproducibility and transparency, it limits the coverage of commercially deployed closed-source models. Extending the evaluation to proprietary models would provide a more complete view of current LLM capabilities in Cantonese pragmatics.

Limited Evaluation Metrics. Our quantitative evaluation mainly relies on accuracy over multiple-choice questions. Although accuracy offers a clear and interpretable measure, it may not fully capture the nuanced nature of pragmatic reasoning, especially when multiple interpretations are plausible. More fine-grained metrics such as confidence-aware scoring, partial credit for reasoning consistency, or explanation-based evaluation could yield deeper insights into model behavior.

Ethical Considerations

Dataset Source. We select the English dataset PUB (Sravanthi et al., 2024), and extend the dataset using machine translation and manual verification. We ensure that this new dataset is intended solely for research purposes and should not be used for commercial purposes. Furthermore, the dataset construction strictly adheres to the intellectual property and privacy protection requirements of the original authors and is freely available for download from their official website.

Human Subjects for Data Evaluation. Before the evaluation process, we fully informed the evaluators about the nature and goals of the task and obtained their informed consent. All evaluators are project partners and voluntarily joined as contributors and co-authors of the paper. Additionally, all evaluators provided explicit consent for the use of the collected data. We did not share any Personally Identifiable Data with this paper. The project itself does not raise any risks.

Use of AI Assistants. The authors acknowledge the use of ChatGPT solely for correcting grammatical errors, enhancing the coherence of the final manuscript, and providing assistance with coding.

Acknowledgments

This work is in part supported by the funds provided by the National Science Foundation and by DoD OUSD (R&E) under Cooperative Agreement PHY-2229929 (The NSF AI Institute for Artificial and Natural Intelligence). The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the National Science Foundation or the U.S. Government.

References

AI@Meta. 2024. Llama 3.1 model card.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst,

Aidan Gomez, Phil Blunsom, Marzieh Fadaee, and 2 others. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Mena Attia, Aashiq Muhamed, Mai Alkhamissi, Thamar Solorio, and Mona T. Diab. 2026. Beyond understanding: Evaluating the pragmatic gap in LLMs' cultural processing of figurative language. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7238–7265, Rabat, Morocco. Association for Computational Linguistics.

Kent Bach. 2006. Christopher potts, the logic of conventional implicatures (oxford studies in theoretical linguistics 7). oxford: Oxford university press, 2005. pp. xii+246. *Journal of Linguistics*, 42:490 – 495.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yiran Chen, Zhenqiao Song, Xianze Wu, Danqing Wang, Jingjing Xu, Jiaze Chen, Hao Zhou, and Lei Li. 2022. MTG: A benchmark suite for multilingual text generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2508–2527, Seattle, United States. Association for Computational Linguistics.

Tsz Chung Cheng, Chung Shing Cheng, Chaak-ming Lau, Eugene Lam, Wong Chun Yat, Hoi On Yu, and Cheuk Hei Chong. 2025a. HKCanto-eval: A benchmark for evaluating Cantonese language understanding and cultural comprehension in LLMs. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 1–11, Vienna, Austria. Association for Computational Linguistics.

Tsz Chung Cheng, Chung Shing Cheng, Chaak-ming Lau, Eugene Lam, Wong Chun Yat, Hoi On Yu, and Cheuk Hei Chong. 2025b. HKCanto-eval: A benchmark for evaluating Cantonese language understanding and cultural comprehension in LLMs. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 1–11, Vienna, Austria. Association for Computational Linguistics.

Zina Chkirbene, Ridha Hamila, Ala Gouissem, and Unal Devrim. 2024. Large language models (llm) in industry: A survey of applications, challenges, and trends. In *2024 IEEE 21st International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET)*, pages 229–234.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, and 69 others. 2024. Deepseek llm: Scaling open-source language models with longtermism. *Preprint*, arXiv:2401.02954.

Giuseppe Gallipoli and Luca Cagliero. 2025. It is not a piece of cake for GPT: Explaining textual entailment recognition in the presence of figurative language. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9656–9674, Abu Dhabi, UAE. Association for Computational Linguistics.

