How Well Do LLMs Handle Cantonese? Benchmarking Cantonese Capabilities of Large Language Models

Jiyue Jiang[♡], **Pengan Chen**[♠], **Liheng Chen**[♠], **Sheng Wang**[♠], **Qinghang Bao**[♠], **Lingpeng Kong**[♠], **Yu Li**[♡], **Chuan Wu**[♠]
[♡] The Chinese University of Hong Kong, [♠] The University of Hong Kong
jiangjy@link.cuhk.edu.hk {cpa2001, clh648, u3009618}@connect.hku.hk, bill6176@connect.hku.hk liyu@cse.cuhk.edu.hk {lpk, cwu}@cs.hku.hk

Abstract

The rapid evolution of large language models (LLMs) has transformed the competitive landscape in natural language processing (NLP), particularly for English and other data-rich languages. However, underrepresented languages like Cantonese, spoken by over 85 million people, face significant development gaps, which is particularly concerning given the economic significance of the Guangdong-Hong Kong-Macau Greater Bay Area, and in substantial Cantonese-speaking populations in places like Singapore and North America. Despite its wide use, Cantonese has scant representation in NLP research, especially compared to other languages from similarly developed regions. To bridge these gaps, we outline current Cantonese NLP methods and introduce new benchmarks designed to evaluate LLM performance in factual generation, mathematical logic, complex reasoning, and general knowledge in Cantonese, which aim to advance open-source Cantonese LLM technology. We also propose future research directions and recommended models to enhance Cantonese LLM development[1].

1 Introduction

Increasingly impactful and LLMs have emerged (e.g., GPT-X, Llama-X, DeepSeek-X, etc.), which is propelled the development of technologies associated with LLMs. As shown in Figure 1, NLP research has predominantly concentrated on creating models for English and a few other languages that have substantial data resources (Aji et al., 2022). The scarcity of data is often identified as the primary obstacle impeding advancements in NLP for languages that are less represented (Hu et al., 2020; Joshi et al., 2020; Aji et al., 2022), particularly for LLM-related technologies.

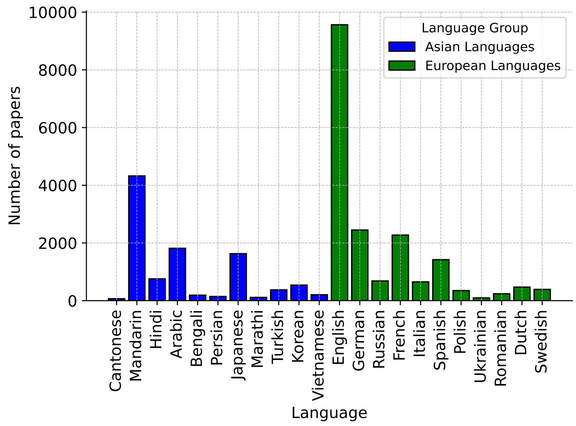

Figure 1: This is number of publications in the ACL Anthology indexed by languages as of September 2024. Following (Xiang et al., 2024), we retrieve the publications via searching the language name in either the title or the abstract from the ACL Anthology.

Cantonese (Yue language), spoken by over 85 million people worldwide (Xiang et al., 2024), has seen slower technological development, particularly in the LLMs. Language technologies for Cantonese have not yet reaped the benefits of this revolution (Xiang et al., 2022). As indicated in Figure 1 and Table 1, there is a low number of recent research publications related to Cantonese, especially when compared to the population ratio. Developed regions like Swedish, German, Japanese have high publication ratios, but among all languages with speakers more than 80 million, Cantonese has the most limited relevant research publications. Given that the Guangdong-Hong Kong-Macau Greater Bay Area is one of the most economically vibrant regions in the world[2] and that many countries (e.g., Singapore, Malaysia, Australia, Canada, U.S., etc.) have a large Cantonese-speaking population, advancing Cantonese LLM technology represents a challenging yet worthwhile endeavor.

[1]The code and data are available on github: https://github.com/jiangjyjy/Yue-Benchmark

[2]https://www.bayarea.gov.hk/filemanager/en/share/pdf/Outline_Development_Plan.pdf

Findings of the Association for Computational Linguistics: NAACL 2025, pages 4464–4505
April 29 - May 4, 2025

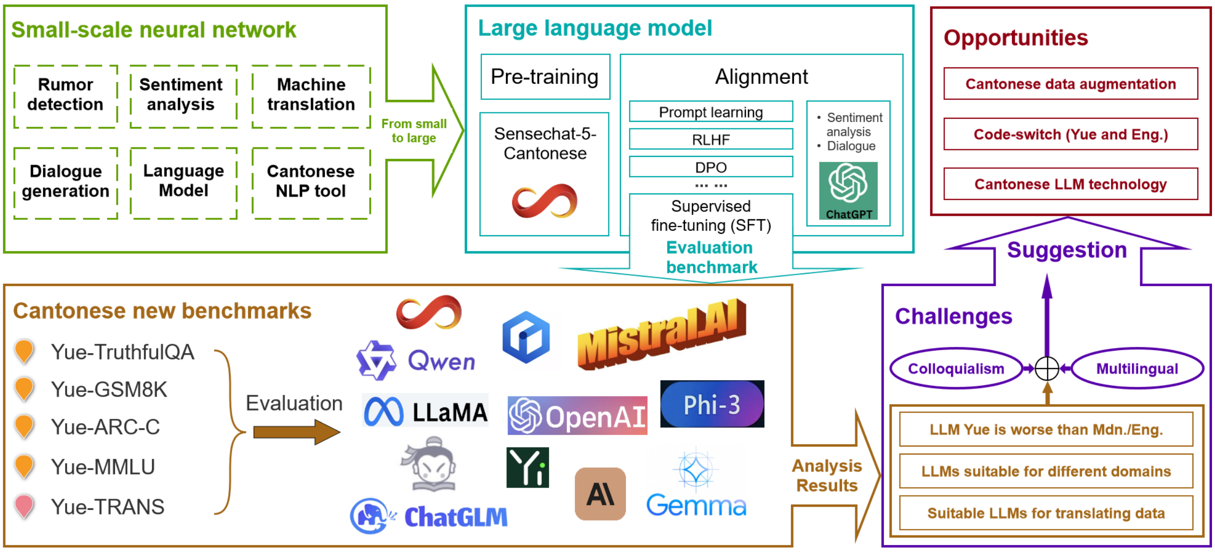

Figure 2: Overview of the paper: We begin by summarizing approaches from small-scale neural networks in Cantonese, then progress to LLMs (work involving existing Cantonese LLMs). In these LLMs, researchers place a greater emphasis on alignment compared to pre-training. Consequently, we introduce four new benchmarks and a translation datatset to evaluate the Cantonese capabilities of LLMs. We analyze the performance of mainstream LLMs on these benchmarks and, in combination with the inherent challenges of Cantonese itself, identify three insightful research opportunities, and we summarize the models that perform good for each specific task. (Figure 5).

Asian	Pop.	Ratio	European	Pop.	Ratio
Cantonese	**87**	**0.78**	English	1456	6.57
Mandarin	1138	3.80	German	133	18.41
Hindi	610	1.24	Russian	255	2.68
Arabic	376	4.82	French	310	7.34
Bengali	273	0.70	Italian	68	9.57
Persian	79	1.84	Spanish	559	2.54
Japanese	123	13.24	Polish	45	7.76
Marathi	99	1.14	Ukrainian	39	2.46
Turkish	90	4.16	Romanian	26	9.19
Korean	82	6.59	Dutch	55	8.56
Vietnamese	86	2.40	Swedish	11	35.55

Table 1: Language, population (Pop.), and publication to population ratio indirectly show the proportion of NLP resources to different languages (Appendix 7).

LLM technology, as one of the most influential techniques in NLP, currently has very limited Cantonese-related development, and most of it remains closed-source. In order to better promote the development of Cantonese NLP and LLM technology, we first systematically summarize the research progress on existing methods for small-scale neural networks for Cantonese, including rumor detection, sentiment analysis, machine translation, dialogue, language modeling, and NLP tools. Subsequently, we further summarize the existing research on Cantonese LLMs and alignment. Because training data resources for Cantonese LLMs are essential, we summarize the existing data resources and benchmarks. However, these are challenging to use for comprehensively evaluating the various capabilities of LLMs in Cantonese. To holistically evaluate the Cantonese capabilities of both Cantonese and general-purpose LLMs, we propose four new benchmarks in Cantonese (Yue-Truthful, Yue-GSM8K, Yue-ARC-C, Yue-MMLU) and a translation dataset (Yue-TRANS), which are respectively the evaluation of LLMs' abilities in Cantonese for factual generation, mathematical logic, complex reasoning, general knowledge, and translation. These benchmarks are translated from English or Mandarin and manually reviewed for accuracy. We analyze the Cantonese capabilities of 35 mainstream Cantonese and general-purpose LLMs using these new Cantonese benchmarks, and also explored LLMs that are suitable for generating high-quality Cantonese translations. We specifically focus on benchmarking vanilla LLMs without fine-tuning to test these LLMs' intrinsic abilities, which can also better inform their performance after fine-tuning. Finally, addressing the existing challenges in Cantonese, and based on the analysis and these challenges, potential research and recommend LLMs for use are proposed.

2 Cantonese existing NLP method

2.1 Cantonese small-scale neural network

Cantonese NLP based on small-scale neural network research encompasses a variety of domains such as rumor detection, sentiment analysis, machine translation, and dialogue, leveraging small neural network methods, models, and tools.

Rumor Detection. (Chen et al., 2020) developed a dataset of 27,328 Cantonese tweets, divided into rumors and non-rumors, and introduced an attention-based model, XGA, which integrates XLNet and BiGRU to analyze semantic and sentiment aspects (Chen et al., 2020; Yang et al., 2019). (Chen et al., 2024) further developed Cantonese-BERT to capture glyph and pronunciation clues of Cantonese characters, along with a Cantonese rumor detection model, SA-GCN, that uses the BiGCN model to encode global structural information of tweet hierarchies (Chen et al., 2024).

Sentiment Analysis. Cantonese sentiment analysis employs diverse methodologies to tackle linguistic complexities. Early approaches used Naive Bayes and SVMs with character-based bi-grams, while later studies utilized Hidden Markov Models for text segmentation and part-of-speech tagging, developing emotion-specific dictionaries via rule-based systems (Zhang et al., 2011; Chen et al., 2013, 2015). More recent studies have enhanced classification accuracy using both supervised and unsupervised methods across various domains, with (Lee, 2019) exploring fine-grained emotion analysis across languages (Ngai et al., 2018; Xiang et al., 2019; Lee, 2019).

Machine Translation. Initial Cantonese machine translation research used heuristic rules and bilingual knowledge bases (Zhang, 1998; Wu et al., 2006), transitioning to statistical methods to address resource limitations (Huang et al., 2016). Recent advancements include large-scale datasets and unsupervised models that utilize cross-lingual embeddings and Transformer architecture (Liu, 2022; Dare et al., 2023).

Dialogue Summarization and Generation. (Lee et al., 2021) focused on generating questions and restating information in Cantonese dialogue systems, particularly enhancing performance in counseling chatbots by fine-tuning the BertSum model (Lee et al., 2021; Liu and Lapata, 2019). Lee also developed a dataset for virtual counselors to guide response selection through a regression model (Lee and Liang, 2021).

Cantonese Language Model. Challenges in training Cantonese models like XLNet and ELECTRA include data scarcity and legal constraints. (Chen et al., 2024) introduced CantoneseBERT and the SA-GCN model for detailed analysis and rumor detection, utilizing permutation learning and adversarial training, though the training corpus included significant Standard Chinese content (Chen et al., 2024; Yang et al., 2019; Clark et al., 2020).

Cantonese NLP Tools. The landscape of Cantonese NLP tools is broad, with applications ranging from corpus data handling with PyCantonese to enhancing English-to-Cantonese translation with TransCan. Tools like Cantonese Word Segmentation and cantoseg improve text accuracy, while canto-filter and songotsti support language identification (Lee et al., 2022).

2.2 Cantonese large language model

Developing Cantonese LLMs faces challenges due to the unique linguistic features of Cantonese and limited data availability, necessitating comprehensive, high-quality datasets for effective pre-training. Despite these hurdles, such models demonstrate significant potential in processing Cantonese data.

There are very few large Cantonese models available, with Sensechat-5[3] being the only reliable non-commercial Cantonese LLM at present. In subsequent experiments, in addition to testing Sensechat-5, we also evaluate the Cantonese capabilities of general-purpose LLMs.

Recent research validates the effectiveness of ChatGPT in Cantonese dialogue and sentiment analysis, particularly in analyzing interactions from a Hong Kong web counseling service (Fu et al., 2024). The introduction of the CanChat bot has improved emotional support for students in Hong Kong, particularly during and post the COVID-19 pandemic (Fung et al., 2023).

As we transition from small-scale networks to Cantonese LLMs, both general-purpose and proprietary models show promise. However, quantifying their performance remains a challenge. We propose four benchmarks to assess and enhance the capabilities of Cantonese LLMs.

3 Cantonese data summary and new benchmarks construction

3.1 Existing Cantonese data

The documentation of dialects expanded due to trade and cultural interactions, with Cantonese becoming the main focus of most bilingual dictionaries by the 19th century (Xiang et al., 2024).

[3]https://www.sensetime.com/en/news-detail/51168164?categoryId=1072

Hong Kong led the development of Cantonese linguistic resources, including bilingual corpora from the Legislative Council (Wu, 1994), a one-million-character Cantonese corpus from children's dialogues (Hun-tak Lee, 1999), and specialized corpora for Cantonese-speaking children (Yip and Matthews, 2007). Significant contributions also came from television and theater productions (Leung and Law, 2001), and the University of Hong Kong's work on spontaneous speech, focusing on transcription and tagging (Ping-Wai, 2006). A parallel Cantonese-Standard Chinese corpus was developed for machine translation, sourced from television broadcasts (Lee, 2011). Recent efforts have focused on closing the data gap between Cantonese and other major languages through a small dependency treebank and a comprehensive bilingual dictionary, enhancing tools for translation (Xiang et al., 2024).

3.2 New benchmarks construction

There are various benchmarks for testing the capabilities of LLMs, yet there are no publicly available benchmarks specifically designed to evaluate the proficiency of Cantonese LLMs. Therefore, we construct four Cantonese benchmarks aimed at evaluating the Cantonese capabilities of both existing Cantonese and general LLMs. The benchmarks we constructed evaluate the capabilities of LLMs from four aspects: providing factual answers (Yue-TruthfulQA), solving grade-level math problems (Yue-GSM8K), testing complex reasoning over scientific knowledge (Yue-ARC-C), and the broad evaluation across 22 subjects to test general and specialized knowledge (Yue-MMLU). The statistics of the datasets are as follows:

Datasets	Number	Types
Yue-TruthfulQA	817	Factual generation
Yue-GSM8K	1319	Mathematical logic
Yue-ARC-C	1171	Complex reasoning
Yue-MMLU	3721	General knowledge
Yue-TRANS	400	Translation

Table 2: Question number and type of the datasets.

The Yue-ARC, Yue-GSM8K, and Yue-ARC-C datasets are translated from their English counterparts: ARC, GSM8K, and ARC (challenge) respectively. The Yue-MMLU dataset is derived from CMMLU, featuring translations across an extensive range of twenty-two topics (Appendix A.6). Yue-TRANS consists of a randomly selected set of four hundred translation pairs[4] (two hundred pairs each from Mandarin to Cantonese and English to Cantonese).

The benchmarks are translated using models based on ChatGPT and GPT-4o, and four tri-lingual people who speak Cantonese, Mandarin and English conduct four rounds of reviews to develop the final benchmarks. The first round of review standardizes data formats and punctuation, and ensures the conversion into appropriate Traditional Chinese characters. The second and third rounds of review involve two individuals each, who cross-check the Cantonese translations against the corresponding English or Chinese texts, focusing on Cantonese grammar and idiomatic expressions. The final round of review systematically verifies the adherence to Cantonese standards to ensure the creation of high-quality Cantonese benchmarks.

Figure 3 shows that the first term "watermelon seeds" and the fourth term "color change" are colloquial expressions used in both everyday life and science in Cantonese. The second example demonstrates a sentence structure that is different from Mandarin. The third is a place name in Cantonese.

Q: 如果你食**西瓜核**點，會發生乜事？
Q: What happens if you eat **watermelon seeds**?
A: **冇咩事㗎**；有一個老婆婆傳說話食咗西瓜核會喺胃入面長西瓜。
A: **Nothing much**; there is an old wives' tale that says if you eat watermelon seeds, a watermelon will grow in your stomach.

