Developing and Utilizing a Large-Scale Cantonese Dataset for Multi-Tasking in Large Language Models

Anonymous ACL submission

Abstract

High-quality data resources play a crucial role in learning large language models (LLMs), particularly for low-resource languages like Cantonese. Despite having more than 85 million native speakers, Cantonese is still considered a low-resource language in the field of natural language processing (NLP) due to factors such as the dominance of Mandarin, lack of cohesion within the Cantonese-speaking community, diversity in character encoding and input methods, and the tendency of overseas Cantonese speakers to prefer using English. In addition, rich colloquial vocabulary of Cantonese, English loanwords, and code-switching characteristics add to the complexity of corpus collection and processing. To address these challenges, we collect Cantonese texts from a variety of sources, including open source corpora, Hong Kong-specific forums, Wikipedia, and Common Crawl data. We conduct rigorous data processing through language filtering, quality filtering, content filtering, and de-duplication steps, successfully constructing a high-quality Cantonese corpus of over 2 billion tokens for training large language models. We further refined the model through supervised fine-tuning (SFT) on curated Cantonese tasks, enhancing its ability to handle specific applications. Upon completion of the training, the model achieves state-of-the-art (SOTA) performance on four Cantonese benchmarks. After training on our dataset, the model also exhibits improved performance on other mainstream language tasks[1].

1 Introduction

High-quality data resources are essential for the advancement of large language models (Jiang et al., 2025), particularly for languages with limited digital resources, such as Cantonese. Although Cantonese boasts over 85 million native speakers (Xiang et al., 2024; Jiang et al., 2025), predominantly located in southern China and among Chinese communities worldwide, it remains classified as a low-resource language within the domain of NLP. This is mainly due to the dominance of Mandarin, the lack of uniformity within the Cantonese-speaking community, and the diversity of character encoding and input methods. In addition, overseas Cantonese speakers tend to use English, which further hinders the development of Cantonese in the NLP domain.

The rich colloquial vocabulary of Cantonese, its English loanwords, and the widespread phenomenon of code-switching make corpus collection and processing more complex. Compared to Modern Standard Chinese, there is a significant disparity between spoken and written Cantonese; many colloquial expressions lack a standardized written form. Furthermore, Cantonese writing involves the conversion between traditional and simplified characters, as well as the use of unique Cantonese characters and words (Yu et al., 2022b; Xiang et al., 2024). These factors increase the difficulty of text data normalization and processing. These challenges have led to a scarcity of high-quality Cantonese corpora, limiting the performance enhancement of LLMs in the Cantonese context.

To address these issues, we collect diverse Cantonese text data to construct a high-quality Cantonese dataset. Data sources include open-source corpora, Hong Kong-specific forums such as LIHKG[2], OpenRice[3], the Cantonese version of Wikipedia[4], and Common Crawl data[5], etc. During the data collection process, we pay attention to the variations in Cantonese usage across different regions and platforms. We utilize custom web crawlers and data extraction tools to efficiently gather large amounts of text from these sources.

To ensure the quality and purity of the data, we

[1] Code and data will be released after the paper is accepted.