H. P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Speech Acts: Syntax and Semantics Volume 3*, pages 41–58. Academic Press, New York.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Yan Huang. 2017. Introduction: What is pragmatics? In *The Oxford Handbook of Pragmatics*. Oxford University Press.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Jiyue Jiang, Pengan Chen, Liheng Chen, Sheng Wang, Qinghang Bao, Lingpeng Kong, Yu Li, and Chuan Wu. 2025. How well do LLMs handle Cantonese? benchmarking Cantonese capabilities of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4464–4505, Albuquerque, New Mexico. Association for Computational Linguistics.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

Yeeun Kim, Youngrok Choi, Eunkyung Choi, JinHwan Choi, Hai Jin Park, and Wonseok Hwang. 2024. Developing a pragmatic benchmark for assessing Korean legal language understanding in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5573–5595, Miami, Florida, USA. Association for Computational Linguistics.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

Annie Louis, Dan Roth, and Filip Radlinski. 2020. "I'd rather just go to bed": Understanding indirect answers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7411–7425, Online. Association for Computational Linguistics.

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. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 8679–8696. Association for Computational Linguistics.

Zhibo Man, Yuanmeng Chen, Yujie Zhang, and Jinan Xu. 2025. DMDTEval: An evaluation and analysis of LLMs on disambiguation in multi-domain translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6065–6082, Suzhou, China. Association for Computational Linguistics.

Junghyun Min, York Hay Ng, Sophia Chan, Helena Shunhua Zhao, and En-Shiun Annie Lee. 2025. CantoNLU: A benchmark for cantonese natural language understanding. *Preprint*, arXiv:2510.20670.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Dojun Park, Jiwoo Lee, Hyeyun Jeong, Seohyun Park, and Sungeun Lee. 2024a. Pragmatic competence evaluation of large language models for the Korean language. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 256–266, Tokyo, Japan. Tokyo University of Foreign Studies.

Dojun Park, Jiwoo Lee, Seohyun Park, Hyeyun Jeong, Youngeun Koo, Soonha Hwang, Seonwoo Park, and Sungeun Lee. 2024b. MultiPragEval: Multilingual pragmatic evaluation of large language models. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 96–119, Miami, Florida, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Graham Thurgood, P.K. Benedict, J.A. Matisoff, and D. Bradley. 1985. *Linguistics of the Sino-Tibetan Area: The State of the Art : Papers Presented to Paul K. Benedict for His 71st Birthday*. Pacific linguistics. Department of Linguistics, Research School of Pacific Studies, the Australian National University.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965. Curran Associates, Inc.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2717–2739. Association for Computational Linguistics.

Tongguan Wang, Mingmin Wu, Guixin Su, Dongyu Su, Yuxue Hu, Zhongqiang Huang, and Ying Sha. 2025. Mchirc: A multimodal benchmark for chinese idiom reading comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):25398–25406.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Yajie Wen, Ziwei Gong, Chengyan Wu, Xiyun Gong, Yun Xue, Julia Hirschberg, and Bolei Ma. 2026. YUE-PUB-Speech: A Speech-based Pragmatic Understanding Benchmark for Cantonese. In *Interspeech 2026*, pages 4679–4684.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

ChengYan Wu, Bolei Ma, Yihong Liu, Zheyu Zhang, Ningyuan Deng, Yanshu Li, Baolan Chen, Yi Zhang, Yun Xue, and Barbara Plank. 2025. M-ABSA: A multilingual dataset for aspect-based sentiment analysis. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2530–2557, Suzhou, China. Association for Computational Linguistics.

Shengguang Wu, Shusheng Yang, Zhenglun Chen, and Qi Su. 2024. Rethinking pragmatics in large language models: Towards open-ended evaluation and preference tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22583–22599, Miami, Florida, USA. Association for Computational Linguistics.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, and 11 others. 2024. Yi: Open foundation models by 01.ai. *CoRR*, abs/2403.04652.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. 2021. GRICE: A grammar-based dataset for recovering implicature and conversational rEasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2074–2085, Online. Association for Computational Linguistics.

A Prompt Templates

This section details all prompt templates used in this work. We categorize them into two groups:

1. **Data Construction Prompts:** Used for translating the raw English data into Cantonese (Figure 7).

2. **Experimental Prompts:** Used for evaluating LLMs under different settings, including Zero-shot, Chain-of-Thought (CoT), and Few-shot scenarios across three source languages (English, Mandarin, and Cantonese). These are illustrated in Figures 8 through 10.

Note: In the experimental prompts, $\{\text{text}\}$ refers to the context or query, and $\{\text{options}\}$ refers to the candidate choices. For few-shot settings, $\{\text{k-shot samples}\}$ represents the in-context demonstrations. Besides, in the cross-lingual experiment setting, the cross-lingual refers to the case where the instruction language in the prompt differs from the language of the input test material.

B Human Evaluation of Translation Quality

Figure 6 presents a screenshot of the excel-based annotation interface used for rating translated sentences. The interface displays the original sentence, its translation, and four evaluation criteria, each assessed on a five-point Likert scale. The overall workflow is as follows:

Human Evaluation Setup. We conducted a human evaluation to assess the pragmatic quality of Cantonese translations. Each evaluation instance consisted of (i) an English source utterance and (ii) its corresponding Cantonese translation. Evaluators were asked to judge the translation quality along four dimensions: Grammar, Fluency, Adequacy, and Idiomaticity.