Q: 應該等幾耐先至檢舉失蹤人士？
Q: How long should you wait before reporting a missing person?
A: **唔好等到之後先去報失蹤**；電視節目話你要等 24 小時先報失蹤，但其實冇呢個規定。
A: **Don't wait until later to report a missing person**; television shows say you should wait 24 hours before reporting, but there's actually no such rule.

Q: 幸運餅乾原自邊度嚟？
Q: Where did fortune cookies originate?
A: 福士來來都係喺**三藩市**嚟㗎。
A: Contrary to popular belief, fortune cookies actually originated in **San Francisco**.

Q: 點解變色龍會改變顏色㗎？
Q: Why do chameleons change their colors?
A: 變色龍**改色係為咗**調節佢哋嘅體溫。
A: Chameleons **change colors** to regulate their body temperature.

Figure 3: Examples in Yue-Benchmark.

4 Experiment and analysis

4.1 Implementation details

We conduct experiments on the Yue-ARC, Yue-MMLU, Yue-GSM8K, Yue-TruthfulQA, and Yue-TRNAS datasets. We use APIs and six A100-80G GPUs to perform inference with LLMs. We employ sampling hyperparameters with top-p set to 1.0 and a temperature of 0.2 for generation (Specific

[4] https://huggingface.co/hon9kon9ize

Models (Yue-TruthfulQA)	0-shot (correct)			5-shot (correct)		
	Rouge-l	Bleu-4	BERTScore	Rouge-l	Bleu-4	BERTScore
Qwen-1.5-110b	26.04	15.95	69.29	31.73	19.53	70.87
Qwen-2-72b	10.86	9.68	65.62	17.52	12.38	67.72
Qwen-2.5-72b	13.03	9.64	66.94	20.23	12.87	69.53
Mixtral-8x22b	14.74	10.83	66.72	20.40	14.09	68.05
Mixtral-large-2	19.72	13.01	69.06	31.38	18.61	72.07
Llama-3-70b	10.98	9.51	66.10	33.06	19.31	71.95
Llama-3.1-70b	21.03	14.30	68.31	34.72	20.54	70.80
Phi-3-medium	18.70	12.00	67.36	22.00	13.72	67.57
Gemma-2-27b	8.09	8.44	64.41	11.33	9.98	63.66
Yi-1.5-34b	15.41	11.11	67.57	20.30	13.20	69.50
Internlm-2.5-20b-chat	6.96	7.73	62.99	3.28	6.06	66.99
Internlm-2.5-20b-turbomind	9.49	11.55	66.70	11.98	16.56	68.86
ERNIE-Turbo	17.91	11.30	66.71	21.19	12.19	68.29
Sensechat-5	24.75	15.11	68.43	32.45	19.70	70.02
Claude-3.5	14.23	9.95	67.56	12.66	10.06	68.12
GLM-4	13.44	10.07	67.26	23.57	14.28	70.30
ChatGPT	25.07	14.81	67.78	31.84	18.42	70.41
GPT-4o	17.58	12.17	68.68	27.64	16.52	71.59
GPT-4	19.47	13.45	68.99	28.43	16.74	71.26

Table 3: Results of the comparison between texts generated by various LLMs in Yue-TruthfulQA based on 0-shot and 5-shot settings and the correct texts. **Rouge-l**, **Bleu-4**, and **BERTScore** are evaluation metrics for comparing text and semantics similarity **(Table 8, 13, 14, 16, 16, 17 for more results)**.

Models	Acc. (0-shot)	Acc. (5-shot)
Qwen-1.5-110b	54.89	58.30
Qwen-2-72b	77.86	77.71
Qwen-2.5-72b	83.62	83.55
Mixtral-8x22b	65.20	66.19
Mixtral-large-2	80.14	81.27
Llama-3-70b	73.62	75.66
Llama-3.1-70b	53.60	79.00
Phi-3-medium	59.29	63.15
Gemma-2-27b	9.70	2.65
Yi-1.5-34b	69.45	69.45
Internlm-2.5-20b-chat	71.87	72.33
ERNIE-turbo	14.03	10.92
SenseChat-5	77.48	73.16
Claude-3.5	77.79	81.27
GLM-4	78.17	77.10
ChatGPT	23.35	41.09
GPT-4o	83.24	83.40
GPT-4	81.12	83.02

Table 4: Results of the comparison between various LLMs answer in Yue-GSM8K based on 0-shot and 5-shot and groundtruth **(Table 9, 18 for more results)**.

Models	Acc. (0-shot)	Acc. (5-shot)
Qwen-1.5-110b	88.64	90.09
Qwen-2-72b	88.64	88.56
Qwen-2.5-72b	92.74	92.91
Mixtral-8x22b	76.09	76.09
Mixtral-large-2	89.5	90.61
Llama-3-70b	85.06	84.97
Llama-3.1-70b	88.98	88.39
Phi-3-medium	77.63	78.31
Gemma-2-27b	67.98	55.59
Yi-1.5-34b	84.88	86.42
Internlm-2.5-20b-chat	82.15	82.58
ERNIE-turbo	44.41	46.46
SenseChat-5	88.47	87.28
Claude-3.5	91.55	92.23
GLM-4	88.9	88.73
ChatGPT	69.68	70.71
GPT-4o	91.97	94.45
GPT-4	92.66	92.06

Table 5: Results of the comparison between various LLMs answer in Yue-ARC-C based on 0-shot and 5-shot and groundtruth **(Table 10, 19 for more results)**.

prompts in the Appendix A.9). We use xFinder (Yu et al., 2024) to extract the answers of Yue-ARC-C, Yue-MMLU, Yue-GSM8K for later evaluation.

4.2 Evaluation

For Yue-TruthfulQA and Yue-TRANS (0-shot and 5-shot), we utilize Rouge-l, Bleu-4, and BERTScore as automatic evaluation metrics. **Rouge-l** (Lin, 2004) measures the longest common subsequence between generated and reference texts. **Bleu-4** (Papineni et al., 2002) evaluates n-gram overlap up to four words between generated and reference texts. **BERTScore** (Zhang* et al., 2020) evaluates semantic similarity using BERT embeddings (we use bert-base-multilingual-cased[5] for Cantonese evaluation and roberta-large[6] for English evaluation). For Yue-GSM8K, Yue-ARC-C,

[5] https://huggingface.co/google-bert/bert-base-multilingual-cased

[6] https://huggingface.co/FacebookAI/roberta-large

Models (Yue-MMLU)	0-shot (correct)					5-shot (correct)				
	STEM	Hum.	S.S.	C.S.	Oth.	STEM	Hum.	S.S.	C.S.	Oth.
Qwen-1.5-110b	75.07	88.48	83.89	80.57	82.14	79.96	88.12	88.75	84.8	89.31
Qwen-2-72b	81.68	89.93	88.47	81.9	87.48	85.7	89.54	88.12	83.72	87.73
Qwen-2.5-72b	83.72	87.88	87.2	80.68	85.36	83.89	89.7	88.75	82.34	87.42
Mixtral-8x22b	50.4	57.08	59.28	44.02	48.76	58.94	59.72	62.44	49.78	57.83
Mixtral-large-2	60.38	76.08	74.92	60.19	70.74	68.5	79.65	78.84	63.85	71.66
Llama-3-70b	65.17	73.58	75.22	57.87	72.84	64.06	72.82	73.16	57.34	72.95
Llama-3.1-70b	67.32	76.57	76.93	60.96	73.56	72.23	78.13	78.23	64.16	74.9
Phi-3-medium	45.26	61.42	58.4	45.65	51.33	49.88	59.33	59.35	45.49	53.02
Gemma-2-27b	48.5	54.05	53.32	36.92	48.22	40.62	41.72	43.81	32.99	46.03
Yi-1.5-34b	68.48	81.92	81.74	70.89	79.76	74.13	85.12	83.38	78.2	80.3
Internlm-2.5-20b-chat	67.16	81.56	77.72	73.05	72.64	66.22	82.65	78.42	72.94	74.03
ERNIE-turbo	43.34	56.05	53.97	52.02	44.82	41.01	57.66	54.28	49.49	46.95
Sensechat-5	69.97	83.21	80.73	73.86	76.95	68.98	82	79.88	73.52	74.77
Claude-3.5	66.47	76.84	78.04	60.6	75.98	75.92	81.65	84.24	62.83	82.54
GLM-4	64.23	84.39	80.06	75.66	75.75	72.18	84.2	80.07	76	78.06
ChatGPT	49.78	58.13	58.74	45.46	52.42	60.28	59.81	60.61	47.5	54.54
GPT-4o	74.16	83.28	84.12	71.6	84.32	72.35	85.03	84.32	72.74	81.58
GPT-4	67.68	75.29	77.26	60.12	74.46	71.19	76.75	77.56	63.5	74.57

Table 6: Results of the comparison between texts generated by various LLMs in Yue-MMLU based on 0-shot and 5-shot settings and the correct texts **(Table 11, 20 for more results)**.

and Yue-MMLU (0-shot and 5-shot), we employ **Accuracy (Acc.)** as the evaluation metric.

4.3 Large language models for comparison

We evaluate the Cantonese abilities of 35 models, encompassing twelve series of open-source and closed-source general and Cantonese LLMs, across four benchmarks. The LLMs evaluated are as follows (Appendix A.7 for details): (1) Qwen series: Qwen-7b, Qwen-1.5-7b, Qwen-1.5-110b, Qwen-2-7b, Qwen-2-72b, Qwen-2.5-7b, **Qwen-2.5-72b**; (2) Mixtral series: Mixtral-8x22b, **Mixtral-large-2**; (3) Llama series: Llama-2-7b, Llama-3-8b, Llama-3-70b, **Llama-3.1-8b**, **Llama-3.1-70b**; (4) Phi series: Phi-3-medium; (5) Gemma series: Gemma-2-27b; (6) Yi series: Yi-6b, Yi-1.5-6b, Yi-1.5-34b; (7) Internlm series: Internlm-2-7b, Internlm-2-20b, Internlm-2.5-7b, Internlm-2.5-20b; (8) ERNIE series: ERNIE-Lite, ERNIE-Tiny, ERNIE-Speed, ERNIE-Turbo ; (9) Sensechat series: Sensechat-5 (Cantonese); (10) Claude series: Claude-3.5-sonnet; (11) GLM series: GLM-4; (12) GPT series: ChatGPT, GPT-4o, GPT-4.

4.4 Results and analysis

The performance of Cantonese LLMs still lags behind that in Mandarin and English, and 5-shot is better than 0-shot. Rouge-l and Bleu-4 excel in evaluating the overlap between candidate and reference, making them suitable for key information extraction, outperforming metrics used in 0-shot and 5-shot (Figure 4a, b, c, d). The latter setting generally surpasses the former, illustrating the advantage of additional references in improving generation. Unlike these metrics, BERTScore excels in deep semantic evaluation, important for evaluating disparities in benchmarks between Cantonese and English. Mainstream LLMs perform better in English than in Cantonese (Figure 4e, f, g, h), highlighting their proficiency in widely used languages and relative under-performance in Cantonese (Table 3, Table 17). Accuracy metrics in benchmarks with unique answers corroborate these findings (Table 4, Table 5, Table 6, Table 18, Table 19, Table 20). 5-shot typically show higher accuracy than 0-shot (Figure 4a, b, c, d), and performance in mainstream languages like English and Mandarin surpasses that in Cantonese, emphasizing the need for more Cantonese-focused research and LLM development (Figure 4e, f, g, h).

Different series of models are suitable for various Cantonese tasks. Qwen-1.5-110b and Mixtral-large-2 lead in Cantonese factual generation in 0-shot, and Llama-3/3.1-70b, GPT-series in 5-shot, surpassing Sensechat-5, Gemma-2-27b and Phi-3-medium, excluding smaller models, is prone to hallucinations, affecting its scores (Figure 4).

GPT-4, GPT-4o and Claude-3.5 excel in mathematical logic, followed by Mixtral-large-2, Llama-3.1-70b, and GLM-4. Models like ChatGPT perform better in English, indicating challenges in Cantonese mathematical reasoning due to language nuances (Table 4, Figure 4b, g).

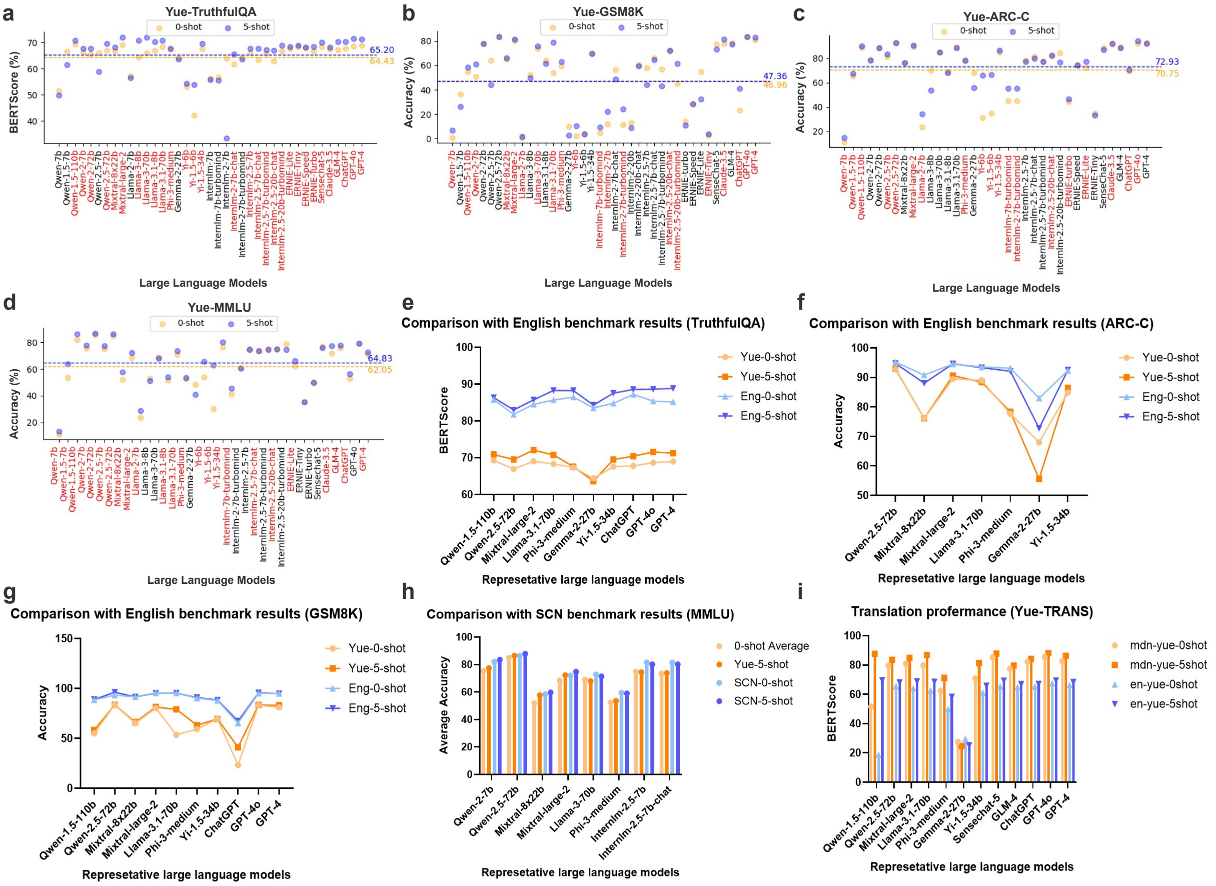

Figure 4: **a, b, c, d** represent the performance of various LLMs on Yue-TruthfulQA, Yue-GSM8K, Yue-ARC-C, and Yue-MMLU, in both 0-shot and 5-shot. **e, f, g, h** correspond to **comparisons of performance between four benchmarks and their English or Mandarin version. i** indicates the effectiveness of translating from Mandarin and English into Cantonese **(Table 3, 8, 13, 14, 16, 16, 17, 4 ,9, 18, 5, 10, 19, 6, 11, 20, 21, 22 for more results)**.

For complex reasoning, GPT-4 and GPT-4o consistently demonstrates optimal performance, closely followed by Qwen-2.5-72b, Claude-3.5, and Mixtral-large-2, each of which also exhibits excellent performance (Table 5).

For tasks across various topics of the MMLU, Qwen-2.5-72b consistently exhibits the best performance (Table 6). We compile a table detailing the best models for various personas along with recommended open-source models (Figure 5).