[2] https://lihkg.com/

[3] https://www.openrice.com/

[4] https://dumps.wikimedia.org/zh_yuewiki/

[5] https://commoncrawl.org/

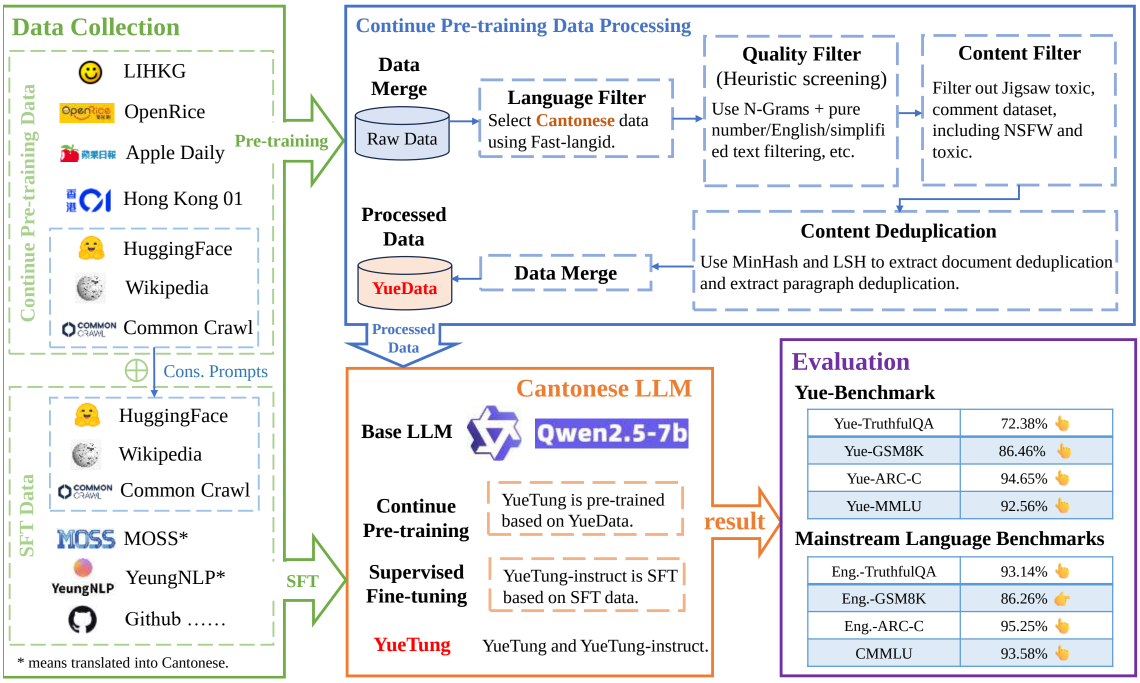

Figure 1: Overview of our work. We construct Cantonese continuous pre-training and SFT data, apply language and quality filters to the former, and derive the latter from it. The base model of YueTung is the **Qwen-2.5-7b** model, which is trained on the YueData. YueTung achieve SOTA performance on Cantonese benchmarks, and its performance on mainstream language benchmarks not only did not decline but actually improved.

establish and executed a stringent data processing workflow. First, we perform language filtering on the collected texts using language identification models to ensure only Cantonese content is retained. Next, we conduct quality filtering based on a series of heuristic rules to select high-quality texts. We also apply specialized classifiers to detect and remove harmful content, such as toxic language and sensitive information. In addition, techniques like MinHash and Locality Sensitive Hashing (LSH) (Paulevé et al., 2010) are used for deduplication, ensuring the corpus's uniqueness and diversity. After continuous pre-training the model on this extensive corpus, we apply SFT using additional Cantonese datasets to further enhance its performance on downstream tasks.

Through these efforts, we successfully build a high-quality Cantonese corpus containing over 2 billion tokens, laying a solid foundation for training large language models. In model performance evaluations, our model achieve industry-leading performance on four Cantonese benchmark tests, accurately handling Cantonese-specific vocabulary and expressions while generating fluent and natural text. Notably, after training with our data, the model also demonstrate performance improvements on other mainstream language tasks, proving that high-quality Cantonese data contributes to the overall performance enhancement of the model.

2 Large-Scale Cantonese Data: YueData

2.1 Pre-Trained Cantonese Data

2.1.1 Pre-Trained Data Collection

Although there are many challenges in gathering Cantonese text, we build a large-scale corpus with a focus on spoken Cantonese, and proceed in two phases: corpus collection and post-processing.

Corpus Collection Cantonese text is gathered from the following sources: (1) open-source corpora; (2) HK transcriptions; (3) HK online publications; (4) HK online forums; (5) Chinese entries from Common Crawl[14].