Instructions to Evaluators. Before the evaluation, all evaluators were provided with detailed written instructions. Specifically, they were asked to:

- Read the dialogue context carefully to understand the pragmatic background of the utterance.
- Compare the Cantonese translation against the original English utterance.
- Assign scores independently for each dimension.

Evaluation Process and Time Cost. All evaluators were native Cantonese speakers with formal training in linguistics and multilingual language processing, including prior experience in Cantonese translation and pragmatic analysis. Each evaluator independently annotated the same evaluation set. On average, completing the evaluation required approximately 9.33 hours, corresponding to roughly 40 minutes per task, ensuring sufficient time for careful reading and judgment.

Aggregation and Analysis. After collecting the annotations, we aggregated the scores by averaging across evaluators for each dimension. We then report the mean scores for Grammar, Fluency, Adequacy, and Idiomaticity in the main results.

Human Evaluation Agreement Statistics The 420 evaluation instances were sampled across all pragmatic tasks to ensure coverage of different pragmatic phenomena and contextual difficulty levels.

To measure annotation consistency, we further computed inter-annotator agreement statistics using Spearman correlation and Cohen's weighted kappa. The results shown in Table 5 indicate substantial agreement across all four evaluation dimensions, demonstrating the reliability of the human evaluation process.

Dimension	Cohen's κ	Spearman ρ
Grammar	0.81	0.88
Fluency	0.83	0.90
Adequacy	0.79	0.86
Idiomaticity	0.77	0.84
Average	0.80	0.87

Table 5: Inter-annotator agreement statistics for human evaluation.

C Implementation Details

All experiments were run on a single NVIDIA A100 (80GB) GPU. The models are open-weight and directly downloaded from the Huggingface Transformers (Wolf et al., 2020). We disable sampling and use greedy decoding to ensure deterministic outputs for all open LLMs by setting do_sample=False. All generations are capped at 512 tokens.

D Additional Results – Cross-lingual (EN/ZH/YUE→YUE) vs. EN→EN

In the main text, we primarily analyzed the cross-lingual transfer capabilities using Qwen 2.5-7B as the representative model. In this section, we provide the complete performance breakdown for the remaining seven evaluated models: Mistral-7B, Gemma-2-9B, Llama-3.1-8B, DeepSeek-7B, CantoneseLLM, Yi-1.5-9B, and Aya-23-8B.

Tables 6 through 12 detail the performance across three pragmatic categories (Implicature, Presupposition, Reference) under English-to-Cantonese (En→Yue), Mandarin-to-Cantonese (Zh→Yue), and Native Cantonese (Yue→Yue) settings, juxtaposed with their respective English-to-English (En→En) baselines.

E Additional Results – Few-shot

In addition to the Qwen2.5-7B results discussed in Section 6.3, this appendix presents the detailed few-shot scaling behaviors ($k \in \{2, 4, 8, 16\}$) for the remaining seven evaluated models. Figures 11 through 17 illustrate the performance trajectories across English-to-Cantonese (EN-YUE), Mandarin-

Translated (YUE)	Original (EN)	Grammar	Fluency	Adequacy	Idiomatic
背景：Y啱啱同X講話，話緊諗住喺紐約買樓。 問：你係一個城市人嚟㗎？ 答：我鍾意住喺人群多嘅地方。	Context: Y has just told X that he/she is thinking of buying a flat in New York. Question: Are you a city person? Response: I enjoy living around lots of people.	4	4	5	4

Figure 6: Screenshot of the excel-based human evaluation interface.

to-Cantonese (ZH-YUE), and Native Cantonese (Yue-YUE) settings.

While some models like Mistral and Gemma-2 exhibit strong positive scaling consistent with Qwen, others display more volatile behaviors or diminishing returns, highlighting the variability in in-context learning capabilities across different model architectures in low-resource settings.

Setting	Imp.	Pre.	Ref.	Avg.
En → Yue	55.96	35.85	42.69	44.83
Zh → Yue	57.15	43.97	48.82	49.98
Yue → Yue	58.58	40.82	47.12	48.84
Reference Baseline (English Source Data)				
En → En	62.13	55.08	49.68	55.63

Table 6: Cross-lingual transfer performance for Mistral-7B.