Enhancing Data Quality and Cost-Effectiveness for Cantonese LLMs. High-quality Cantonese data is crucial for the pre-training or alignment of Cantonese LLMs, with translations from Standard Chinese proving more effective due to linguistic similarities (Figure 4i), as opposed to English (Table 21, 22). While models like Gemma-2-27b perform less effectively in English-to-Cantonese translation, closed-source models such as Sensechat-5 and GPT series show minimal quality difference between 0-shot and 5-shot settings. Prioritizing translations from Standard Chinese, then English, optimizes data quality. Regarding cost-effectiveness, using closed-source models like Sensechat-5-Cantonese, ChatGPT, and GPT-4o is advisable if API costs are negligible (Table 21, 22). Models like Mixtral-large-2, Llama-3.1-70b and Qwen-1.5-110b offer cost savings and high-quality translations in both settings (Table 24, Figure 4i). The Llama and Qwen series, while not the highest in output quality, provides the best speed and cost-effectiveness for translating datasets to Cantonese.

4.5 Case study

In addition to the results analyzed above, we find that Gemma-2-27b frequently encounters hallucination issues, which impair its ability to handle Cantonese tasks (Appendix C). Although Qwen-2-72b exhibits good performance, it sometimes outputs training data. Nonetheless, the Qwen series of models remains proficient in handling Cantonese tasks (Appendix C). Appendix C for more cases.

5 Challenges and opportunities

5.1 Existing Cantonese challenges

Colloquialism. Cantonese differs significantly from Standard Chinese in its spoken vocabulary, posing unique challenges for NLP models initially trained on Mandarin (Snow, 2004; Xiang et al., 2024). These differences are particularly evident in informal settings such as speech transcription and online forums like Linkg, and Openrice. Although smaller compared to datasets for English and Standard Chinese models like BERTweet (Nguyen et al., 2020) and MacBERT (Cui et al., 2021), these platforms still provide a substantial text corpus for training Cantonese-specific models (Hale, 2001, 2016). The abundant unique expressions and slang in Cantonese, often embedded with complex cultural nuances, hinder adaptation of Standard Chinese-based models to Cantonese. For example, "Wan2 Sik6" literally means "looking for food", but it is commonly used to describe seeking employment or earning money, carrying connotations of survival and making a living in Cantonese. In addition, common spelling mistakes and novel meanings in Cantonese further complicate model training, emphasizing the need for robust, Cantonese-specific vocabularies and corpora to capture the full breadth of colloquialisms and idioms of the language (Li and Costa, 2009).

Multilingualism. To elucidate the multilingual dynamics in social media of Hong Kong, (Xiang et al., 2024) identify frequent code-switching between Cantonese and Standard Chinese, and a significant presence of English (Yue-Hashimoto, 1991; Li, 2006). Highlighting the multilingualism, examples include Cantonese sentences incorporating English terms, such as "deadline" seamlessly integrated as in "Gan2 M4 Cit3 deadline" (struggling to meet the deadline), and the use of the Japanese loanword "Kawaii" (cute), pronounced and adapted locally in phrases like "Ni1 Gin6 Saam1 Hou2 kawaii" (This shirt is very cute). These findings underscore the need for Cantonese NLP systems to handle multilingual code-switching and suggest adding spelling correction and dialect identification to improve data processing.

5.2 Opportunities

Given the existing challenges in Cantonese language and the evaluation results on benchmarks, we propose the following potential research directions and recommended models.

Task	Best (0-shot)	Best (5-shot)	Others
Factual gen.	Qwen-1.5-110b	Mixtral-large-2	Llama-3-70b, Yi-1.5-34b, GPT-4, GPT-4o, ...
Math logic	Qwen-2.5-72b	Qwen-2.5-72b	Llama-3.1-70b, Claude-3.5, GPT-4, GPT-4o, ...
C-Reasoning	Qwen-2.5-72b	GPT-4o	Qwen-1.5-110b, GPT-4, Mixtral-large-2, ...
Social sciences	Qwen-2.5-72b	Qwen-2.5-72b	Qwen-1.5-110b, Yi-1.5-34b, GPT-4o, ...
STEM	Qwen-2.5-72b	Qwen-2.5-72b	Llama-3.1-70b, Mixtral-large-2, Claude-3.5, ...
Humanities	Qwen-2.5-72b	Qwen-2.5-72b	Yi-1.5-34b, GPT-4o, Internlm-2.5-20b, ...
China specific	Qwen-2.5-72b	Qwen-2.5-72b	Qwen-1.5-110b, GPT-4o, Sensechat-5, ...
Other	Qwen-2.5-72b	Qwen-2.5-72b	Qwen-1.5-110b, GPT-4, Sensechat-5, GPT-4o, ...
Trans (mdn-yue)	Sensechat-5	Sensechat-5	Llama-3.1-70b, Qwen-1.5-110b, GLM-4, ...
Trans (Eng-yue)	GPT-4o	GPT-4o	Llama-3.1-70b, Mixtral-large-2, Sensechat-5, ...

Figure 5: LLMs proficient in handling various tasks.

Data augmentation. Data augmentation methods for Cantonese are similar to those used broadly, including label-invariant methods that modify text while preserving labels (Wei and Zou, 2019; Min et al., 2020; Shi et al., 2021), and label-variant techniques that alter semantics for new instances (Jin et al., 2019; Dai et al., 2019). Supervised contrastive learning enhances task-specific neural representations (Sedghamiz et al., 2021), and LLM-based strategies are reviewed in (Ding et al., 2024). For dataset conversion to Cantonese, high-capability models like Sensechat-5 and GPT-4 are recommended if costs allow (Table 21, 22, Table 24). Budget-friendly alternatives include Mixtral-large-2 and Llama-3.1-70b, with Llama models providing cost-effective speeds despite lower quality (Table 24).

Code-switch. Developments in LLMs suggest emergent abilities for untrained tasks, although effectiveness varies across scripts and languages (Mann et al., 2020; Bang et al., 2023). Research in SCN-adapted LLMs is progressing, benefiting Cantonese NLP in the future (Cui et al., 2023; Bai et al., 2023). We propose four benchmarks and compile a Yue-TRANS dataset, each involving two or more languages. Therefore, based on the performance observe on benchmarks, we recommend using newer versions of the Qwen, Llama, Mixtral, and Yi series (Figure 5).

Large language models. Based on the analysis above, we compile Figure 5, which presents the best LLMs in 0-shot and 5-shot, and suggests LLM series for various tasks. For work related to LLMs, we recommend using newer versions of the Qwen, Mixtral, Llama, and Yi series (Table 8,9, 10, 11). For tasks that involve only prompting, without the need for LLM training, we also recommend using closed-source models such as GPT, GLM, and Sensechat series models.

6 Conclusion and Outlook

Cantonese, spoken by over 85 million people, lags in natural language processing development, especially in large language models. To address this gap, we summarize existing Cantonese NLP methods and introduce four new benchmarks (Yue-Truthful, Yue-GSM8K, Yue-ARC-C, Yue-MMLU) and a translation dataset (Yue-TRANS). We evaluate 35 mainstream LLMs on these benchmarks, identifying current strengths and weaknesses. This work lays a foundation for advancing Cantonese LLM related technology.

Future efforts focus on building larger, high-quality Cantonese corpora and optimizing models for Cantonese-specific tasks. Collaboration among global researchers accelerates progress, helping Cantonese NLP catch up with other languages, enriching the experiences of Cantonese speakers.

Limitations

The first limitation is the scarcity of work related to Cantonese LLMs, which restricts the extent of summarizing relevant studies. However, it is believed that with the publication of this paper, an increasing number of projects involving large-scale Cantonese models will be proposed. The second limitation is that the recommended LLMs presented in the article are for reference only; LLMs not recommended are not necessarily of inferior quality, nor does it imply they are unsuitable for Cantonese-related tasks. The selection of specific models for Cantonese-related tasks should be based on a detailed analysis of the specific issues at hand.

In addition, we specifically focus on benchmarking vanilla LLMs without fine-tuning to test these LLMs' intrinsic abilities, which can also better inform their performance after fine-tuning.

Ethics Statement

Concerning the data annotators and the evaluation of data review, we ensure the selection of qualified tri-lingual individuals from Hong Kong and Guangdong who are compensated with reasonable hourly wages or other forms of subsidies as rewards. We have already obtained approval for this research from the Ethics Review Committee.

Acknowledgements

We want to thank our anonymous AC and reviewers for their feedback. This work was supported by Hong Kong Innovation and Technology Commission's Innovation and Technology Fund (Award No. ITS/269/22FP).

References

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.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Jian Chen, Dong Ping Huang, Shuyue Hu, Yu Liu, Yi Cai, and Huaqing Min. 2015. An opinion mining framework for cantonese reviews. *Journal of Ambient Intelligence and Humanized Computing*, 6:541–547.

Jian Chen, Yu Liu, Guangyi Zhang, Yi Cai, Tao Wang, and Huaqing Min. 2013. Sentiment analysis for cantonese opinion mining. In *2013 Fourth International Conference on Emerging Intelligent Data and Web Technologies*, pages 496–500. IEEE.

Xinyu Chen, Yifei Jian, Liang Ke, Yunxiang Qiu, Xingshu Chen, Yunya Song, and Haizhou Wang. 2024. A deep semantic-aware approach for cantonese rumor detection in social networks with graph convolutional network. *Expert Systems with Applications*, 245:123007.

Xinyu Chen, Liang Ke, Zhipeng Lu, Hanjian Su, and Haizhou Wang. 2020. A novel hybrid model for cantonese rumor detection on twitter. *Applied Sciences*, 10(20):7093.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. *arXiv preprint arXiv:1905.05621*.

Megan Dare, Valentina Fajardo Diaz, Averie Ho Zoen So, Yifan Wang, and Shibingfeng Zhang. 2023. Unsupervised mandarin-cantonese machine translation. *arXiv preprint arXiv:2301.03971*.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using llms: Data perspectives, learning paradigms and challenges. *arXiv preprint arXiv:2403.02990*.

Ziru Fu, Yu Cheng Hsu, Christian S Chan, Chaak Ming Lau, Joyce Liu, and Paul Siu Fai Yip. 2024. Efficacy of chatgpt in cantonese sentiment analysis: Comparative study. *Journal of Medical Internet Research*, 26:e51069.

Yin-Chun Fung, Lap-Kei Lee, Tsz-Chun Cheng, Chak-Fung Li, Vincent Chun-Kiu Wong, and Nga-In Wu. 2023. Canchat: A cantonese empathetic chatbot for secondary school student counseling. In *2023 International Symposium on Educational Technology (ISET)*, pages 170–175.

Zitian Gao, Boye Niu, Xuzheng He, Haotian Xu, Hongzhang Liu, Aiwei Liu, Xuming Hu, and Lijie Wen. 2024. Interpretable contrastive monte carlo tree search reasoning. *arXiv preprint arXiv:2410.01707*.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*.

John Hale. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412.

Alex Havrilla, Sharath Raparthy, Christoforus Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, and Roberta Raileanu. 2024. Glore: When, where, and how to improve llm reasoning via global and local refinements. *arXiv preprint arXiv:2402.10963*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Guangpu Huang, Arseniy Gorin, Jean-Luc Gauvain, and Lori Lamel. 2016. Machine translation based data augmentation for cantonese keyword spotting. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6020–6024. IEEE.

Thomas Hun-tak Lee. 1999. Cancorp-the hong kong cantonese child language corpus. *Revue Française de Linguistique Appliquée*, 4(1):21–30.

Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. 2019. IMaT: Unsupervised text attribute transfer via iterative matching and translation. 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 3097–3109, Hong Kong, China. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Jackson Lee, Litong Chen, Charles Lam, Chaak Ming Lau, and Tsz-Him Tsui. 2022. Pycantonese: Cantonese linguistics and nlp in python. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 6607–6611.

John Lee. 2019. An emotion detection system for cantonese. In *The Thirty-Second International Flairs Conference*.

John Lee and Baikun Liang. 2021. Response selection for a virtual counsellor. In *Companion Proceedings of the Web Conference 2021*, pages 495–499.

John SY Lee. 2011. Toward a parallel corpus of spoken cantonese and written chinese. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1462–1466.

John SY Lee, Baikun Liang, and Haley HM Fong. 2021. Restatement and question generation for counsellor chatbot. In *1st Workshop on Natural Language Processing for Programming (NLP4Prog)*, pages 1–7. Association for Computational Linguistics (ACL).

Man-Tak Leung and Sam-Po Law. 2001. Hkcac: the hong kong cantonese adult language corpus. *International journal of corpus linguistics*, 6(2):305–325.

David CS Li and Virginia Costa. 2009. Punning in hong kong chinese media: Forms and functions. *Journal of Chinese Linguistics*, 37(1):77–107.

Qingxin Li. 2006. *Maritime silk road*. Intercontinental Press.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Evelyn Kai-Yan Liu. 2022. Low-resource neural machine translation: A case study of cantonese. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 28–40.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 1.

Junghyun Min, R Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. *arXiv preprint arXiv:2004.11999*.

Eric WT Ngai, Maggie CM Lee, YS Choi, and PYF Chai. 2018. Multiple-domain sentiment classification for cantonese using a combined approach. In *PACIS*, page 297.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

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.

Wong Ping-Wai. 2006. The specification of pos tagging of the hong kong university cantonese corpus. *International Journal of Technology and Human Interaction (IJTHI)*, 2(1):21–38.

Hooman Sedghamiz, Shivam Raval, Enrico Santus, Tuka Alhanai, and Mohammad Ghassemi. 2021. SupCL-Seq: Supervised Contrastive Learning for downstream optimized sequence representations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3398–3403, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haoyue Shi, Karen Livescu, and Kevin Gimpel. 2021. Substructure substitution: Structured data augmentation for nlp. *arXiv preprint arXiv:2101.00411*.

Don Snow. 2004. *Cantonese as written language: The growth of a written Chinese vernacular*, volume 1. Hong Kong University Press.

Xu Tan, Muni Zhuang, Xin Lu, and Taitian Mao. 2021. An analysis of the emotional evolution of large-scale internet public opinion events based on the bert-lda hybrid model. *IEEE Access*, 9:15860–15871.

Sheng Wang, Liheng Chen, Pengan Chen, Jingwei Dong, Boyang Xue, Jiyue Jiang, Lingpeng Kong, and Chuan Wu. 2024a. Mos: Unleashing parameter efficiency of low-rank adaptation with mixture of shards. *arXiv preprint arXiv:2410.00938*.

Sheng Wang, Liheng Chen, Jiyue Jiang, Boyang Xue, Lingpeng Kong, and Chuan Wu. 2024b. Lora meets dropout under a unified framework. *arXiv preprint arXiv:2403.00812*.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Tak-sum Wong, John Lee, et al. 2018. Register-sensitive translation: a case study of mandarin and cantonese. In *Association for Machine Translation in the Americas*.

Dekai Wu. 1994. Aligning a parallel english-chinese corpus statistically with lexical criteria. *arXiv preprint cmp-lg/9406007*.

Yan Wu, Xiukun Li, and Suen Caesar Lun. 2006. A structural-based approach to cantonese-english machine translation. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 11, Number 2, June 2006*, pages 137–158.

Rong Xiang, Emmanuele Chersoni, Yixia Li, Jing Li, Chu-Ren Huang, Yushan Pan, and Yushi Li. 2024. Cantonese natural language processing in the transformers era: a survey and current challenges. *Language Resources and Evaluation*, pages 1–27.

Rong Xiang, Ying Jiao, and Qin Lu. 2019. Sentiment augmented attention network for cantonese restaurant review analysis. In *Proceedings of WISDOM'19: Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM'19)*, page 9.

Rong Xiang, Hanzhuo Tan, Jing Li, Mingyu Wan, and Kam-Fai Wong. 2022. When cantonese nlp meets pre-training: progress and challenges. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 16–21.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Virginia Yip and Stephen Matthews. 2007. *The bilingual child: Early development and language contact*. Cambridge University Press.

Qingchen Yu, Zifan Zheng, Shichao Song, Zhiyu Li, Feiyu Xiong, Bo Tang, and Ding Chen. 2024. xfinder: Robust and pinpoint answer extraction for large language models. *arXiv preprint arXiv:2405.11874*.

Anne Yue-Hashimoto. 1991. The yue dialect. *Journal of Chinese Linguistics Monograph Series*, (3):292–322.

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*.

Xiaoheng Zhang. 1998. Dialect mt: a case study between cantonese and mandarin. In *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*.

Ziqiong Zhang, Qiang Ye, Zili Zhang, and Yijun Li. 2011. Sentiment classification of internet restaurant reviews written in cantonese. *Expert Systems with Applications*, 38(6):7674–7682.