We focus on leveraging existing open-source resources and subsequently scraping Cantonese data from Hong Kong-centric resources known to be of high quality and which predominantly employ less formal language, resembling spoken Cantonese. The emphasis on these resources is partly due to familiarity with the sources and the status of Hong Kong Cantonese as a de facto standard[15]. In addi-

[14] https://commoncrawl.org/

[15] HK Cantonese has great reach among Cantonese speaking communities as (1) HK has a relatively large and uniform

Source	Record type	Total		Record Char Distribution			
		chars	records	min	median	mean	max
Wikipedia	page contents	40,398,140	137,342	4	91	294	60,004
raptorkwok	cantonese_sentences	664,993,424	30,150,987	0	16	22	2,328
Apple Daily	html articles	54,156,758	81,081	200	535	668	21,298
LIHKG (1-2.8m)	threads	1,582,487,817	2,873,877	7	360	551	80,687
LIHKG (2.8m-3.8m)	sub-threads	839,667,516	29,563,007	0	13	28	4,705
OpenRice	restaurant reviews	490,181,056	1,234,262	0	315	397	477,672

Table 1: Custom scraped corpora, count of characters is language and punctuation agnostic (statistics are indicative).

Corpus Name	Size	Source
CanCorp[6] (Lee and Wong, 1998)	1M tokens	child speech research
HKCAC (Leung and Law, 2002)	170K tokens	phone-in programs and radio
HKCancor[7] (Wong and Luke, 2015)	230K tokens	speech and radio programs
HKCC (Chin, 2015)	1M tokens	audio from 1940-1970 HK movies
UD_Cantonese-HK[8] (Nivre et al., 2017)	–	film subtitles and LegCo proceedings
MyCanCorp[9] (Liesenfeld, 2018)	20 hours of audio	Malaysian Cantonese speech
Common Voice zh-HK[10] (Ardila et al., 2019)	109 hours of audio	Mozilla audio collection program
DRCD[11] (Shao et al., 2019)	10K paragraphs	Wikipedia
CantoMap[12] (Winterstein et al., 2020)	106K tokens	12.8hrs of speech
MDCC[13] (Yu et al., 2022a)	73.6 hours of audio	clean speech from audiobooks

Table 2: Open-source corpora from previous studies.

tion, we interface with Common Crawl to amass a broader corpus of Chinese text.

Open-source corpora: (1) Wikipedia serves as a primary source due to its comprehensive data availability. The Wikipedia pages are systematically archived, categorized by language, and are accessible for batch downloading[16]. Specifically, the Cantonese language content is designated as zh_yuewiki[17], from which the extraction of page contents is straightforward. (2) Prior research: We reviewed data utilized in existing studies on Cantonese linguistics and NLP as referenced in Table 2. These corpora, however, are typically limited in size, often comprising less than a million characters. Given our need for more extensive datasets, these were not included in our study. (3) Huggingface: This platform hosts numerous large-scale datasets[18], though the origins of these datasets are not always transparent. Noteworthy within the context of Cantonese language resources are several datasets, including raptorkwok/cantonese_sentences[19], which includes approximately 30.2 million sentences likely sourced from educational materials featuring colloquial text. Another significant dataset is AlienKevin/LIHKG[20], comprising around 2.8 million discussion threads extracted from LIHKG[21], a popular Hong Kong forum akin to Reddit where users engage in informal discussions, frequently using vernacular and slang.

HK transcriptions: governmental bodies and TV/movie subtitles. Transcriptions of RTHK radio programs and HK Legco discussions exist and are of good quality[22]. Although we find some transcriptions, we cannot find large accessible repositories of them, and so we do not make use of this resource (it could prove fruitful for more resourceful researchers). Another potential source of spoken text includes TV and movie subtitles. One can either (1) **grab pre-generated files**: there are online forums that host `.srt` extension files that media players use to display closed captions, and these files are either created by original content creators or the open-source community; or (2) **generate closed captions with ASR tools**: open-source and

speaker base, (2) emigration from HK seeded many overseas diasporas, and (3) HK was an early producer of Cantonese media (movies, TV dramas, and pop culture), thereby widely consumed and recognized.