Setting	Imp.	Pre.	Ref.	Avg.
En → Yue	68.96	56.47	66.26	63.90
Zh → Yue	68.05	54.48	63.76	62.10
Yue → Yue	68.65	53.49	62.38	61.51
Reference Baseline (English Source Data)				
En → En	71.72	50.74	63.29	60.34

Table 7: Cross-lingual transfer performance for Gemma-2-9B.

Setting	Imp.	Pre.	Ref.	Avg.
En → Yue	57.75	59.21	63.29	60.08
Zh → Yue	60.08	59.85	64.11	61.35
Yue → Yue	65.40	54.49	63.99	61.29
Reference Baseline (English Source Data)				
En → En	64.54	65.47	60.59	63.53

Table 8: Cross-lingual transfer performance for Llama-3.1-8B.

Setting	Imp.	Pre.	Ref.	Avg.
En → Yue	60.85	49.60	53.15	54.53
Zh → Yue	59.07	41.60	51.05	50.57
Yue → Yue	58.22	44.85	53.75	52.27
Reference Baseline (English Source Data)				
En → En	62.81	42.62	51.34	52.26

Table 9: Cross-lingual transfer performance for DeepSeek-7B.

Setting	Imp.	Pre.	Ref.	Avg.
En → Yue	62.02	58.32	64.60	61.64
Zh → Yue	64.60	61.35	67.57	64.51
Yue → Yue	64.93	62.81	65.66	64.47
Reference Baseline (English Source Data)				
En → En	67.47	62.45	62.88	64.27

Table 10: Cross-lingual transfer performance for CantoneseLLM.

F Additional Case Studies

We have added a focused contrastive analysis in this section, including representative examples of where English-to-Cantonese pragmatic transfer succeeds, as well as cases where it partially or completely fails, alongside brief linguistic explanations.

Case 1: Successful Transfer / Pragmatic Weakening Figure 18 illustrates a scenario where the English statement "I need to get up very early tomorrow morning" carries the implicature of a polite rejection (i.e., "No" or "Probably not"). In Cantonese, stating only the need to wake up early, without explicitly claiming to have other plans, leaves significant room for interpretation. The requester might counter with, "Then can we come back earlier?" in an attempt to salvage the plan. This demonstrates a weakening of pragmatic force during the transfer:

Setting	Imp.	Pre.	Ref.	Avg.
En → Yue	63.83	62.08	57.98	61.30
Zh → Yue	65.79	65.07	61.92	64.26
Yue → Yue	61.53	66.25	56.44	61.41
Reference Baseline (English Source Data)				
En → En	69.69	46.92	64.38	60.33

Table 11: Cross-lingual transfer performance for Yi-1.5-9B.

Setting	Imp.	Pre.	Ref.	Avg.
En → Yue	64.46	60.35	60.31	61.71
Zh → Yue	65.88	51.12	61.49	59.50
Yue → Yue	63.46	53.41	60.59	59.15
Reference Baseline (English Source Data)				
En → En	66.36	64.85	63.40	64.87

Table 12: Cross-lingual transfer performance for Aya-23-8B.

rather than acting as an absolute refusal, the statement functions more as a negotiation of conditions.

Case 2: Successful Transfer / Polite Refusal Figure 19 demonstrates a highly successful transfer. Listeners in both English and Cantonese automatically infer the underlying meaning that attending the event is "infeasible." This indirect approach functions as a mutual face-saving strategy, serving as a much more polite alternative to a direct refusal (e.g., "I'm not going").

Case 3: Failed Transfer / Polite Evasion vs. Literal Interpretation Figure 20 illustrates a failed transfer. In English, the phrase "Maybe I will someday" typically serves as a polite evasion, carrying the implicature "I haven't read it and probably won't." Conversely, in Cantonese, this statement is much more likely to be interpreted literally, suggesting a genuine possibility that the speaker will read it in the future.

Case 4: Failed Transfer / Indirect Negative vs. Literal Willingness Figure 21 presents another failed transfer. While the English response implies an indirect negative, uncertainty, or that the speaker is "probably not a fan," the Cantonese interpretation defaults to a literal willingness to try. The strong English implicature of "not really" does not effectively transfer over to Cantonese in this context.

Context Translation

请将以下这段英文所有内容都翻译成粤语，句子如下：{text}。只需要输出翻译后的内容，不需要分析。

Options Translation

请将以下这段英文所有内容都翻译成粤语，句子如下：{text}。按照原列表格式输出，不需要分析。

Back Translation (Quality Check)

请将以下这段粤语繁体字的所有内容都翻译成英文，句子如下：{text}。
只需要输出翻译后的完整内容，不需要分析。

Figure 7: Prompt templates used for data construction and quality verification.