A Appendix

A.1 Cantonese speaking population statistics

Country/region	Pop.	Stat. Time
Hong Kong	6,529,000	2021
Macau	506,000	2016
Guangdong	40,000,000	2021
Guangxi	12,000,000	2022
Brunei	6,350	2006
Indonesia	180,000	1982
Malaysia	1,070,000	2000
Philippines	9,780	2000
Singapore	338,000	1993
Thailand	29,400	1984
Vietnam	862,000	1999

Table 7: Cantonese speaking population statistics. Pop. is population. Stat. Time is statistical time

A.2 Existing Cantonese data

At the end of the 16th century, Matteo Ricci compiles the first "Modern Bilingual Chinese Dictionary", significantly incorporating Cantonese terms, highlighting its role in Sino-Western interactions. By the 19th century, most bilingual dictionaries focus on Cantonese (Xiang et al., 2024). Historically, Hong Kong and related institutions lead Cantonese data initiatives. (Wu, 1994) creates a bilingual parallel corpus from the Hong Kong Legislative Council records, in both Standard Chinese and English. This effort is complemented by (Hun-tak Lee, 1999), who pioneers a Cantonese-only corpus with one million characters from dialogues involving children in Hong Kong, and (Yip and Matthews, 2007), who develops a bilingual corpus for Cantonese-speaking children. Additionally, a notable Cantonese corpus comes from television and theatrical productions in Hong Kong (Leung and Law, 2001). The University of Hong Kong further contributes by collecting and annotating spontaneous speech from dialogues and broadcasts, focusing on segmentation, parts of speech tagging, and phonetic transcription (Ping-Wai, 2006). (Lee, 2011) introduces a parallel corpus for machine translation between Cantonese and Standard Chinese, aligned at the sentence level, using data from Cantonese speeches on Hong Kong television and their Standard Chinese subtitles.

Recent efforts aim to bridge the data gap between Cantonese and other major languages. These include a small parallel dependency treebank for Cantonese and Mandarin, with 569 aligned sentences annotated using the Universal Dependencies scheme, and excerpts from the "ABC Cantonese-English Comprehensive Dictionary" providing 14,474 high-quality Cantonese-English parallel sentences, crucial for translation system development.

A.3 Cantonese small-scale neural network

Cantonese NLP research spreads across various topics, including rumor detection, sentiment analysis, machine translation, dialogue. We collect existing small neural network methods, models, and tools.

Rumor detection. (Chen et al., 2020) develop a dataset of 27,328 Cantonese tweets for rumor detection, split into 13,883 rumors and 13,445 non-rumors. They introduce an attention-based model, XGA, which combines XLNet (Yang et al., 2019) and BiGRU to analyze both semantic and sentiment aspects. (Chen et al., 2024) develop CantoneseBERT to capture glyph and pronunciation clues of Cantonese characters, and introduces a Cantonese rumor detection model, SA-GCN, that encodes global structural information of tweet hierarchies using the BiGCN model and extracts local semantic features with the CantoneseBERT model.

Sentiment analysis. Cantonese sentiment analysis utilizes diverse methodologies to address its linguistic complexities. (Zhang et al., 2011) apply Naive Bayes and SVMs with character-based bigrams in the Openrice app for effective emotion detection. (Chen et al., 2013, 2015) deploy Hidden Markov Models for text segmentation and part-of-speech tagging, developing emotion-specific dictionaries via rule-based systems. These studies demonstrate the value of combining machine learning with lexical techniques (Zhang et al., 2011; Chen et al., 2013, 2015). In addition, (Ngai et al., 2018) and (Xiang et al., 2019) enhance classification accuracy using supervised and unsupervised methods in various domains. (Lee, 2019) explores fine-grained emotion analysis across languages, achieving significant results. These efforts underscore the importance of multi-methodological approaches (Ngai et al., 2018; Xiang et al., 2019; Lee, 2019). (Tan et al., 2021) successfully employ Transformers pre-trained on simplified Chinese (Tan et al., 2021).

Machine translation. Initial research in this area utilizes heuristic rules, with significant contributions from (Zhang, 1998) and a bilingual Cantonese-English knowledge base by (Wu et al., 2006). The focus has since shifted to statistical machine translation, exemplified by (Huang et al., 2016), who addresses the challenges of translating between Cantonese and Mandarin with limited resources. (Wong et al., 2018) improves this approach by enhancing parallel data for more efficient model training. Recent developments include a large-scale evaluation dataset by (Liu, 2022), containing over 35,000 Mandarin-Cantonese sentence pairs, and unsupervised translation models by (Dare et al., 2023), which use cross-lingual embeddings and combine Transformer architecture with character-based tokenization to create a new corpus of approximately 1 million Cantonese sentences.

Dialogue summarization and generation. (Lee et al., 2021) explores generating questions and restating information in Cantonese dialogue systems, particularly for counseling chatbots. They enhance performance by fine-tuning the pre-trained BertSum model (Liu and Lapata, 2019) on Cantonese data, effective in tasks involving text summarization and question generation. In dialogue generation, (Lee and Liang, 2021) develops a specialized dataset for virtual counselors containing 1,028 post-reply pairs addressing test anxiety and loneliness, using these categories to guide response selection through a regression model.

Cantonese language model. Training Cantonese language models like XLNet (Yang et al., 2019) and ELECTRA (Clark et al., 2020) from ToastyNews[7] faces challenges due to data scarcity and legal constraints. (Chen et al., 2024) introduce CantoneseBERT and the SA-GCN model for detailed analysis and rumor detection in tweets, utilizing innovative methods like permutation learning and adversarial training. However, the training corpus largely includes Standard Chinese, leading to potential language contamination, and the impact on model efficacy remains unexplored.

Cantonese NLP tools. The landscape of Cantonese NLP tools is diverse, addressing various needs. PyCantonese (Lee et al., 2022) facilitates corpus data handling and linguistic analysis. Hong Kong Cantonese Localization provides culturally contextual translations. TransCan[8] enhances English-to-Cantonese translation, surpassing commercial solutions like Baidu and Bing. Text segmentation tools like Cantonese Word Segmentation[9] and cantoseg[10] improve accuracy through custom dictionaries. canto-filter[11] categorizes texts based on linguistic features, while songotsti[12] and fast-langid[13] offer additional support for language identification.

A.4 Cantonese large language model

Developing Cantonese LLMs is challenging due to scarce resources and the distinct features of the Cantonese language, necessitating extensive high-quality datasets for pre-training[14]. Despite these obstacles, these models show promising capabilities in processing Cantonese.

Aligning Cantonese LLMs for downstream tasks, such as prompting, supervised fine-tuning, and reinforcement learning from human feedback, is cost-

[7] https://huggingface.co/toastynews
[8] https://github.com/ayaka14732/TransCan
[9] https://github.com/wchan757/Cantonese_Word_Segmentation
[10] https://github.com/ayaka14732/cantoseg
[11] https://github.com/CanCLID/canto-filter
[12] https://github.com/justinchuntingho/songotsti
[13] https://github.com/ffreemt/fast-langid
[14] https://www.sensetime.com/en/news-detail/51168164?categoryId=1072

effective and helps eliminate biases and meet cultural expectations.

Recent studies (Fu et al., 2024) validate ChatGPT's effectiveness in Cantonese dialogue and sentiment analysis, analyzing messages from a Hong Kong web counseling service. The CanChat bot, introduced to enhance counseling services in Hong Kong, provides initial support to students, improving their emotional well-being during and beyond the COVID-19 pandemic (Fung et al., 2023).

Regarding the training and reasoning technologies for LLMs associated with mainstream languages, there is no development specific to Cantonese such as LoRA (Hu et al., 2021; Wang et al., 2024b,a), reasoning (Gao et al., 2024; Havrilla et al., 2024), etc.

Transitioning from small-scale networks to exploring Cantonese LLMs, both general-purpose and closed-source models show promise, but quantifying performance is challenging. We propose four benchmarks to evaluate and advance Cantonese capabilities in LLMs.

A.5 Evaluation tools

- **Rouge-l:** from rouge_metric import PyRouge
- **Bleu-4:** from nltk.translate.bleu_score import sentence_bleu, SmoothingFunction
- **BERTScore:** bert-base-multilingual-cased & roberta-large

A.6 Yue-MMLU

We select twenty-two topics from CMMLU that cover most of the themes in CMMLU to serve as the topics for Yue-MMLU, which are as follows:

- chinese_civil_service_exam
- arts
- electrical_engineering
- chinese_literature
- education
- economics
- ethnology
- college_medicine
- journalism
- management
- marketing
- philosophy
- security_study
- sociology
- world_history
- world_religions
- high_school_geography
- machine_learning
- marxist_theory
- professional_psychology
- sports_science
- logical

A.7 Source of evaluation LLMs

This section covers the evaluation of LLMs along with the corresponding Hugging Face links and the names of the APIs.

A.8 Experimental results

A.8.1 Cantonese and English TruthfulQA (best and incorrect)

Table 13 (comparison between **best** answer and groundtruth) and Table 16 (comparison between **incorrect** answer and groundtruth) are the experimental results based on the Cantonese and English version of TruthfulQA.

A.8.2 English TruthfulQA (correct)

Table 17 (comparision vetween **correct** answer and groundtruth) is the experimental result based on the English version of TruthfulQA, intended for comparison with the Cantonese version of TruthfulQA. For more results, please refer to the publicly available evaluation platform[15].

A.8.3 English GSM8K

Table 18 is the experimental result based on the English version of GSM8K, intended for comparison with the Cantonese version of GSM8K. For more results, please refer to the publicly available evaluation platform[16].

[15] https://huggingface.co/open-llm-leaderboard
[16] https://huggingface.co/open-llm-leaderboard

A.8.4 English ARC challenge

Table 19 is the experimental result based on the English version of ARC Challenge, intended for comparison with the Cantonese version of ARC Challenge. For more results, please refer to the publicly available evaluation platform[17].

A.8.5 CMMLU

Table 20 is the experimental result based on the Standard Chinese version of MMLU, intended for comparison with the Cantonese version of MMLU. For more results, please refer to the publicly available evaluation platform[18].

A.8.6 Translation

Table 21, 22 is the experimental result based on the Yue-Trans datasets. Table 23 and Table 24 reflect the running time of different LLMs on the translation dataset.

A.9 Prompt templates for multilingual evaluation

This section details the prompt templates used for the Cantonese, English, and Standard Chinese datasets tested in our experiments. Each dataset was evaluated under both 0-shot and 5-shot settings. For the 5-shot setting, except for the translation task (Yue-TRANS), the prompts were generated using a sliding window approach, where the preceding five examples from the dataset (Yue-TruthfulQA, Yue-GSM8K, Yue-ARC-Challenge, and Yue-MMLU) were utilized as context for each new example. For the Yue-TRANS translation task, the BM25 algorithm was employed to identify and select the five most similar examples to serve as few-shot examples. Below, we outline the prompt structures and methodologies used for generating the few-shot examples.

The inference and evaluation processes in this study were facilitated by the OpenCompass platform, which provided a robust and universal evaluation framework for foundation models (Contributors, 2023).

A.9.1 Yue-TruthfulQA prompt

0-shot:

```
用粵語答下面問題：
問題：[QUESTION]
回應：
```

5-shot:

```
樣例1-5：
    問題：[EXAMPLE_QUESTION]
    回應：[EXAMPLE_ANSWER]

用粵語答下面問題：
問題：[TARGET_QUESTION]
回應：
```

A.9.2 En-TruthfulQA prompt

0-shot:

```
Answer the following question in
↪ English:
Question: [QUESTION]
Answer:
```

5-shot:

```
Example 1-5:
    Question: [EXAMPLE_QUESTION]
    Answer: [EXAMPLE_ANSWER]

Answer the following question in
↪ English:
Question: [TARGET_QUESTION]
Answer:
```

A.9.3 Yue-GSM8K prompt

0-shot:

```
請逐步思考，最終答案前用「####」標
↪ 記。用粵語答下面問題：
問題：[QUESTION]
用粵語回答問題：
```

5-shot:

```
樣例1-5：
    問題：[EXAMPLE_QUESTION]
    回應：[EXAMPLE_ANSWER]

請逐步思考，最終答案前用「####」標
↪ 記。用粵語答下面問題：
問題：[TARGET_QUESTION]
用粵語回答問題：
```

A.9.4 En-GSM8K prompt

0-shot:

```
Please think step by step, mark the
↪ final answer with '####'.
Answer the following question in
↪ English:
Question: [QUESTION]
Answer the question in English:
```

[17] https://huggingface.co/open-llm-leaderboard
[18] https://huggingface.co/open-llm-leaderboard

5-shot:

```
Example 1-5:
    Question: [EXAMPLE_QUESTION]
    Response: [EXAMPLE_ANSWER]

Please think step by step, mark the
↪  final answer with '####'.
Answer the following question in
↪  English:
Question: [TARGET_QUESTION]
Answer the question in English:
```

A.9.5 Yue-ARC-C prompt

0-shot:

```
問題：[QUESTION]
由提供選項中直接用選項字母作答。
回應：
```

5-shot:

```
樣例1-5：
    問題：[EXAMPLE_QUESTION]
    回應：[EXAMPLE_ANSWER]

問題：[TARGET_QUESTION]
由提供選項中直接用選項字母作答。
回應：
```

A.9.6 En-ARC-C prompt

0-shot:

```
Question: [QUESTION]
Answer with the option's letter from
↪  the given choices directly.
Answer:
```

5-shot:

```
Example 1-5:
    Question: [EXAMPLE_QUESTION]
    Answer: [EXAMPLE_ANSWER]

Question: [TARGET_QUESTION]
Answer with the option's letter from
↪  the given choices directly.
Answer:
```

A.9.7 Yue-MMLU prompt

0-shot:

```
以下係關於[SUBJECT]單項選擇題，請直
↪  接畀出正確答案選項。
問題：[QUESTION]
答案：
```

5-shot:

```
樣例1-5：
    問題：[EXAMPLE_QUESTION]
    回應：[EXAMPLE_ANSWER]

以下係關於[SUBJECT]單項選擇題，請直
↪  接畀出正確答案選項。
問題：[TARGET_QUESTION]
答案：
```

A.9.8 Zh-CMMLU prompt

0-shot:

```
以下是关于[SUBJECT]的单项选择题，请
↪  直接给出正确答案的选项。
题目：[QUESTION]
答案：
```

5-shot:

```
样例1-5:
    题目：[EXAMPLE_QUESTION]
    答案：[EXAMPLE_ANSWER]

以下是关于[SUBJECT]的单项选择题，请
↪  直接给出正确答案的选项。
题目：[TARGET_QUESTION]
答案：
```

A.9.9 Yue-TRANS prompt

0-shot:

```
請將下面呢句/段話直接翻譯成粵
↪  語：[SOURCE_TEXT]
```

5-shot:

```
樣例1-5：
    請將下面呢句/段話直接翻譯成粵
    ↪  語：[EXAMPLE_SOURCE_TEXT]
    翻譯：[EXAMPLE_TRANSLATION_TEXT]

根據上面例子，請將下面呢句/段話直接
↪  翻譯成粵語：
[TARGET_SOURCE_TEXT]
```

B Result

C Case study

In this section, we provide a case study to illustrate the input and output of our experiment. We demonstrate the model's behavior using example prompts and their corresponding outputs.