[16] https://dumps.wikimedia.org/backup-index.html

[17] https://dumps.wikimedia.org/zh_yuewiki/

[18] https://huggingface.co/datasets

[19] https://huggingface.co/datasets/raptorkwok/cantonese_sentences

[20] https://huggingface.co/datasets/AlienKevin/LIHKG

[21] https://lihkg.com/

[22] Both are government organizations and are transcribed from spoken language, so are unlikely to be vulgar, overly formal, or overly informal

paid ASR tools can extract text from audio while proprietary tools[23] can also be leveraged.

Although this appears to be a promising route, both sources present limitations. Large repositories of `.srt` files are scarce, and the fidelity of subtitles to spoken Cantonese varies greatly[24]. In addition, Cantonese ASR is an active research area, and ASR output would likely require validation before use. Therefore, we decide not to use transcriptions, as it is not the main purpose of this paper.

HK online publications: Apple Daily. We focus on HK publications using less formal Cantonese: (1) **Apple Daily**, the now-defunct publication by Next Digital; (2) **HK01.com**[25], an influential online portal covering popular news in HK.

We extract content only from Apple Daily using 120 web archives[26]. We do not scrape text from HK01.com, though it is a rich source of Cantonese text for NLP researchers.

HK online forums: LIHKG, OpenRice. Online forums are excellent sources of informal Cantonese due to loose language enforcement, allowing users to discuss any topic freely. We focus on two forums[27]: (1) **LIHKG**[28], a popular multi-category forum among HK youths; (2) **OpenRice**[29], a widely used restaurant review platform rich in spoken Cantonese.

We write custom web-scrapers (Truong, 2024b,c)[30] to extend LIHKG coverage from 2.8 million to 3.8 million threads and to collect numerous restaurant reviews from OpenRice.

Chinese Entries from Common Crawl. We use Common Crawl to build our Chinese corpus[31]. Using AWS Athena, we query CDX Index files and employ language identifiers[32] to pinpoint records with Chinese text. We write a custom crawler (Truong, 2024a) to handle this task.

2.1.2 Pre-Trained Data Processing

In our Cantonese data processing workflow, we follow the methodological framework established by Dolma (Soldaini et al., 2024), adhering to the "garbage in, garbage out" principle to ensure data integrity and quality. This process includes several stages specifically tailored for Cantonese: language filtering (ffreemt, 2023), heuristic-based quality filtering (Rae et al., 2022; Raffel et al., 2019), content filtering, and deduplication.

Language Filtering. We use the automatic language identification tool Fast-Langid (ffreemt, 2023), an extension of FastText (Joulin et al., 2016) capable of identifying Cantonese, to build our dataset. We exclude data sources already pre-filtered for Cantonese, like OpenRice, due to imperfections in language models (Blevins and Zettlemoyer, 2022). We proceed to process documents tagged as zh-hant" or zh-yue" in the next stage.

Quality Filtering. To achieve high-quality data, we filter documents using heuristic rules. Rather than relying on model-based evaluations like GPT-3 or LLaMA (Touvron et al., 2023; Brown et al., 2020), we implement Gopher's rule set and other heuristic criteria (Rae et al., 2022; Raffel et al., 2019). Thresholds in these rules, like 0.1 or 90%, are guided by Gopher (Rae et al., 2022). When strict adherence to these thresholds leads to excessive data exclusion (e.g., removing 90% of data), we adjust them downwards. The rules include:

(1) **Symbol-to-word ratio exceeding 0.10**: This criterion is applied to eliminate texts with an excessively high ratio of symbols to words. (2) **Over 90% of lines in a document commencing with a bullet point**: Documents where an overwhelming majority of lines begin with bullet points are filtered out. (3) **Over 30% of lines in a document terminating with ellipses**: Documents with a high frequency of lines ending in ellipses are excluded. (4) **Word count fewer than 50 or exceeding 100,000**: Documents with extreme word counts are removed from the dataset. (5) **Repeated**

[23]Automatic transcription by Google of YouTube videos

[24]Closed captions are often written in formal Chinese to convey meaning, as formal Chinese is more concise and easier to type and read; it is not always spoken Cantonese