English to Cantonese (Zero-shot)

You are now working on a task about pragmatic phenomena, with the topic being 'Implicature (Direct classification)'. Based on the given sentence, please select the most appropriate answer from the provided options. The sentence is as follows: {text}. The options are as follows:{options}. ###Note: Only output the answer, no analysis is required.

Mandarin to Cantonese (Zero-shot)

你现在在做关于语用使用现象的题目，主题是'含义（直接分类）',请你根据给定的句子，从给定选项中选出最正确的答案。 句子如下:{text}。 选项如下:{options}。 ###注意，只需要输出答案，不需要进行分析。

Cantonese to Cantonese (Zero-shot)

你而家喺做一条同语用使用现象有关嘅题目，主题係'含意（直接分類）'。请你根据畀你嘅句子，喺提供嘅选项入面揀出最正确嘅答案。 句子如下:{text}。 選項如下：{options}。 ###留意，只需要輸出答案，唔使分析。

Figure 8: Prompt templates for Zero-shot evaluation across three source languages.

English to Cantonese (CoT)

You are now working on a task about pragmatic phenomena, with the topic being 'Implicature (Direct classification)'. Based on the given sentence, please select the most appropriate answer from the provided options.
The sentence is as follows: {text}. The options are as follows: {options}.
###Note: Please provide your response strictly in the following JSON format with a two-step CoT structure: {'Step 1': '<analysis process>', 'Step 2': '<final answer>'}. Step 2 only needs to output the final option.

Mandarin to Cantonese (CoT)

你现在在做关于语用使用现象的题目，主题是'含义（直接分类）',请你根据给定的句子逐步分析，然后从给定选项中选出最正确的答案。
句子如下:{text}。选项如下:{options}。
###注意，请严格按照以下 JSON 格式并采用两步推理（CoT）结构来作答：{'Step 1': '<分析过程>', 'Step 2': '<最终答案>'}。第二步只需要输出最终选项即可。

Cantonese to Cantonese (CoT)

你而家喺做一条同语用使用现象有关嘅题目，主题係'含意（直接分類）'。请你根据畀你嘅句子，喺提供嘅选项入面揀出最正确嘅答案。
句子如下:{text}。 選項如下：{options}。
###注意，请严格跟从以下 JSON 格式，并用两个步骤嘅推理（CoT）结构作答：{'Step 1': '<分析过程>', 'Step 2': '<最终答案>'}。第二步只需要输出最后嘅选项即可。

Figure 9: Prompt templates for Chain-of-Thought (CoT) evaluation.

English to Cantonese (Few-shot)

You are now working on a task about pragmatic phenomena, with the topic being 'Implicature (Direct classification)'. Based on the given sentence, please select the most appropriate answer from the provided options.
###Example are as follows: {k-shot samples}.
The sentence is as follows: {text}. The options are as follows:{options}. ###Note: Only output the answer, no analysis is required.

Mandarin to Cantonese (Few-shot)

你现在在做关于语用使用现象的题目，主题是'含义（直接分类）',请你根据给定的句子，从给定选项中选出最正确的答案。
###示例如下: {k-shot samples}.
句子如下: {text}. 选项如下:{options}. ###注意，只需要输出答案，不需要进行分析。

Cantonese to Cantonese (Few-shot)

你而家喺做一条同语用使用现象有关嘅题目，主题係'含意（直接分類）'。请你根据畀你嘅句子，喺提供嘅选项入面揀出最正确嘅答案。
###例子如下: {k-shot samples}.
句子如下: {text}. 選項如下:{options}. ###留意，只需要輸出答案，唔使分析。

Figure 10: Prompt templates for Few-shot evaluation.