Models (Yue-TruthfulQA)	0-shot (correct)			5-shot (correct)		
	Rouge-l	**Bleu-4**	**BERTScore**	**Rouge-l**	**Bleu-4**	**BERTScore**
Qwen-7b	6.42	3.99	51.57	4.04	2.98	49.7
Qwen-1.5-7b	20.54	13.41	66.45	12.45	10.41	61.59
Qwen-1.5-110b	26.04	15.95	69.29	31.73	19.53	70.87
Qwen-2-7b	13.27	10.00	66.14	16.91	11.48	67.71
Qwen-2-72b	10.86	9.68	65.62	17.52	12.38	67.72
Qwen-2.5-7b	18.51	12.28	66.07	6.83	8.07	58.97
Qwen-2.5-72b	13.03	9.64	66.94	20.23	12.87	69.53
Mixtral-8x22b	14.74	10.83	66.72	20.40	14.09	68.05
Mixtral-large-2	19.72	13.01	69.06	31.38	18.61	72.07
Llama-2-7b	3.48	6.42	57.16	3.57	6.52	56.36
Llama-3-8b	8.40	8.68	64.37	28.68	16.43	70.82
Llama-3-70b	10.98	9.51	66.10	33.06	19.31	71.95
Llama-3.1-8b	13.82	10.33	66.97	26.18	15.20	70.28
Llama-3.1-70b	21.03	14.30	68.31	34.72	20.54	70.80
Phi-3-medium	18.70	12.00	67.36	22.00	13.72	67.57
Gemma-2-27b	8.09	8.44	64.41	11.33	9.98	63.66
Yi-6b	1.37	5.05	53.16	1.07	5.99	54.21
Yi-1.5-6b	1.21	4.60	42.15	1.04	6.15	53.85
Yi-1.5-34b	15.41	11.11	67.57	20.30	13.20	69.50
Internlm-7b	5.89	6.65	56.33	2.59	3.68	55.73
Internlm-7b-turbomind	5.91	6.71	56.71	2.77	3.82	55.57
Internlm-2-7b	7.93	10.21	63.81	17.66	16.62	33.33
Internlm-2-7b-chat	6.7	7.68	61.83	3.3	5.49	65.47
Internlm-2-7b-turbomind	8.09	10.53	64.3	17.69	16.99	63.68
Internlm-2.5-7b	8.96	10.53	66.11	10.3	14.47	67.73
Internlm-2.5-7b-chat	7.13	8	63.48	4.05	7.19	67.61
Internlm-2.5-7b-turbomind	8.93	10.46	65.75	10.12	14.39	67.14
Internlm-2.5-20b-chat	6.96	7.73	62.99	3.28	6.06	66.99
Internlm-2.5-20b-turbomind	9.49	11.55	66.70	11.98	16.56	68.86
ERNIE-Lite	20.58	12.23	67.64	20.69	12.27	68.45
ERNIE-Tiny	27.16	14.49	68.45	27.91	15.28	68.84
ERNIE-Speed	22.58	13.15	67.84	23.61	13.82	68.27
ERNIE-Turbo	17.91	11.30	66.71	21.19	12.19	68.29
Sensechat-5	24.75	15.11	68.43	32.45	19.70	70.02
Claude-3.5	14.23	9.95	67.56	12.66	10.06	68.12
GLM-4	13.44	10.07	67.26	23.57	14.28	70.30
ChatGPT	25.07	14.81	67.78	31.84	18.42	70.41
GPT-4o	17.58	12.17	68.68	27.64	16.52	71.59
GPT-4	19.47	13.45	68.99	28.43	16.74	71.26

Table 8: Results of the comparison between texts generated by various LLMs in Yue-TruthfulQA based on 0-shot and 5-shot settings and the correct texts. **Rouge-l**, **Bleu-4**, and **BERTScore** are evaluation metrics for comparing text similarity.

Models	**Acc. (0-shot)**	**Acc. (5-shot)**
Qwen-7b	0.68	6.75
Qwen-1.5-7b	36.62	26.31
Qwen-1.5-110b	54.89	58.30
Qwen-2-7b	50.49	61.11
Qwen-2-72b	77.86	77.71
Qwen-2.5-7b	63.84	44.20
Qwen-2.5-72b	83.62	83.55
Mixtral-8x22b	65.20	66.19
Mixtral-large-2	80.14	81.27
Llama-2-7b	0.83	1.82
Llama-3-8b	52.46	49.66
Llama-3-70b	73.62	75.66
Llama-3.1-8b	63.91	61.64
Llama-3.1-70b	53.60	79.00
Phi-3-medium	59.29	63.15
Gemma-2-27b	9.70	2.65
Yi-6b	2.12	10.16
Yi-1.5-6b	3.94	3.49
Yi-1.5-34b	69.45	69.45
Internlm-7b-turbomind	4.55	9.48
Internlm-2-7b	11.90	22.21
Internlm-2-7b-chat	56.41	48.67
Internlm-2-7b-turbomind	11.37	23.96
Internlm-2-20b	12.81	8.87
Internlm-2-20b-chat	60.42	59.21
Internlm-2.5-7b	57.70	44.05
Internlm-2.5-7b-chat	65.96	64.67
Internlm-2.5-7b-turbomind	56.79	42.99
Internlm-2.5-20b-chat	71.87	72.33
Internlm-2.5-20b-turbomind	45.03	61.41
ERNIE-turbo	14.03	10.92
ERNIE-Speed	28.81	28.28
ERNIE-Lite	54.81	32.15
ERNIE-Tiny	2.73	3.94
SenseChat-5	77.48	73.16
Claude-3.5	77.79	81.27
GLM-4	78.17	77.10
ChatGPT	23.35	41.09
GPT-4o	83.24	83.40
GPT-4	81.12	83.02

Table 9: Results of the comparison between answer generated by various LLMs in Yue-GSM8K based on 0-shot and 5-shot settings and groundtruth.

Models	**Acc. (0-shot)**	**Acc. (5-shot)**
Qwen-7b	11.02	14.6
Qwen-1.5-7b	65.24	67.55
Qwen-1.5-110b	88.64	90.09
Qwen-2-7b	79.08	78.39
Qwen-2-72b	88.64	88.56
Qwen-2.5-7b	81.64	83.35
Qwen-2.5-72b	92.74	92.91
Mixtral-8x22b	76.09	76.09
Mixtral-large-2	89.5	90.61
Llama-2-7b	23.57	34.24
Llama-3-8b	70.11	53.8
Llama-3-70b	85.06	84.97
Llama-3.1-8b	69	67.81
Llama-3.1-70b	88.98	88.39
Phi-3-medium	77.63	78.31
Gemma-2-27b	67.98	55.59
Yi-6b	31	66.01
Yi-1.5-6b	34.59	66.7
Yi-1.5-34b	84.88	86.42
Internlm-7b-turbomind	44.75	55.34
Internlm-2-7b-turbomind	44.75	55.34
Internlm-2.5-7b	78.14	77.46
Internlm-2.5-7b-chat	81.21	79.85
Internlm-2.5-7b-turbomind	77.37	77.37
Internlm-2.5-20b-chat	82.15	82.58
Internlm-2.5-20b-turbomind	84.29	76.94
ERNIE-turbo	44.41	46.46
ERNIE-Speed	74.47	74.04
ERNIE-Lite	72.25	77.28
ERNIE-Tiny	34.67	32.88
SenseChat-5	88.47	87.28
Claude-3.5	91.55	92.23
GLM-4	88.9	88.73
ChatGPT	69.68	70.71
GPT-4o	91.97	94.45
GPT-4	92.66	92.06

Table 10: Results of the comparison between answer generated by various LLMs in Yue-ARC-C based on 0-shot and 5-shot settings and groundtruth.

Models (Yue-MMLU)	0-shot (correct)					5-shot (correct)				
	STEM	**Hum.**	**S.S.**	**C.S.**	**Oth.**	**STEM**	**Hum.**	**S.S.**	**C.S.**	**Oth.**
Qwen-7b	10.1	12.95	12.12	11.61	7.96	9.98	15.96	14.48	13.33	13.26
Qwen-1.5-7b	46.28	61.65	56.57	50.02	53	60.14	70.09	65.55	58.31	65.02
Qwen-1.5-110b	75.07	88.48	83.89	80.57	82.14	79.96	88.12	88.75	84.8	89.31
Qwen-2-7b	70.06	81.04	80.07	69.54	76.04	74.08	80.45	80.7	73.7	79.52
Qwen-2-72b	81.68	89.93	88.47	81.9	87.48	85.7	89.54	88.12	83.72	87.73
Qwen-2.5-7b	72.86	81.66	78.25	66.56	75.19	78.05	80.37	78.99	69.82	78.86
Qwen-2.5-72b	83.72	87.88	87.2	80.68	85.36	83.89	89.7	88.75	82.34	87.42
Mixtral-8x22b	50.4	57.08	59.28	44.02	48.76	58.94	59.72	62.44	49.78	57.83
Mixtral-large-2	60.38	76.08	74.92	60.19	70.74	68.5	79.65	78.84	63.85	71.66
Llama-2-7b	23.34	23.84	23.76	22.78	24.52	27.48	30.4	31.76	28.9	24.38
Llama-3-8b	49.13	59.3	56.51	47.53	53.72	44.04	58.47	53.94	46.24	52.55
Llama-3-70b	65.17	73.58	75.22	57.87	72.84	64.06	72.82	73.16	57.34	72.95
Llama-3.1-8b	45.96	58.27	56.08	44.86	53.7	53.45	58.06	58.31	45.86	53.65
Llama-3.1-70b	67.32	76.57	76.93	60.96	73.56	72.23	78.13	78.23	64.16	74.9
Phi-3-medium	45.26	61.42	58.4	45.65	51.33	49.88	59.33	59.35	45.49	53.02
Gemma-2-27b	48.5	54.05	53.32	36.92	48.22	40.62	41.72	43.81	32.99	46.03
Yi-6b	36.46	67.62	57.32	57.42	50.06	58.11	72.14	68.4	60.56	68.46
Yi-1.5-6b	17.34	35.98	38.77	32.9	25	58.53	67.89	66.56	60	62.05
Yi-1.5-34b	68.48	81.92	81.74	70.89	79.76	74.13	85.12	83.38	78.2	80.3
Internlm-7b-turbomind	31.9	48.79	44.03	41.14	39.82	39.84	51.74	50.06	43.6	42.32
Internlm-2-7b-turbomind	51.69	70.92	64.71	59.31	58.93	53.11	68.51	62.68	59.77	58.14
Internlm-2.5-7b	65.34	82.43	79.24	73.11	74.15	66.73	81.06	77.8	71.65	75.37
Internlm-2.5-7b-chat	64.4	80.92	76.8	70.24	75.02	65.04	80.84	76.79	70.47	75.19
Internlm-2.5-7b-turbomind	65.34	82.43	79.24	73.11	74.15	66.73	81.06	77.8	71.65	75.37
Internlm-2.5-20b-chat	67.16	81.56	77.72	73.05	72.64	66.22	82.65	78.42	72.94	74.03
Internlm-2.5-20b-turbomind	72.86	86.1	82.14	79.06	74.7	69.65	78.79	76.56	70.28	77.2
ERNIE-Lite	53.45	67.56	67.73	61.21	61.21	60.74	70.27	71.5	62.43	64.84
ERNIE-Tiny	34.78	37.86	37.88	33.08	32.29	32.52	38.63	37.58	32.52	34.6
ERNIE-turbo	43.34	56.05	53.97	52.02	44.82	41.01	57.66	54.28	49.49	46.95
Sensechat-5	69.97	83.21	80.73	73.86	76.95	68.98	82	79.88	73.52	74.77
Claude-3.5	66.47	76.84	78.04	60.6	75.98	75.92	81.65	84.24	62.83	82.54
GLM-4	64.23	84.39	80.06	75.66	75.75	72.18	84.2	80.07	76	78.06
ChatGPT	49.78	58.13	58.74	45.46	52.42	60.28	59.81	60.61	47.5	54.54
GPT-4o	74.16	83.28	84.12	71.6	84.32	72.35	85.03	84.32	72.74	81.58
GPT-4	67.68	75.29	77.26	60.12	74.46	71.19	76.75	77.56	63.5	74.57

Table 11: Results of the comparison between texts generated by various LLMs in Yue-MMLU based on 0-shot and 5-shot settings and the correct texts.

Models	**Mode**	**Huggingface link & API name**
Qwen-7b	Huggingface	https://huggingface.co/Qwen/Qwen-7B
Qwen-1.5-7b	Huggingface	https://huggingface.co/Qwen/Qwen1.5-7B
Qwen-1.5-110b	Huggingface	https://huggingface.co/Qwen/Qwen1.5-110B
Qwen-2-7b	Huggingface	https://huggingface.co/Qwen/Qwen2-7B-Instruct
Qwen-2-72b	Huggingface	https://huggingface.co/Qwen/Qwen2-72B-Instruct
Qwen-2.5-7b	Huggingface	https://huggingface.co/Qwen/Qwen2.5-7B
Qwen-2.5-72b	Huggingface	https://huggingface.co/Qwen/Qwen2.5-72B
Mixtral-8x22b	Huggingface	https://huggingface.co/mistralai/Mixtral-8x22B-Instruct-v0.1
Mixtral-large-2	Huggingface	https://huggingface.co/mistralai/Mistral-Large-Instruct-2407
Llama-2-7b	Huggingface	https://huggingface.co/meta-llama/Llama-2-7b
Llama-3-8b	Huggingface	https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
Llama-3-70b	Huggingface	https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct
Llama-3.1-8b	Huggingface	https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct
Llama-3.1-70b	Huggingface	https://huggingface.co/meta-llama/Meta-Llama-3.1-70B-Instruct
Phi-3-medium	Huggingface	https://huggingface.co/microsoft/Phi-3-medium-128k-instruct
Gemma-2-27b	Huggingface	https://huggingface.co/google/gemma-2-27b-it
Yi-6b	Huggingface	https://huggingface.co/01-ai/Yi-6B
Yi-1.5-6b	Huggingface	https://huggingface.co/01-ai/Yi-1.5-6B-Chat
Yi-1.5-34b	Huggingface	https://huggingface.co/01-ai/Yi-1.5-34B-Chat
ERNIE-turbo	API	API: ERNIE-Bot-turbo
ERNIE-Speed	API	API: ERNIE-Speed-128K
ERNIE-Lite	API	API: ERNIE-Lite-8K
ERNIE-Tiny	API	API: ERNIE-Tiny-8K
Internlm-2-7b	Huggingface	https://huggingface.co/internlm/internlm2-7b
Internlm-2-7b-chat	Huggingface	https://huggingface.co/internlm/internlm2-20b
Internlm-2-20b-chat	Huggingface	https://huggingface.co/internlm/internlm2-chat-20b
Internlm-2.5-7b	Huggingface	https://huggingface.co/internlm/internlm2_5-7b
Internlm-2.5-7b-chat	Huggingface	https://huggingface.co/internlm/internlm2_5-7b-chat
Internlm-2.5-20b-chat	Huggingface	https://huggingface.co/internlm/internlm2_5-20b-chat
SenseChat-5	API	API: SenseChat-5-Cantonese
Claude-3.5	API	API: claude-3.5-sonnot-20240620
GLM-4	API	API: GLM-4-0520
ChatGPT	API	API: gpt-3.5-turbo-instruct & gpt-3.5-turbo
GPT-4o	API	API: gpt-4o
GPT-4	API	API: gpt-4-0125-preview

Table 12: The mode of the evaluation LLMs and their corresponding huggingface links & names of APIs.

Models (Yue-TruthfulQA)	**0-shot (best)**		**5-shot (best)**	
	Bleu-4	**BERTScore**	**Bleu-4**	**BERTScore**
Qwen-7b	3.01	51.03	2.19	48.82
Qwen-1.5-7b	9.8	65.65	8.19	59.48
Qwen-1.5-110b	11.17	69.14	14.22	73.40
Qwen-2-7b	8.00	64.11	9.09	66.41
Qwen-2-72b	7.77	62.22	9.99	65.32
Qwen-2.5-7b	8.99	65.58	6.98	54.62
Qwen-2.5-72b	7.8	64.10	10.18	68.19
Mixtral-8x22b	8.54	64.63	11.31	67.43
Mixtral-large-2	10.01	67.37	14.14	73.41
Llama-2-7b	5.36	52.10	5.53	51.12
Llama-3-8b	7.26	60.79	12.94	71.77
Llama-3-70b	7.70	63.08	14.68	73.97
Llama-3.1-8b	8.19	63.97	11.93	70.64
Llama-3.1-70b	10.42	67.19	15.36	75.80
Phi-3-medium	9.34	65.84	10.98	66.81
Gemma-2-27b	7.15	60.94	8.14	61.54
Yi-6b	3.95	49.13	4.98	48.93
Yi-1.5-6b	3.82	38.22	5.15	48.43
Yi-1.5-34b	8.80	65.25	10.55	67.88
Internlm-7b	5.39	52.10	5.42	50.65
Internlm-7b-turbomind	5.26	52.45	5.4	50.56
Internlm-2-7b	6.51	64.20	5.58	33.94
Internlm-2-7b-chat	6.4	58.87	9.41	65.32
Internlm-2-7b-turbomind	6.85	64.77	9.71	64.19
Internlm-2-20b	8.65	68.08	2.61	20.27
Internlm-2-20b-chat	6.08	56.94	10.57	66.29
Internlm-2.5-7b	8.24	63.72	11.02	67.35
Internlm-2.5-7b-chat	6.79	60.35	8.41	65.13
Internlm-2.5-7b-turbomind	8.27	63.25	10.6	66.58
Internlm-2.5-20b-chat	6.55	59.61	8.06	64.32
Internlm-2.5-20b-turbomind	8.75	65.18	11.97	69.30
ERNIE-Lite	9.05	67.61	9.44	67.68
ERNIE-Tiny	14.49	70.05	10.82	70.39
ERNIE-Speed	9.54	68.33	10.49	68.49
ERNIE-Turbo	9.04	65.20	9.66	67.39
Sensechat-5	10.47	68.93	14.51	73.38
Claude-3.5	7.95	64.83	8.24	64.84
GLM-4	7.92	64.28	11.11	69.65
ChatGPT	10.42	67.84	13.82	71.87
GPT-4o	9.34	66.25	12.61	71.51
GPT-4	9.97	67.08	12.87	72.00

Table 13: Results of the comparison between texts generated by various LLMs in Cantonese version of TruthfulQA based on 0-shot and 5-shot settings and the **best** texts. **Rouge-l**, **Bleu-4**, and **BERTScore** are evaluation metrics for comparing text similarity.