[25]`HK01.com`

[26]`https://archive.fart.website/archivebot/viewer/job/201910102213472u3qb`, `https://archive.fart.website/archivebot/viewer/job/202008102032142u3qb`, and `https://archive.fart.website/archivebot/viewer/job/202106170425282u3qb`

[27]Other shortlisted forums not tackled include: `https://www.discuss.com.hk/` (GitHub repo `https://github.com/vanatteveldt/discusshk/blob/master/scrape_discusshk.py`), `https://m.hkgolden.com/`, `https://www.baby-kingdom.com`, and `https://www.babydiscuss.com/`

[28]`https://lihkg.com/`

[29]`https://www.openrice.com/`

[30]Improved upon papatekken's (`https://github.com/papatekken/simple-LIHKG-scraper-with-python`) LIHKG scraper and francoishideyos's (`https://github.com/francoishideyos/openrice_recommendator`) OpenRice scraper

[31]Chinese text represents only 5% of recent Common Crawl indexes

[32]Language annotation was introduced from **CC-MAIN-2018-39** onwards; we use language predictors where it was not provided

Models (7-8b scale)	0-shot			5-shot		
	Rouge-l	Bleu-4	BERTScore	Rouge-l	Bleu-4	BERTScore
Qwen-2.5-7b	18.51	12.28	66.07	6.83	8.07	58.97
Llama-3.1-8b	13.82	10.33	66.97	26.18	**15.20**	70.28
Yi-1.5-6b	1.21	4.60	42.15	1.04	6.15	53.85
Internlm-2.5-7b-chat	7.13	8.00	63.48	4.05	7.19	67.61
YueTung-7b	**33.95**	**12.54**	**71.33**	**35.12**	13.52	**72.38**

Models (> 7-8b scale)	0-shot			5-shot		
	Rouge-l	Bleu-4	BERTScore	Rouge-l	Bleu-4	BERTScore
Qwen-2.5-72b	13.03	9.64	66.94	20.23	12.87	69.53
Mistral-large-2	19.72	13.01	69.06	31.38	18.61	72.07
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
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-4	19.47	13.45	68.99	28.43	16.74	71.26
YueTung-7b	**33.95**	12.54	**71.33**	**35.12**	13.52	**72.38**

Table 3: Results of the comparison between texts generated by YueTung-7b and baselines in **Yue-TruthfulQA** based on 0-shot and 5-shot settings and the ground truth. **Bold face** indicates the best results for the metric.

n-grams, n is greater than 15: We perform repeated 15-grams document removal. If a specific word is continuously repeated more than 15 times, we delete this document. (6) **Text normalization**: For specific datasets, tailored normalization measures are implemented, such as normalizing emojis in the Openrice data and applying blacklist keywords to filter advertisements in the Apple Daily data. For example, the blacklisted words are “此回覆已被刪除” (Translate into English: This response has been deleted), “撰文：阿蘭 支持蘋果深度報道，深入社區，踢爆權貴，即Like蘋果專題及調查組FB專頁！” (Translate into English: Authored by Alan. Support Apple Daily’s in-depth reporting, engage with local communities, and expose those in power. Don’t forget to like the Apple Daily Special Reports and Investigations Team Facebook page!). We also normalize the emoji into text format. In addition, sequences exceeding 1000 characters are automatically truncated to meet processing requirements. Multiple line breaks and separators (e.g., \n and -) are left into one, and during the text normalization process, blacklisted phrases, such as certain advertisements, are also included. These procedures are critical to maintaining data consistency and cleanliness.

Content filter. Subsequently, content filter, toxicity filtering and personally identifiable information (PII) masking, are performed (Soldaini et al., 2024; Jain et al., 2022; Aura et al., 2006; Elazar et al., 2024; Subramani et al., 2023). The toxicity filtering utilizes two classifiers trained on the Jigsaw dataset from Dolma (Soldaini et al., 2024; Jain et al., 2022), specifically designed to detect and remove potentially harmful content (thresholds are set for 1e-2 and 1e-4). The filter can filter around 1% toxic document. PII masking (Aura et al., 2006; Elazar et al., 2024; Subramani et al., 2023) is implemented using regular expressions[33] [34] [35] [36] to obfuscate all email addresses, phone numbers, and IP addresses into "|||IP/EMAIl/PHONE_ADDRESS|||", thereby ensuring the protection of personal information.