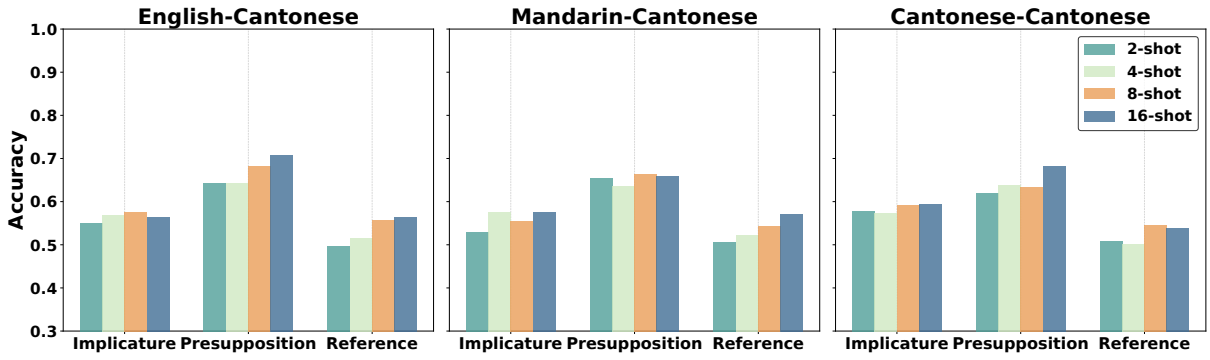


Figure 11: Performance scaling of **Mistral-7B**. The results demonstrate a consistent performance improvement as context increases (EN-YUE: $r = 0.91, p = 0.09$; ZH-YUE: $r = 0.96, p = 0.04$; YUE-YUE: $r = 0.98, p = 0.02$).

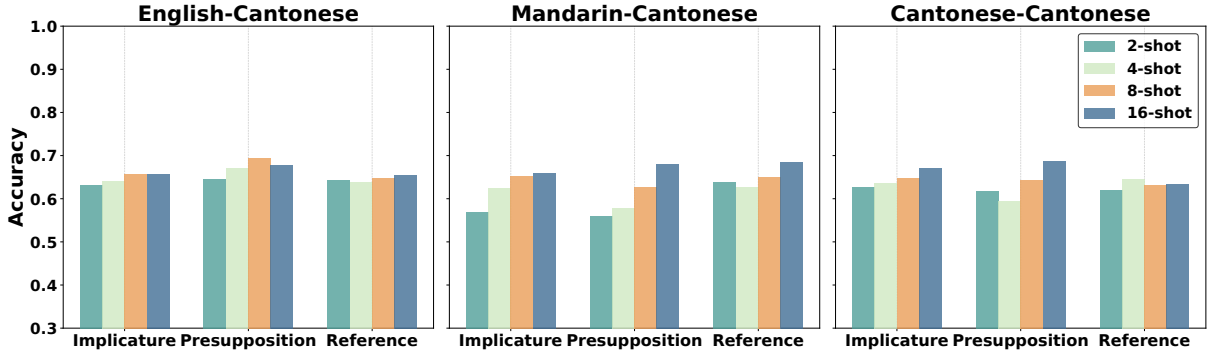


Figure 12: Performance scaling of **Gemma-2-9B**. The results demonstrate strong positive scaling, particularly in native and Mandarin transfer settings (EN-YUE: $r = 0.77, p = 0.13$; ZH-YUE: $r = 0.97, p = 0.02$; YUE-YUE: $r = 0.99, p = 0.004$).

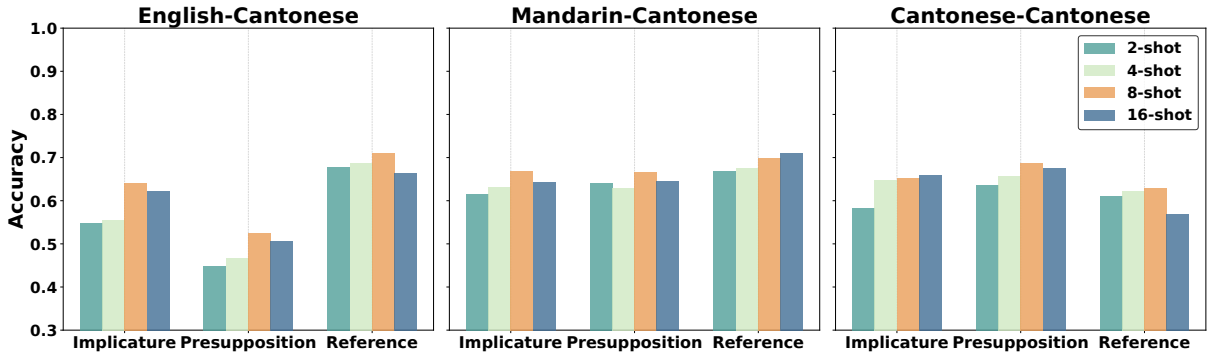


Figure 13: Performance scaling of **Llama-3.1-8B**. Unlike other models, Llama-3.1 shows mixed or negative scaling trends in certain settings (EN-YUE: $r = 0.52, p = 0.30$; ZH-YUE: $r = -0.89, p = 0.10$; YUE-YUE: $r = -0.77, p = 0.22$).