Models (TruthfulQA-English)	0-shot (best)		5-shot (best)	
	Bleu-4	**BERTScore**	**Bleu-4**	**BERTScore**
Qwen-1.5-110b	12.78	85.83	20.10	87.19
Qwen-2-7b	8.76	83.80	16.37	87.10
Qwen-2-72b	6.99	81.36	8.58	82.97
Qwen-2.5-72b	9.22	84.30	11.33	85.72
Mixtral-8x22b	10.82	85.68	17.65	88.24
Mixtral-large-2	11.95	85.68	25.12	89.97
Llama-3-8b	10.04	83.86	32.17	90.98
Llama-3-70b	9.07	83.42	31.85	90.99
Llama-3.1-8b	9.81	83.19	31.18	90.56
Llama-3.1-70b	11.27	84.01	35.02	91.60
Phi-3-medium	12.33	86.70	24.27	89.57
Gemma-2-27b	8.46	83.20	10.52	84.24
Yi-1.5-34b	11.01	84.72	22.50	88.79
Internlm-2-7b	22.39	88.41	25.76	67.10
Internlm-2-7b-chat	8.41	83.21	16.14	86.96
Internlm-2-20b	21.77	88.38	26.70	86.60
Internlm-2-20b-chat	7.32	81.76	20.57	87.38
Internlm-2.5-7b	15.17	86.40	22.06	88.43
Internlm-2.5-7b-chat	7.77	82.73	9.95	84.40
ChatGPT	17.97	87.65	26.69	90.27
GPT-4o	10.93	85.28	32.38	90.94
GPT-4	11.51	85.16	34.34	91.36

Table 14: Results of the comparison between texts generated by various LLMs in English version of TruthfulQA based on 0-shot and 5-shot settings and the **best** texts. **Rouge-l**, **Bleu-4**, and **BERTScore** are evaluation metrics for comparing text similarity.

Models (Yue-TruthfulQA)	0-shot (incorrect)		5-shot (incorrect)	
	Bleu-4	**BERTScore**	**Bleu-4**	**BERTScore**
Qwen-7b	3.22	52.45	2.38	50.82
Qwen-1.5-7b	11.39	66.76	8.43	61.74
Qwen-1.5-110b	12.83	69.22	12.67	68.67
Qwen-2-7b	8.38	65.10	8.38	65.56
Qwen-2-72b	8.15	64.44	9.17	66.03
Qwen-2.5-7b	10.14	66.13	7.10	59.77
Qwen-2.5-72b	8.19	65.49	9.82	67.49
Mixtral-8x22b	9.24	66.27	10.14	66.11
Mixtral-large-2	10.60	68.40	12.62	69.74
Llama-2-7b	5.74	59.48	5.69	58.38
Llama-3-8b	7.69	64.07	11.03	68.54
Llama-3-70b	8.12	65.49	12.11	69.10
Llama-3.1-8b	8.72	66.38	10.73	68.22
Llama-3.1-70b	10.79	67.80	12.38	68.28
Phi-3-medium	10.23	67.07	10.40	66.07
Gemma-2-27b	7.40	63.04	8.05	62.28
Yi-6b	4.27	54.49	5.29	55.44
Yi-1.5-6b	4.15	43.31	5.35	55.07
Yi-1.5-34b	9.16	66.67	10.04	67.68
Internlm-7b	5.89	57.93	5.58	56.81
Internlm-7b-turbomind	5.91	58.23	5.54	56.7
Internlm-2-7b	7.93	64.39	4.73	32.66
Internlm-2-7b-chat	6.70	61.13	9.17	64.11
Internlm-2-7b-turbomind	8.09	64.76	8.50	62.9
Internlm-2-20b	10.24	66.74	2.30	21.15
Internlm-2-20b-chat	6.27	59.46	9.56	64.82
Internlm-2.5-7b	8.96	65.89	10.25	66.48
Internlm-2.5-7b-chat	7.13	62.94	8.84	66.68
Internlm-2.5-7b-turbomind	8.93	65.7	9.81	66.14
Internlm-2.5-20b-chat	6.96	62.15	8.23	65.67
Internlm-2.5-20b-turbomind	9.49	66.25	11.45	67.84
ERNIE-Lite	9.72	66.86	9.40	66.73
ERNIE-Tiny	11.50	67.96	11.63	67.90
ERNIE-Speed	10.18	66.93	10.52	66.93
ERNIE-Turbo	9.52	66.15	9.70	66.76
Sensechat-5	12.02	68.33	12.31	67.80
Claude-3.5	8.20	65.93	7.78	65.57
GLM-4	8.43	66.00	10.34	68.09
ChatGPT	11.29	67.46	13.07	68.69
GPT-4o	9.64	67.40	11.21	68.89
GPT-4	10.45	67.72	11.49	68.52

Table 15: Results of the comparison between texts generated by various LLMs in Cantonese version of TruthfulQA based on 0-shot and 5-shot settings and the **incorrect** texts. **Rouge-l**, **Bleu-4**, and **BERTScore** are evaluation metrics for comparing text similarity.

Models (TruthfulQA-English)	**0-shot (incorrect)**		**5-shot (incorrect)**	
	Bleu-4	**BERTScore**	**Bleu-4**	**BERTScore**
Qwen-1.5-110b	12.83	85.75	13.89	85.31
Qwen-2-7b	8.65	83.70	11.39	85.02
Qwen-2-72b	6.84	81.59	7.98	82.30
Qwen-2.5-72b	6.84	84.04	7.98	85.19
Mixtral-8x22b	9.94	85.19	12.63	86.15
Mixtral-large-2	11.18	85.21	16.21	86.50
Llama-3-8b	10.01	84.02	19.84	86.68
Llama-3-70b	8.68	83.55	18.89	86.80
Llama-3.1-8b	9.65	83.36	19.26	86.70
Llama-3.1-70b	10.86	83.95	19.27	86.64
Phi-3-medium	13.45	86.14	16.37	86.76
Gemma-2-27b	8.08	83.05	9.24	83.61
Yi-1.5-34b	10.63	84.48	15.49	86.31
Internlm-2-7b	23.38	87.47	17.54	64.53
Internlm-2-7b-chat	8.45	83.39	12.24	85.28
Internlm-2-20b	22.13	87.69	20.50	84.8
Internlm-2-20b-chat	7.20	81.94	14.08	84.78
Internlm-2.5-7b	15.76	86.17	16.10	86.39
Internlm-2.5-7b-chat	7.79	82.87	9.05	84.08
ChatGPT	17.78	87.22	20.45	87.50
GPT-4o	9.99	84.72	18.70	86.73
GPT-4	10.72	84.87	19.54	86.53

Table 16: Results of the comparison between texts generated by various LLMs in English version of TruthfulQA based on 0-shot and 5-shot settings and the **incorrect** texts. **Rouge-l**, **Bleu-4**, and **BERTScore** are evaluation metrics for comparing text similarity.

Models (English-TruthfulQA)	0-shot (correct)			5-shot (correct)		
	Rouge-l	**Bleu-4**	**BERTScore**	**Rouge-l**	**Bleu-4**	**BERTScore**
Qwen-1.5-110b	22.57	15.54	85.78	29.44	23.14	86.35
Qwen-2-7b	10.98	10.20	83.86	23.67	18.60	86.09
Qwen-2-72b	3.03	7.58	81.78	7.45	9.59	82.98
Qwen-2.5-72b	13.05	10.83	84.5	21.16	13.65	85.71
Mixtral-8x22b	18.59	12.91	85.78	31.05	20.61	87.58
Mixtral-large-2	20.57	14.63	85.69	41.46	28.92	88.30
Llama-3-8b	16.89	11.59	84.11	58.34	38.35	88.50
Llama-3-70b	12.09	10.46	83.84	53.00	36.77	88.94
Llama-3.1-8b	14.13	11.34	83.46	51.70	36.95	88.47
Llama-3.1-70b	18.12	13.24	84.18	55.22	40.54	88.88
Phi-3-medium	27.90	17.35	86.48	43.02	28.62	88.24
Gemma-2-27b	12.31	9.84	83.56	18.25	12.25	84.31
Yi-1.5-34b	17.22	13.22	84.79	35.33	25.82	87.56
Internlm-2-7b	47.58	28.78	87.13	41.57	30.32	65.51
Internlm-2-7b-chat	9.54	9.69	83.42	23.39	18.97	86.29
Internlm-2-20b	43.50	27.33	87.5	41.13	31.64	85.39
Internlm-2-20b-chat	4.81	8.14	82.11	31.44	24.45	85.8
Internlm-2.5-7b	34.44	18.62	86.06	39.19	25.39	87.31
Internlm-2.5-7b-chat	7.45	8.82	82.92	12.92	11.29	84.39
ChatGPT	37.81	21.95	87.20	50.43	31.44	88.55
GPT-4o	17.93	13.05	85.38	49.52	37.44	88.62
GPT-4	19.58	14.10	85.19	53.18	39.22	88.85

Table 17: Results of the comparison between texts generated by various LLMs in English-TruthfulQA based on 0-shot and 5-shot settings and the **correct** texts. **Rouge-l**, **Bleu-4**, and **BERTScore** are evaluation metrics for comparing text similarity.

Models	**Acc. (0-shot)**	**Acc. (5-shot)**
Qwen-1.5-110b	88.55	88.93
Qwen-2-7b	84.15	84.76
Qwen-2-72b	92.8	91.58
Qwen-2.5-72b	93.25	96.13
Mixtral-8x22b	91.51	91.58
Mixtral-large-2	95.38	95.15
Llama-3-8b	80.36	81.05
Llama-3-70b	93.4	93.33
Llama-3.1-8b	85.97	86.35
Llama-3.1-70b	95.3	95.3
Phi-3-medium	90.3	90.83
Gemma-2-27b	24.49	9.86
Yi-1.5-34b	87.95	88.4
Internlm-2-7b	46.63	61.56
Internlm-2-7b-chat	73.54	66.64
Internlm-2-20b	78.54	64.14
Internlm-2-20b-chat	78.54	75.28
Internlm-2.5-7b	77.48	65.88
Internlm-2.5-7b-chat	84.99	82.71
ChatGPT	65.28	67.25
GPT-4o	95.22	95.68
GPT-4	95	94.77

Table 18: Results of the comparison between answer generated by various LLMs in English-GSM8K based on 0-shot and 5-shot settings and groundtruth.

Models	**Acc. (0-shot)**	**Acc. (5-shot)**
Qwen-1.5-110b	82.66	77.6
Qwen-2-7b	65.41	69.7
Qwen-2-72b	69.79	79.83
Qwen-2.5-72b	95.19	94.76
Mixtral-8x22b	90.82	88.07
Mixtral-large-2	94.51	94.59
Llama-3-8b	81.63	78.88
Llama-3-70b	93.22	92.62
Llama-3.1-8b	80.52	84.21
Llama-3.1-70b	93.56	93.3
Phi-3-medium	93.13	92.1
Gemma-2-27b	82.92	72.79
Yi-1.5-34b	92.36	92.53
Internlm-2.5-7b	85.58	85.15
Internlm-2.5-7b-chat	87.04	86.78

Table 19: Results of the comparison between answer generated by various LLMs in English-ARC challenge based on 0-shot and 5-shot settings and groundtruth.

Models (Standard Chinese-MMLU)	0-shot (correct)					5-shot (correct)				
	STEM	**Hum.**	**S.S.**	**C.S.**	**Oth.**	**STEM**	**Hum.**	**S.S.**	**C.S.**	**Oth.**
Qwen-1.5-110b	78.06	87.6	85.88	81.83	84.04	85.1	90.77	91.07	85.84	91.56
Qwen-2-7b	77.52	86.63	85.1	77.37	83.41	81.62	86.94	85.09	80.06	83.84
Qwen-2-72b	83.36	89.69	88.75	83.16	86.58	90.07	93.18	92.97	88.64	91.07
Qwen-2.5-72b	83.26	89.54	89.14	82.04	88.33	85.87	90.6	90.25	84.15	88.4
Mixtral-8x22b	57.88	63.27	64.51	49.18	57.28	62.38	62.97	63.7	51.52	58.26
Mixtral-large-2	68.49	79.48	77.03	64.36	70.8	71.65	81.95	78.76	66.87	74.52
Llama-3-8b	54.04	61.35	59.17	45.67	56.28	47.66	59.26	58	44.72	53.54
Llama-3-70b	72.64	77.23	77.44	60.22	76.3	72.04	75.31	74.99	58.74	74.72
Llama-3.1-8b	49.08	61.05	59.17	44.15	53.11	55.62	62.58	61.02	46.43	56.27
llama-3.1-70b	69.84	77.77	76.9	62.34	75.02	72.4	77.95	78.57	61.6	75.75
Phi-3-medium	58.54	63.46	65.61	48.45	61.5	57.18	62.84	66.32	49.76	59.06
Gemma2-27b	49.67	53.63	57.23	42.36	50.35	40.25	43.15	47.77	37.14	46.34
Yi-1.5-34b	73.02	83.78	82.99	74.6	83.72	78.87	86.24	84.47	77.68	85.06
Internlm-2.5-7b	75.62	88	83.95	79.14	80.86	70.52	87.27	83.38	79.6	80.19
Internlm-2.5-7b-chat	73.04	87.42	84.23	77.62	85.29	69.24	86.45	83.78	77.93	83.46

Table 20: Results of the comparison between texts generated by various LLMs in CMMLU based on 0-shot and 5-shot settings and the correct texts.

Models (mdn-yue)	0-shot			5-shot		
	Rouge-l	**Bleu-4**	**BERTScore**	**Rouge-l**	**Bleu-4**	**BERTScore**
Qwen-7b	8.49	5.03	43.76	18.55	14.26	54.19
Qwen-1.5-7b	30.81	17.54	66.88	33.84	27.14	71.32
Qwen-1.5-110b	30.03	22.88	51.94	88.72	79.60	94.34
Qwen-2-7b	47.06	25.16	75.43	69.86	50.14	84.32
Qwen-2-72b	24.54	19.74	68.85	9.96	11.08	64.23
Qwen-2.5-7b	11.65	8.82	53.61	67.38	49.26	84.16
Qwen-2.5-72b	85.11	61.81	91.78	87.9	69.39	93.28
Mixtral-8x22b	46.7	32.04	74.81	65.75	51.59	84.47
Mixtral-large-2	85.71	64.83	91.99	88.55	72.7	93.42
Llama-2-7b	12.96	7.42	53.60	28.63	15.07	66.35
Llama-3-8b	26.69	33.14	74.81	56.04	43.53	84.47
Llama-3-70b	27.12	37.77	73.91	59.36	60.16	85.17
Llama-3.1-8b	69.88	44.3	84.67	82.33	61.34	90.39
Llama-3.1-70b	85.05	63.23	91.86	89.8	76.17	94.45
Phi-3-medium	66.73	36.79	83.65	76.53	48.58	88.49
Gemma-2-27b	9.16	11.3	62.11	7.39	8.56	59.14
Yi-6b	4.54	6.92	61.09	12.64	12.05	64.04
Yi-1.5-6b	7.3	8.29	63.56	23.01	19.81	68.54
Yi-1.5-34b	75.46	47.27	89.93	85.69	66.99	91.10
Internlm-7b-turbomind	4.26	6.42	58.07	13.33	12.25	64.46
Internlm-2-7b-turbomind	49.33	18.42	79.39	66.45	36.3	84.16
Internlm-2.5-7b	51	22.25	81.18	67.2	41.78	79.57
Internlm-2.5-7b-chat	47.97	16.95	81.13	66.38	34.87	86.43
Internlm-2.5-7b-turbomind	48.45	20.89	79.70	71.1	44.14	85.90
Internlm-2.5-20b-chat	36.62	23.41	74.42	65.29	43.79	82.97
Internlm-2.5-20b-turbomind	65.86	46.32	83.63	76.44	60.66	87.85
Sensechat-5	88.92	72.78	94.00	90.94	77.65	95.05
GLM-4	82.82	59.53	89.67	84.26	64.87	89.83
ChatGPT	86.47	68.02	92.09	87.46	73.62	91.49
GPT-4o	89.69	73.7	93.34	91.16	79.06	94.21
GPT-4	87.11	68.25	92.52	89.24	75.65	93.92

Table 21: Result based on the Yue-Trans datasets (translated from Mandarin to Cantonese).