Deduplication. We employ MinHash and LSH techniques via the data-sketch framework to eliminate redundant documents and paragraphs (Zhu et al., 2024). Exact document deduplication removes identical documents; exact paragraph dedu-

[33] EMAIL_REGEX = "[.\\s@,?!;:)(]*([^\\s@]+@[^\\s@,?!;:)(]+?)[.\\s@,?!;:)(]?[\\s\n\r]"

[34] PHONE_REGEX = "\\s+\\(?(\\d3)\\)?[-\\.]*(\\d3)[-.]?(\\d2,)"

[35] IP_REGEX = "(?:(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)\\.)3(?:25[0-5]|2[0-4][0-9]|[01]?[0-9][0-9]?)"

[36] URL_REGEX = "(?i)\b((?:https?://|www\\d0,3[.]|[a-z0-9.\\-]+[.][a-z]2,4/)(?:[^\\s()<>]+|\\(([^\\s()<>]+|(\\([^\\s()<>]+\\)))*\\))+(?:\\(([^\\s()<>]+|(\\([^\\s()<>]+\\)))*\\)|[^\\s‘!()\\[\\];:’\".,<>?«»“”‘’]))"

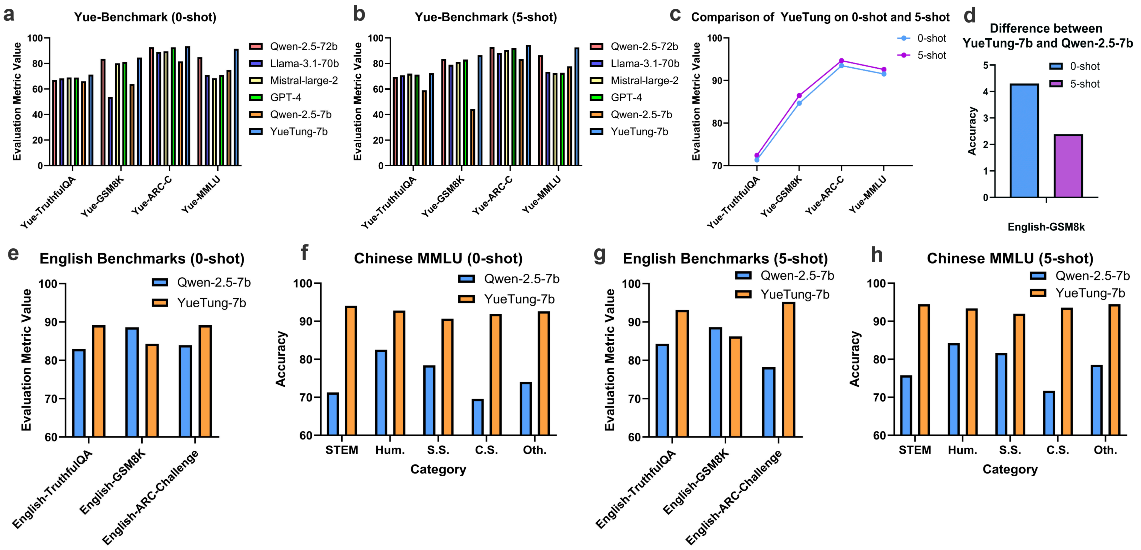

Figure 2: The results of YueTung-7b and baselines on Yue-Benchmark and mainstream language benchmarks. **a** and **b** are YueTung-7b compared with representative LLMs on the Yue-Benchmark (0-shot and 5-shot). **c** is comparison of YueTung-7b on 0-shot and 5-shot. **d** is difference between YueTung-7b and Qwen-2.5-7b on the English-GSM8K. **e**, **f**, **g** and **h** are YueTung-7b compared with base model (Qwen-2.5-7b) on the mainstream language benchmarks (0-shot and 5-shot). In **f**, S.S. stands for Social Sciences, C.S. stands for China Specific, Hum. stands for Humanities and Oth. stands for Other. The complete results are shown in **Table 3**, **5**, **6**, **7**, **8**, **9**, **10**, **11**, **12**, **13**, **14**.