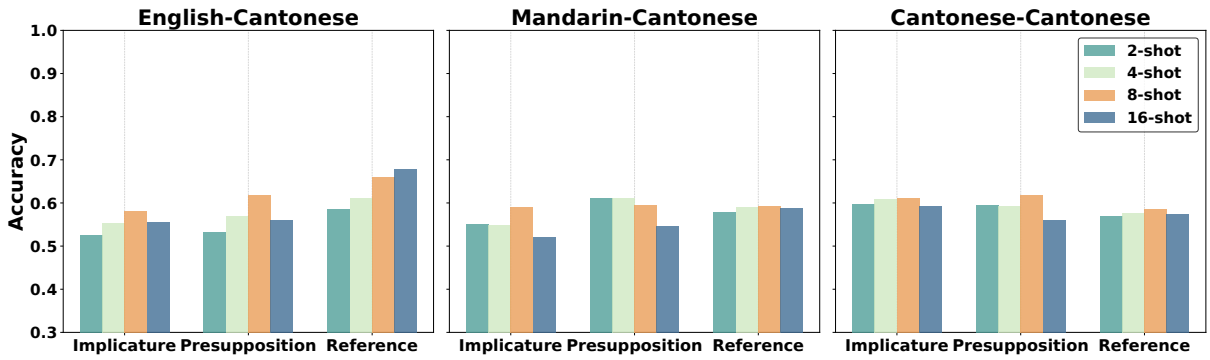


Figure 14: Performance scaling of **DeepSeek-7B**. The model exhibits volatility across different source languages (EN-YUE: $r = 0.52, p = 0.03$; ZH-YUE: $r = -0.92, p = 0.07$; YUE-YUE: $r = 0.77, p = 0.03$).

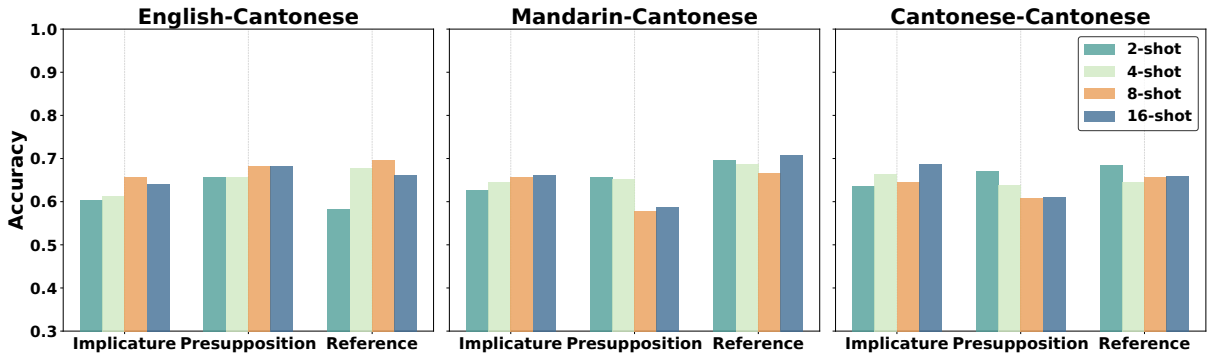


Figure 15: Performance scaling of **CantoneseLLM (CLLM)**. As a specialized model, its few-shot gains are relatively flat or fluctuating compared to generalist models (EN-YUE: $r = 0.58, p = 0.40$; ZH-YUE: $r = -0.54, p = 0.40$; YUE-YUE: $r = -0.28, p = 0.60$).

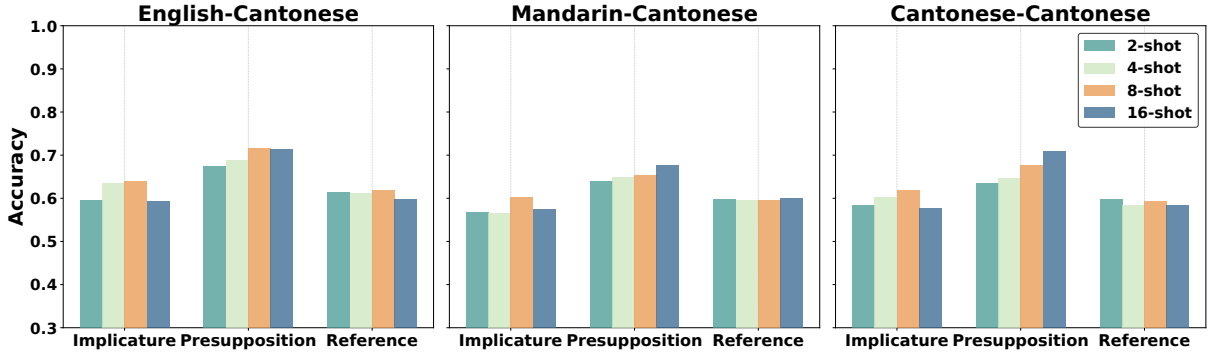


Figure 16: Performance scaling of **Yi-1.5-9B**. The statistical analysis indicates weak or negative correlations in this setting (EN-YUE: $r = 0.58, p = 0.40$; ZH-YUE: $r = -0.54, p = 0.40$; YUE-YUE: $r = -0.287, p = 0.60$).