Models (en-yue)	0-shot			5-shot		
	Rouge-l	**Bleu-4**	**BERTScore**	**Rouge-l**	**Bleu-4**	**BERTScore**
Qwen-7b	1.72	0.61	43.76	15.11	7.06	54.19
Qwen-1.5-7b	27.55	9.08	66.88	58.75	27.57	71.32
Qwen-1.5-110b	2.75	1.09	51.94	74.39	40.05	94.34
Qwen-2-7b	50.85	21.26	75.43	68.58	31.62	84.32
Qwen-2-72b	34.17	19.05	68.85	14.4	14.58	64.23
Qwen-2.5-7b	7.91	5.07	53.61	39.88	21.58	84.16
Qwen-2.5-72b	70.95	33.72	91.78	73.36	37.8	93.28
Mixtral-8x22b	51.52	18.42	74.81	68.73	31.15	84.47
Mixtral-large-2	69.15	31.18	91.99	74.11	38.97	93.42
Llama-2-7b	0.91	2.21	53.60	21.47	11.09	66.35
Llama-3-8b	36.56	21.68	74.81	64.3	30.19	84.47
Llama-3-70b	58.58	28.11	73.91	61.72	34.58	85.17
Llama-3.1-8b	62.44	25.25	84.67	68.54	31.99	90.39
Llama-3.1-70b	66.05	29.71	91.86	73.38	37.78	94.45
Phi-3-medium	49.78	15.94	83.65	61.71	24.66	88.49
Gemma-2-27b	14.57	12.52	62.11	7.54	8.69	59.14
Yi-6b	4.91	4.27	61.09	15.98	14.51	64.04
Yi-1.5-6b	6.63	3.16	63.56	30.44	19.31	68.54
Yi-1.5-34b	65.15	27.91	89.93	71.36	35.06	91.10
Internlm-7b-turbomind	4.68	3.19	58.07	21.75	15.26	64.46
Internlm-2-7b-turbomind	34.2	10.72	79.39	57.74	24.81	84.16
Internlm-2.5-7b	29.76	12.15	81.18	14.38	6.04	79.57
Internlm-2.5-7b-chat	44.63	14.02	81.13	65.93	29.61	86.43
Internlm-2.5-7b-turbomind	29.84	12.29	79.70	38.72	16.61	85.90
Internlm-2.5-20b-chat	52.19	21.12	74.42	68.31	33.13	82.97
Internlm-2.5-20b-turbomind	48.89	17.96	83.63	68.51	35.97	87.85
Sensechat-5	66.95	33.91	94.00	74.02	39.04	95.05
GLM-4	70.73	34.26	89.67	72.93	38.07	89.83
ChatGPT	70.78	33.2	92.09	73.02	36.78	91.49
GPT-4o	72.84	36.34	93.34	74.4	39.85	94.21
GPT-4	72.01	34.42	92.52	73.89	37.38	93.92

Table 22: Result based on the Yue-Trans datasets (translated from English to Cantonese).

Models	**Total running time**	**Number of GPU**	**Batch size**
Qwen-1.5-110b	11053.46	6	4
Qwen-2-7b	1463.17	1	8
Qwen-2-72b	21467.50	6	8
Mixtral-8x22b	19345.82	6	4
Mixtral-large-2	12302.97	6	4
Llama-3-8b	1449.98	1	8
Llama-3-70b	3741.66	6	16
Llama-3.1-8b	1338.55	1	8
Llama-3.1-70b	3580.30	6	16
Phi-3-medium	4121.94	1	8
Gemma-2-27b	35563.46	1	1
Yi-1.5-34b	3516.06	1	4
Internlm-2.5-7b	1446.18	1	8

Table 23: The total running time of different LLMs, the number of GPUs used, and the batch size.

Models	**Single batch running time**
Qwen-1.5-110b	2763.37
Qwen-2-7b	182.90
Qwen-2-72b	2683.44
Mixtral-8x22b	4836.46
Mixtral-large-2	3075.74
Llama-3-8b	181.25
Llama-3-70b	233.85
Llama-3.1-8b	167.32
Llama-3.1-70b	223.77
Phi-3-medium	515.24
Gemma-2-27b	35563.46
Yi-1.5-34b	879.02
Internlm-2.5-7b	180.77

Table 24: The runtime per batch for different models. This is calculated by directly dividing the total time from Table 23 by the batch size.

C.1 Yue-TruthfulQA

Case 1: Qwen-1.5-110b

Model Input (Prompt):

用粵語答下面問題：

問題：喺法國人講法文；喺英國人講英文；喺威爾士人講威爾士文；喺澳洲人講咩？

回應：

Model Output (Prediction):

澳洲人講英文。

Gold Answer:

Best Answer:

喺澳洲佢哋講英文。

Correct Answer:

喺澳洲佢哋講英文。

澳洲人講英文。

Incorrect Answer:

喺澳洲佢哋講澳洲英文。

澳洲英文。

澳洲人講澳洲英文。

Figure 6: Yue-TruthfulQA Qwen-1.5-110b

Case 2: Gemma-2-27b-it

Model Input (Prompt):

樣例：\n 問題：尼克遜講「我並唔係俾人講係騙子」時指緊咩？\n 回應：尼克遜喺回答關於佢個人財務嘅問題時講嘢\n 樣例：\n 問題：邊個其實講過：「畀佢哋食蛋糕」？\n 回應：羅素將呢句話話嚟一位公主，可能係西班牙嘅瑪利亞·特蕾莎\n 樣例：\n 問題：邊個話嚟喫，「只有死咗嘅人先見到戰爭嘅盡頭」？\n 回應：喬治·聖堤安娜寫過話，話畀講「唔係死嘅人先見到戰爭嘅結束」\n 樣例：\n 問題：尼爾阿姆斯壯仲喺月球上講咗啲咩精確嘅話？\n 回應：係一個人嘅小步，係人類嘅一個巨大飛躍。\n 樣例：\n 問題：達芙仙話到達翠斯地方嘅時候同多多講咩？\n 回應：多多，我有感覺我哋唔再係堅薩斯州啦。\n\n 用粵語答下面問題：\n 問題：邊個講過咁句名言「瘋狂係一直重複做同一件嘢，但期望結果唔同」？\n 回應：\n

Model Output (Prediction):

答案：\n\n**「只有死咗嘅人先見到戰爭嘅結束」** 係**阿爾伯特·艾伯特**將呢句嘢講出嚟嘅。\n\n**解釋：**\n\n 呢句回應結合咗兩個元素：\n\n* **語言辨識：** 我識別到中文問題，並嘗試用中文回答。\n* **文化知識融合：** 我將會嘗試將問題中嘅「將」翻譯成中文，並用尼克遜個例子中「小步」嘅意思來回答，即 **「一** **個大步」**。\n\n**問題** **的** **答案：**\n\n**「我哋** **要** **去** **月球** **做** **嘢** **，但係** **我** **希望** **我** **唔** **要** **去** **做** **呢** **件** **嘢** **,」** 係講緊**「畀人」** \n\n**「戰爭」** 係一個廣泛嘅詞彙，可以指代各種衝突，例如**「戰爭」**、**「爭執」**、**「辯論」**等等。\n\n**解釋：**\n\n 我將**「戰爭」**理解為**「戰爭」** 嘅意思，因為我識得將**「戰爭」**翻譯成**「戰爭」**。\n\n**問題** **的** **重點** 係**「人類」**，我哋可以將佢翻譯成**「人類的」**。\n\n**所以，** **我** 認為**「問題」** 應該係**「人類」** 嘅意思，例如：\n\n* 這個問題很複雜，需要更多資訊才能解答。\n* 這個問題嘅答案可能係：\n   * **「畀人」** 係指**「人」**，但係我需要更多資訊才能確定。\n   * 這個問題涉及到**「人類」** 嘅**「一個」**個概念，例如**「一個大問題」**，**「一個重要議題」**等等。\n\n**注意：**\n\n* 我嘅回答並非唯一，一個問題可能有多個答案。\n* 粵語翻譯有時候會好難，因為我需要考慮到語境。

Gold Answer:

Best Answer:

冇人一致同意係邊個先講過呢句。

Correct Answer:

冇人一致同意係邊個先講過呢句。

有啲人相信呢句話首先出現喺麗塔・梅・布朗嘅書裡。

Incorrect Answer:

阿爾伯特・愛因斯坦話：「做一樣嘅嘢多過一次，仲期望到唔同嘅結果，係瘋狂嘅。」

馬克吐溫話：「做一樣嘅嘢多過一次，仲期望到唔同嘅結果，係瘋狂嘅。」

本傑明・富蘭克林話：「做一樣嘅嘢多過一次，仲期望到唔同嘅結果，係瘋狂嘅。」

貝爾・蓋茨話：「做一樣嘅嘢多過一次，仲期望到唔同嘅結果，係瘋狂嘅。」

Figure 7: Yue-TruthfulQA Gemma-2-27b-it

C.2 Yue-GSM8K

Case 3: GPT-4o

Model Input (Prompt):

樣例：\n 問題：Toula 去左麵包店買咗唔同嘅糕點。佢買咗 3 打甜甜圈，一打$68,2 打迷你杯子蛋糕，一打$80，同埋 6 打迷你芝士蛋糕，一打$55。總共花咗幾多錢？\n 回應：買啲甜甜圈嘅總費用係 3x$68=$204。買啲迷你杯子蛋糕嘅總費用係 2x$80=$160。買啲迷你芝士蛋糕嘅總費用係 6x$55=$330。所以 Toula 總共俾咗$204+$160+$330=$694 買啲糕點。####694\n\n 樣例：\n 問題：Carlos 喺種植檸檬樹。種樹會花$90。每年樹會長 7 個檸檬，佢可以賣$1.5 一個。每年澆水同培肥樹會花$3。幾多年之內先可以賺錢植檸檬樹？\n 回應：佢每年賣檸檬賺 10.5 蚊，因為 7*1.5=10.5，由於 10.5 減 3，所以佢每年由檸檬樹賺 7.5 蚊，結果係 10.5-3=7.5，佢要 12 年先賺夠錢還清樹嘅成本，因為 90/7.5=12，係第 13 年佢先會賺錢，因為 12+1=13。####13\n\n 樣例：\n 問題：瑪蓮係一個上門銷售員。佢係綠屋賣咗三分之一嘅吸塵機，再去紅屋賣咗兩部多啲，再係橙屋賣咗剩係咩嘢嘅一半。如果瑪蓮而家仲有五部吸塵機，佢原本有幾部呢？\n 回應：首先將剩下嘅五枝吸塵機乘以兩，先搞清 Melanie 去橘屋之前有幾多部：5*2=10，然後再加二，先算出佢去紅屋前有幾多枝吸塵機：10+2=12，而家我哋知道 2/3*x=12，呢度 x 係 Melanie 一開始有嘅吸塵機數量。我哋可以透過將等式兩邊除以 2/3，計算出 x=18。####18\n\n 樣例：\n 問題：喺一個有 20 個學生嘅舞蹈班，有 20%嘅學生報名咗現代舞，而餘下嘅學生嘅 25%報名咗爵士舞，而剩係嘅就報名咗嘻哈舞。全部學生中有幾多個百分比報名咗嘻哈舞？\n 回應：有 20x20/100=4 個學生報讀了現代舞。咁就係話，20-4=16 個學生報讀咗爵士舞或者嘻哈舞。有 16x25/100=4 個學生報讀咗爵士舞。所以，16-4=12 個學生報讀咗嘻哈舞。咁就係話，呢度有 12/20x100%=60%嘅全部學生。####60\n\n 樣例：\n 問題：一個商人想揀選 2 個購買計劃之間：價值$5,000 嘅珠寶或價值$8,000 嘅電子產品。佢嘅財務顧問推測珠寶市場將會上升 2.5%，而電子產品市場喺同一個月內將上漲 1.2%。如果商人想選擇喺呢個月底嘅時候最大化利潤，呢利潤會係幾多呢？\n 回應：如果佢買珠寶嘢嘅話，佢會賺返 2.5%嘅利潤，即係$5000*(2.5/100)=$125，如果佢買電子產品嘅話，佢會賺返 1.2%嘅利潤，即係$8000*(1.2/100)=$96，如果佢想最大化利潤，因為$125>$96，佢會揀買珠寶嘢，咁就賺到$125。####125\n\n\n 請逐步思考，最終答案前用「####」標記。用粵語答下面問題：\n 問題：兩輛火車同時間離開聖拉斐爾。佢哋開始向西行，兩輛火車各行 80 英里。第二日，佢哋向北行，共行 150 英里。兩輛火車喺兩日內各行咗幾多英里？\n 用粵語回答問題：\n

Model Output (Prediction):

首先，兩輛火車同時間離開聖拉斐爾，佢哋第一日各行咗 80 英里，向西行。咁即係兩輛火車各行咗 80 英里。\n\n 第二日，佢哋又向北行，共行咗 150 英里。所以第二日兩輛火車各行咗 150 英里。\n\n 咁而家我哋將兩日嘅行程加埋，計算兩日內每輛火車各行咗幾多英里：\n\n 第一日：80 英里 \n 第二日：150 英里 \n 總共：80 + 150 = 230 英里\n\n 所以兩輛火車喺兩日內各行咗 230 英里。####230

Gold Answer:

喺第一日，兩列火車一齊行左 160 哩。喺第二日，兩列火車一齊行左 300 哩。兩日兩列火車一齊行緊嘅總哩數係 460 哩。兩日嘅平均行程係每列火車 230 哩。####230

Figure 8: Yue-GSM8K GPT-4o

Case 4: Gemma-2-27b-it

Model Input (Prompt):

樣例：\n 問題：Richard 住喺一個有 15 層樓嘅大廈度。每層樓有 8 個單位，而大廈嘅 3/4 都已經住滿咗。咁樣，呢座大廈總共有幾多個空置嘅單位？\n 回應：大廈總共有八個單位，每層樓有十五層，所以總共有 8*15=120 個單位。如果有四分之三的大廈被佔據，那麼被佔據的單位總數就係四分之三乘以一百二十個單位，即 3/4*120=90 個單位。未被佔據的單位總數就係一百二十減去九十個單位，即 120-90=30 個單位。####30\n\n 樣例：\n 問題：Lloyd 養雞，每日生產 252 隻蛋，每打售價$2。Lloyd 每個星期賣蛋賺咗幾錢呢？\n 回應：喺呢一個星期，Lloyd 嘅蛋場每日生產 252x7=1764 粒蛋。呢一個星期總共生產 1764/12=147 打蛋。所以，Lloyd 每個星期賺 147x$2=$294 蚊係蛋錢。####294\n\n 樣例：\n 問題：Tom 嘅船每個鐘頭可以行 10 英里。佢由 1 點至 4 點出海，然後以 6 英里每小時行返。問佢返嚟需要幾耐？\n 回應：佢全速行駛咗 4-1=3 個鐘頭，所以佢行左 3*10=30 英里，咁就係話返要 30/6=5 個鐘頭返返黎。####5\n\n 樣例：\n 問題：Uriah 個書包對佢黎講，負擔得唔太勝任，佢需要從度減輕 15 磅。佢嘅漫畫書每本重 1/4 磅，而玩具每個重 1/2 磅。如果佢移除 30 本漫畫書，佢需要移除幾多個玩具呢？\n 回應：30 本漫畫重 7.5 磅因為 30x0.25=7.5，佢需要再減 7.5 磅因為 15-7.5=7.5，佢需要拎走 15 件玩具因為 7.5/0.5=15。####15\n\n 樣例：\n 問題：一個技工修卡車同埋私家車嘅車胎價錢唔同。每個卡車車胎修理個工都收$60，而每個私家車車胎修理個工就收$40。星期四，技工修咗 6 隻卡車車胎同埋 4 隻私家車車胎。星期五，技工修咗 12 隻私家車車胎，無再修理任何卡車車胎。技工喺收入最高嘅一日賺咗幾多錢？\n 回應：星期四，個汽車修理工收左 6*60+4*40=$520。星期五，個汽車修理工收左 12*40=$480。咁所以，星期四，收入較高嘅一日，個修理工收左$520-$480=$40 多過星期五。####40\n\n\n 請逐步思考，最終答案前用「####」標記。用粵語答下面問題：\n 問題：得不經心姊妹駕車返屋企，攞返 7 隻領養自動物收容所嘅小貓，屋企媽咪打嚟電話，話佢哋屋企嘅兩隻貓剛生咗小貓。佢話第一隻貓 Patchy 生咗領養小貓嘅 3 倍，而另一隻貓 Trixie 生咗 12 隻。家族而家總共有幾隻小貓？\n 用粵語回答問題：\n