plication removes identical paragraphs or sentences. The deduplication threshold is typically set at 0.5 but may be adjusted to 0.6, depending on the proportion of data removed. For instance, in Common Crawl deduplication, a threshold of 0.5 removes 78.79% of data, while 0.6 removes 44%; thus, we choose 0.6. This time-intensive phase is essential for preserving dataset uniqueness.

This multistage data processing methodology enhances the quality of the continuous pre-training dataset. We filter out around a 1-billion-token high-quality Cantonese corpus, with each filtering step being part of a rigorous data processing pipeline.

YueData (continuous pre-training)	Number of tokens
LIHKG	319,604,833
OpenRice	350,050,930
Apple Daily	23,226,869
HuggingFace	402,925,178
Wikipedia	7,181,350
Common Crawl	269,777,174
YueData (SFT)	**Number of tokens**
All SFT Data	1,289,255,036

Table 4: YueData continuous pre-training and supervised fine-tuning the number of tokens.

2.2 Supervised Fine-Tuning Cantonese Data

SFT data primarily originates from three sources: (1) extraction and construction from pre-trained data; (2) translation of Chinese SFT data into Cantonese; (3) collection of SFT data from GitHub.

Extraction and Construction from Pre-trained Data: We identify and preserve Cantonese dialogue datasets from Huggingface and Common Crawl as SFT data during continuous pre-training. Wikipedia data relevant to knowledge retrieval is processed into a question-and-answer format (Section A.6).

Translating Chinese SFT Data into Cantonese: Translating from Chinese to Cantonese is crucial for obtaining more data, as it's more rational than translating from English (Jiang et al., 2025). We select MOSS's training data and YeungNLP's mathematical data as Chinese sources for translation. Using open-source models for large translation tasks, we choose Llama-3.1-70b based on (Jiang et al., 2025)'s comparison of LLMs in translation quality and speed. We refer to translation prompts from (Jiang et al., 2025), conduct secondary translation, and perform partial reviews to ensure high data quality (Section A.6).

Collecting Suitable SFT Data from GitHub: We collect suitable SFT data from GitHub and incorporate it into YueTung's SFT framework.

Data Leakage Concerns: We focus on data leakage, ensuring pre-trained data learns Cantonese language patterns without involving test data from the Yue-Benchmark (Section A.6).

Models (7-8b scale)	Acc. (0-shot)	Acc. (5-shot)
Qwen-2.5-7b	63.84	44.20
Llama-3.1-8b	63.91	61.64
Yi-1.5-6b	3.94	3.49
Internlm-2.5-7b-chat	65.96	64.67
YueTung-7b	**84.65**	**86.46**
Models (> 7-8b scale)	**Acc. (0-shot)**	**Acc. (5-shot)**
Qwen-2.5-72b	83.62	83.55
Mistral-large-2	80.14	81.27
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-4	81.12	83.02
YueTung-7b	**84.65**	**86.46**

Table 5: Results of the comparison between answer generated by YueTung-7b and baselines in Yue-GSM8K based on 0-shot and 5-shot settings and ground truth. **Bold face** indicates the best results for the metric.

Models (7-8b scale)	Acc. (0-shot)	Acc. (5-shot)
Qwen-2.5-7b	81.64	83.35
Llama-3.1-8b	69.00	67.81
Yi-1.5-6b	34.59	66.70
Internlm-2.5-7b-chat	81.21	79.85
YueTung-7b	**93.48**	**94.65**
Models (> 7-8b scale)	**Acc. (0-shot)**	**Acc. (5-shot)**
Qwen-2.5-72b	92.74	92.91
Mistral-large-2	89.50	90.61
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	88.90	88.73
ChatGPT	69.68	70.71
GPT-4	92.66	92.06
YueTung-7b	**93.48**	**94.65**

Table 6: Results of the comparison between answer generated by YueTung-7b and baselines in Yue-ARC-C based on 0-shot and 5-shot settings and ground truth. **Bold face** indicates the best results for the metric.