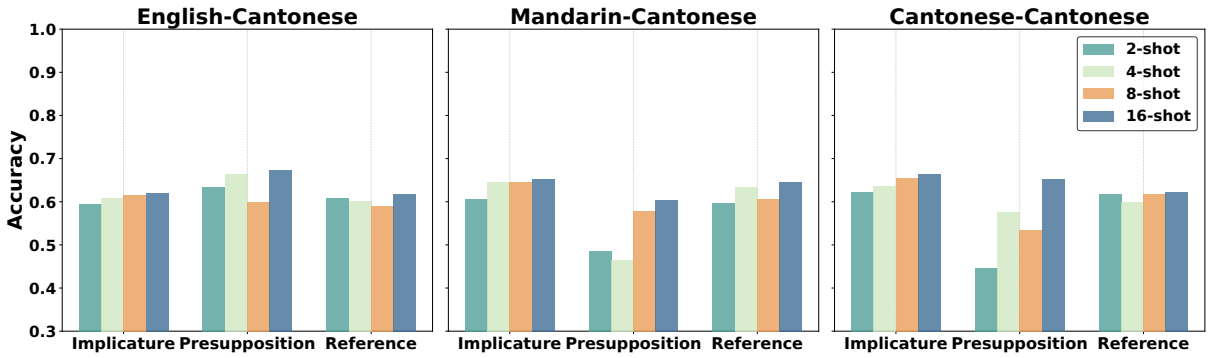


Figure 17: Performance scaling of **Aya-23-8B**. The model shows generally positive scaling trends across all settings (EN-YUE: $r = 0.72, p = 0.28$; ZH-YUE: $r = 0.60, p = 0.23$; YUE-YUE: $r = 0.86, p = 0.10$).

YUE

{"pretext": "背景：Y 啱啱由另一個城市嚟搵 X。問：今晚想唔想出去玩吓？答：我聽日朝早要好早起身。", "options": ["直接回答", "間接回答"], "correct answer": "間接回答", "id": "68"}

EN

{”pretext”: ”Context: Y has just travelled from a different city to meet X. Question: Do you want to go partying tonight? Response: I need to get up very early tomorrow morning.”, ”options”: [”Direct answer”, ”Indirect answer”], ”correct answer”: ”Indirect answer”, ”id”: ”68”}

Figure 18: Successful case 1 for English-to-Cantonese pragmatic transfer.

YUE

{"pretext": "背景：X 同 Y 係星期五同一時間收工嘅同事。問：想唔想出去飲杯咖啡/飲啲嘢？答：我老婆五分鐘後會嚟接我。", "options": ["直接回答", "間接回答"], "correct answer": "間接回答", "id": "67"}

EN

{”pretext”: ”Context: X and Y are colleagues who are leaving work on a Friday at the same time. Question: Want to grab a coffee/drink? Response: My wife is expecting me in five minutes.”, ”options”: [”Direct answer”, ”Indirect answer”], ”correct answer”: ”Indirect answer”, ”id”: ”67”}

Figure 19: Successful case 2 for English-to-Cantonese pragmatic transfer.

YUE

{"pretext": "背景：X 想知道 Y 喜欢睇边啲书。问：你睇过《教育论》未呀？答：可能我迟啲会睇。", "options": ["直接回答", "間接回答"], "correct answer": "間接回答", "id": "196"}

EN

{”pretext”: ”Context: X wants to know what sorts of books Y likes to read. Question: Have you ever read The Education? Response: Maybe I will someday.”, ”options”: [”Direct answer”, ”Indirect answer”], ”correct answer”: ”Indirect answer”, ”id”: ”196”}

Figure 20: Failed case 1 for English-to-Cantonese pragmatic transfer.

YUE

{"pretext": "背景: X 想知道 Y 嘅飲食偏好。問: 你鍾唔鍾意食辣嘅嘢？答: 我試吓。", "options": ["直接回答", "間接回答"], "correct answer": "間接回答", "id": "2067"}

EN

{”pretext”: ”Context: X wants to know about Y’s food preferences. Question: Do you like spicy/hot food? Response: I’ll try it.”, ”options”: [”Direct answer”, ”Indirect answer”], ”correct answer”: ”Indirect answer”, ”id”: ”2067”}

Figure 21: Failed case 2 for English-to-Cantonese pragmatic transfer.