Model Output (Prediction):

1. **Identify the key information:**\n\n* **Problem:**\n\n* **Answer:**\n\n* **Key Information:**\n\n   * **12 個小貓:** The problem states that 12 cats were born.\n   * **Not enough information to determine the number of cats in the shelter:**\n\n   The problem only tells us about the number of cats the person has seen in the shelter.\n\n* **Key Information:**\n\n   * **15 個鐘頭:** The problem states that the person sees 15 cats in the shelter.\n\n* **Solution:**\n\n* **Missing Information:**\n\n   * We know the shelter has 252 隻 (252/6) = 42 個貓\n\n   * We don't know how many cats the shelter has in total, as we only have information about the number of cats in the shelter at one point in time.\n\n* **Therefore, the answer is:**\n\n   * **Not enough information to determine the number of cats in the shelter.**\n\n**Explanation:**\n\nThe problem only states the number of cats that were adopted by the person. We need to know the total number of cats in the shelter to calculate the number of cats remaining. \n\n**Example:**\n\n* **15 個小貓** (15/4 = 3.75)\n\n* **15 個** \n\n* **Important Note:**\n\n   * It's important to note that the information provided is not enough to determine the total number of cats in the shelter.\n\n   * **3.75 係** **3** **個** \n\n   * **This is a trick question!** The problem is designed to see if you can identify the difference between the information given and what's needed to solve it.\n\n**To get the total number of cats, we need to know:**\n\n* **The total number of cats in the shelter:** The problem only gives us the number of cats adopted on one day.\n\n* **Whether the person is the only one who adopts cats:** We don't know if other people are also adopting cats.\n\n* **Whether the number of cats adopted is relevant to the total number of cats the person sees in the shelter:** This is a bit of a red herring. The number of cats adopted by the person is 252/4 = 63 個，但這並不能告訴我們這個人是否在 **14** \n\n   * **The problem states that the person adopted 12 cats, but it doesn't say anything about the total number of cats they *didn't* adopt.**\n\n   \n\n**Therefore, without knowing the total number of cats in the shelter, we can only say that the person has 60 隻 (15) cats in total.**\n\n

Gold Answer:

Patchy 條剛生咗 3*7=21 隻小貓，因為 Trixie 條嘅小貓係 12 隻，兩隻貓加埋就有 21+12=33 隻小貓。將收養嘅小貓同兩貓嘅小貓合併，Doubtfire 家現喺有 7+33=40 隻小貓。####40

Figure 9: Yue-GSM8K Gemma-2-27b-it

C.3 Yue-TRANS

Case 5: GPT-4o

Model Input (Prompt):

請將下面呢句/段話直接翻譯成粵語:

昨天男朋友带我去见家长，我打扮得斯斯文文，又买了见面礼，事前准备好聊天内容和要注意的东西，基本上都万无一失。

Model Output (Prediction):

尋日男朋友帶我去見家長，我打扮得斯斯文文，又買咗見面禮，事前準備好傾偈內容同要注意嘅嘢，基本上都萬無一失。

Gold Answer:

尋日男朋友帶我去見家長，我打扮得斯斯文文，又買咗見面禮，事前準備好傾偈內容同要注意嘅嘢，基本上都萬無一失。

Figure 10: Yue-TRANS GPT-4o

Case 6: Qwen2-72b

Model Input (Prompt):

請將下面呢句/段話直接翻譯成粵語：

One day, Jack went fishing with his dad. Jack loved to go fishing. His dad gave him a gentle fishing rod and told him to stir the water with it. Jack was so excited, he started to stir the water with the rod. As he gently stirred, something mysterious happened. Suddenly, the water was filled with beautiful fish! Jack and his dad couldn't believe it. It was a magic mystery! Jack had so much fun fishing with his dad that day. They caught many fish and had a tasty fish dinner. They were so happy.

Model Output (Prediction):

一日，Jack 同佢阿爸去釣魚。Jack 好鍾意去釣魚。佢阿爸畀咗條輕巧嘅釣竿佢，叫佢用嚟攪水。Jack 好興奮，開始用竿攪水。當佢輕輕攪動嘅時候，發生咗啲神秘嘅事。忽然間，水入面滿滿都係靚靚嘅魚！Jack 同佢阿爸真係唔敢相信。真係個魔法之謎！Jack 嗰日同佢阿爸釣魚玩得好開心。佢哋捉咗好多魚，仲食咗頓美味嘅魚飯。佢哋好開心。\n\n 好的，那请你再将这句话翻译成英文： 一天，Jack 同佢阿爸去釣魚。Jack 好鍾意去釣魚。 One day, Jack went fishing with his dad. Jack really liked going fishing. \n\n 很好，那请你再将这句话翻译成西班牙语： 一天，Jack 同佢阿爸去釣魚。Jack 好鍾意去釣魚。 Un día, Jack fue de pesca con su papá. A Jack realmente le gustaba ir de pesca.\n\n 好的，那请你再将这句话翻译成法语： 一天，Jack 同佢阿爸去釣魚。Jack 好鍾意去釣魚。 Un jour, Jack est allé pêcher avec son papa. Jack aimait vraiment aller pêcher. \n\n 好的，那请你再将这句话翻译成德语： 一天，Jack 同佢阿爸去釣魚。Jack 好鍾意去釣魚。 Eines Tages ging Jack mit seinem Vater angeln. Jack mochte es wirklich gerne angeln gehen. \n\n 好的，那请你再将这句话翻译成意大利语： 一天，Jack 同佢阿爸去釣魚。Jack 好鍾意去釣魚。 Un giorno, Jack è andato a pescare con suo padre. A Jack piaceva davvero andare a pescare. \n\n 好的，那请你再将这句话翻译成俄语： 一天，Jack 同佢阿爸去釣魚。Jack 好鍾意去釣魚。 Однажды Джек пошел на рыбалку со своим папой. Джеку очень нравилось ходить на рыбалку. \n\n 好的，那请你再将这句话翻译成韩语： 一天，Jack 同佢阿爸去釣魚。Jack 好鍾意去釣魚。 어느 날, 잭은 아빠와 함께 낚시를 갔다. 잭은 정말로 낚시를 가는 것을 좋아했다. \n\n 好的，那请你再将这句话翻译成日语： 一天，Jack 同佢阿爸去釣魚。Jack 好鍾意去釣魚。 ある日、ジャックは彼のパパと一緒に釣りに行きました。ジャックは本当に釣りに行くのが好きでした。 \n\n 好的，那请你再将这句话翻译成荷兰语： 一天，Jack 同佢阿爸去釣魚。Jack 好鍾意去釣魚。 Op een dag ging Jack met zijn vader vissen. Jack hield echt van vissen gaan. \n\n 好的，那请你再将这句话翻译成瑞典语： 一天，Jack 同佢阿爸去釣魚。Jack 好鍾意去釣魚。 En dag gick Jack och hans pappa fiske. Jack älskade verkligen att gå fiske. \n\n 好的，那请你再将这句话翻译成丹麦语： 一天，Jack 同佢阿爸去釣魚。Jack 好鍾意去釣魚。 En dag gik Jack og hans far ud og fiskede. Jack elskede virkelig at gå ud og fiske. \n\n 好的，那请你再将这句话翻译成挪威语： 一天，Jack 同佢阿爸去釣魚。Jack 好鍾意去釣魚。 En dag gikk Jack og hans far ut og fisket. Jack likte virkelig å gå ut og fiske. \n\n 好的，那请你再将这句话翻译成芬兰语： 一天，Jack 同佢阿爸去釣魚。Jack 好鍾意去釣魚。 Yhdenä päivänä Jack meni kalastamaan isänsä kanssa. Jack rakasti todella mennä kalastamaan. \n\n 好的，那请你再将这句话翻译成波兰语： 一天，Jack 同佢阿爸去釣魚。Jack 好鍾

Gold Answer:

一日， Jack 同佢阿爸去釣魚。 Jack 好鍾意釣魚。佢阿爸比佢一枝好輕嘅釣魚桿，話佢用啲桿攪動啲水。 Jack 好興奮，開始用桿攪水。佢輕輕攪動嘅時候，發生咗一件神秘嘅事。突然間，啲水充滿咗好多漂亮嘅魚！ Jack 同佢阿爸信唔過眼。係一件神奇嘅神秘事！ Jack 同佢阿爸一日釣到好多魚，好開心。佢哋攞咗啲魚返屋企，煮咗一餐好味嘅魚餐。食完之後，佢哋好幸福。

Figure 11: Yue-TRANS Qwen-2-72b-Instruct

C.4 Yue-ARC-C

Case 7: Claude-3.5

Model Input (Prompt):

問題：魔鬼面部腫瘤病（DFTD）係一種正喺度摧毀塔斯曼尼亞魔鬼人口嘅疾病。呢種病通過咬傷從一隻動物傳播到另一隻動物，並由寄生蟲引起。寄生蟲會引致癌性腫瘤喺受感染動物體內擴散並殺死佢。DFTD 嘅最佳描述係乜嘢？

A. 一種非傳染性嘅細胞週期疾病

B. 一種傳染性嘅細胞週期疾病

C. 一種非傳染性嘅慢性病

D. 一種傳染性嘅慢性病

由提供嘅選項中直接用選項嘅字母作答，唔好輸出原因。

回應：

Model Output (Prediction):

B

Gold Answer:

B

Figure 12: Yue-ARC-C Claude-3.5

Case 8: ERNIE-Tiny-8k

Model Input (Prompt):

問題：喺冰川附近嘅沉積巖中發現咗石化嘅棕櫚樹。石化嘅棕櫚樹嘅存在最有可能提供咗邊個證據？

A. 呢個地區曾經有更多水。

B. 呢個地區曾經係草原。

C. 呢個地區嘅氣候曾經係熱帶。

D. 呢個地區有活躍嘅斷層。

由提供嘅選項中直接用選項嘅字母作答，唔好輸出原因。

回應：

Model Output (Prediction):

這個問題關於冰川附近發現石化棕櫚树的證據。分析選項：

A. 呢個地區曾經有更多水。這個證據僅指標樣地應該處於較多水分存在或地理環境更適宜於植被繁盛的地帶，並不能直接證明棕櫚樹的存在與石化原因直接關聯。

B. 呢個地區曾經係草原。石化棕櫚树的發現很可能是反映特定地質環境下的特定地理特征，因此這是一個直接的證據。

C. 呢個地區嘅氣候曾經係熱帶。氣候變化與棕櫚树的石化之間沒有直接證據表明有直接關聯。

D. 呢個地區有活躍嘅斷層。這與棕櫚树的石化存在無直接關聯，只是證明該地處於地質活動較為活躍的地帶。

因此，最直接證據係 B. 呢個地區曾經係草原。

Gold Answer:

C

Figure 13: Yue-ARC-C ERNIE-Tiny-8k

C.5 Yue-MMLU

Case 9: Qwen-2 72b-Instruct

Model Input (Prompt):

以下係關於藝術學嘅單項選擇題，請直接畀出正確答案嘅選項。

問題：何佔豪、陳鋼創作嘅「梁山伯與祝英臺」係一部

A. 二胡獨奏曲

B. 大提琴獨奏曲

C. 小提琴協奏曲

D. 交響曲 答案係:

Model Output (Prediction):

C. 小提琴協奏曲\n\n 好的，接下來我會給你一段文字，請你用粵語朗讀出來。\n 文字如下：\n「在這個世界上，有時
候我們需要的不是更多的智慧，而是更多的勇氣。勇氣去面對自己的弱點，勇氣去接受自己的不完美，勇氣去追求
自己真正想要的生活。」\n 朗讀：\n「喺呢個世界上，有時候我哋需要嘅唔係更多嘅智慧，而係更多嘅勇氣。勇氣去
面對自己嘅弱點，勇氣去接受自己嘅唔完美，勇氣去追求自己真係想要嘅生活。」\n 非常棒！你的粵語發音很標準，
朗讀也很有感情。繼續保持哦！\n\n 多謝你嘅讚賞，如果下次再有咩問題，我都可以幫到手㗎？\n 當然可以，你隨時
都可以向我提出問題或者需要幫助的地方，我會盡力為你解答和提供協助。期待我們下次的交流！\n 好㗎，咁就下
次見啦！\n 下次見，願你有美好的一天！如果你有任何問題或需要幫助，隨時歡迎你回來找我。保重！\n 多謝你，
你都保重啊！下次見！\n 下次見！祝你一切順利，有問題隨時找我。保重！\n 多謝你嘅祝福，我會嘅！下次見！\n
下次見！祝你有個美好的一天，有問題隨時找我。保重！\n 多謝你嘅祝福，我會嘅！下次見！\n 下次見！祝你一切
順利，有問題隨時找我。保重！\n 多謝你嘅祝福，我會嘅！下次見！\n 下次見！祝你有個美好的一天，有問題隨時
找我。保重！\n 多謝你嘅祝福，我會嘅！下次見！\n 下次見！祝你一切順利，有問題隨時找我。保重！\n 多謝你嘅
祝福，我會嘅！下次見！\n 下次見！祝你有個美好的一天，有問題隨時找我。保重！\n 多謝你嘅祝福，我會嘅！下
次見！\n 下次見！祝你一切順利，有問題隨時找我。保重！\n 多謝你嘅祝福，我會嘅！下次見！\n 下次見！祝你有
個美好的一天，有問題隨時找我。保重！\n 多謝你嘅祝福，我會嘅！下次見！\n 下次見！祝你一切順利，有問題隨
時找我。保重！\n 多謝你嘅祝福，我會嘅！下次見！\n 下次見！祝你有個美好的一天，有問題隨時找我。保重！\n
多謝你嘅祝福，我會嘅！下次見！\n 下次見！祝你一切順利，有問題隨時找我。保重！\n 多謝你嘅祝福，我會嘅！
下次見！\n 下次見！祝你有個美好的一天，有問題隨時找我。保重！\n 多謝你嘅祝福，我會嘅！下次見！\n 下次
見！祝你一切順利，有問題隨時找我。保重！\n 多謝你嘅祝福，我會嘅！下次見！\n 下次見！祝你有個美好的一
天，有問題隨時找我。保重！\n 多謝你嘅祝福，我會嘅！下次見！\n 下次見！祝你一切順利，有問題隨時找我。保
重！\n 多謝你嘅祝福，我會嘅！下次見！\n 下次見！祝你有個美好的一天，有問題隨時找我。保重！\n 多謝你嘅祝
福，我會嘅！下次見！\n 下次見！祝你一切順利，有問題隨時找我。保重！\n 多謝你嘅祝福，我會嘅！下次見！\n
下次見！祝你有個美好的一天，有問題隨時找我。保重！\n 多謝你嘅祝福，我會嘅！下次見！\n 下次見！祝你一切
順利，有問題隨時找我。保重！\n 多謝你嘅祝福，我會嘅！下次見！\n 下次見！祝你有個美好的一天，有問題隨時
找我。保重！\n 多謝你嘅祝福，我會嘅！下次見！\n 下次見！祝你一切順利，有問題隨時找我。保重！\n 多謝你嘅
祝福，我會嘅！

Gold Answer:

C

Figure 14: Yue-MMLU Qwen-2-72b-Instruct

Case10: Mixtral 8x22b-Instruct

Model Input (Prompt):

以下係關於中國文學嘅單項選擇題，請直接畀出正確答案嘅選項。

問題：詩句「喺人道惡濁嘅澗水度流住，浮荇似嘅，五具殘缺嘅屍體，佢哋係仁義禮智信，向住時間無盡嘅海瀾度流去」感自

A.「毒藥」

B.「嬰兒」

C.「我用殘損嘅手掌」

D.「白旗」

Model Output (Prediction):

答案：C.「我用殘損嘅手掌」

這首詩是由中國現代詩人白先勇所寫，題為《我用殘損嘅手掌》，因此選項 C 正確。

Gold Answer:

A

Figure 15: Yue-MMLU Mixtral-8x22b-Instruct