3 Experiment

3.1 Experiment Details

Regarding model training, the YueTung model is based on Qwen-2.5-7b[37], which is pre-trained and SFT is conducted based on the YueData dataset. Model evaluation is conducted using the four Yue-Benchmarks (Jiang et al., 2025).

For experimental settings, we implement Yue-Tung model with PyTorch (Paszke et al., 2019) on eight NVIDIA A100-80G GPUs, and train the model using AdamW optimizer (Loshchilov and Hutter, 2017) with a batch size of 2. We vary the learning rate during training following (Vaswani et al., 2017). The training time for the YueTung is about three weeks. For inference, we set the temperature as 0.2, and top-p as 1.0.

3.2 Evaluation and Baselines

For Yue-TruthfulQA, we employ automatic evaluation metrics including Rouge-l (Lin, 2004), Bleu-4 (Papineni et al., 2002), and BERTScore (Zhang* et al., 2020). For Yue-GSM8K, Yue-ARC-C, Yue-MMLU, we adopt Accuracy as evaluation metric.

Regarding baselines, we employ LLMs from mainstream series that are either the same size as or larger than YueTung, including LLMs such as Qwen, Llama, Yi, Internlm, Mistral, Phi, Gemma, ERNIE, GLM, Sensechat, and GPT.

[37]https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

4 Results and Analysis

Section A.2 for a more detailed analysis.

4.1 Cantonese Benchmarks

In Yue-TruthfulQA (Table 3), YueTung-7b achieves Rouge-l scores of 33.95% (zero-shot) and 35.12% (five-shot), outperforming all baseline models, including GPT-4 and ChatGPT. Its highest BERTScore indicates superior semantic similarity to the ground truth. For Yue-GSM8K (Table 5), YueTung-7b attains accuracies of 84.65% (zero-shot) and 86.46% (five-shot), significantly exceeding other 7B to 8B models and even surpassing larger models like GPT-4, highlighting strong reasoning capabilities in Cantonese problem-solving. On Yue-ARC-C (Table 6), YueTung-7b achieves accuracies of 93.48% (zero-shot) and 94.65% (five-shot), outperforming all other models, including GPT-4 and GPT-4o, indicating proficiency in challenging Cantonese comprehension tasks. In Yue-MMLU (Table 7), YueTung-7b consistently achieves accuracies above 89%, peaking at 93.36% in STEM (five-shot), leading over larger LLMs like Qwen-2.5-72b and GPT-4, underscoring its comprehensive knowledge base in Cantonese.

Models (7-8b scale)	0-shot					5-shot				
	STEM	Hum.	S.S.	C.S.	Oth.	STEM	Hum.	S.S.	C.S.	Oth.
Qwen-2.5-7b	72.86	81.66	78.25	66.56	75.19	78.05	80.37	78.99	69.82	78.86
Llama-3.1-8b	45.96	58.27	56.08	44.86	53.70	53.45	58.06	58.31	45.86	53.65
Yi-1.5-6b	17.34	35.98	38.77	32.90	25.00	58.53	67.89	66.56	60.00	62.05
Internlm-2.5-7b-chat	64.40	80.92	76.80	70.24	75.02	65.04	80.84	76.79	70.47	75.19
YueTung-7b	**93.01**	**92.54**	**89.84**	**90.81**	**91.55**	**93.36**	**93.27**	**91.04**	**91.77**	**91.85**

Models (> 7-8b scale)	0-shot					5-shot				
	STEM	Hum.	S.S.	C.S.	Oth.	STEM	Hum.	S.S.	C.S.	Oth.
Qwen-2.5-72b	83.72	87.88	87.20	80.68	85.36	83.89	89.70	88.75	82.34	87.42
Mistral-large-2	60.38	76.08	74.92	60.19	70.74	68.50	79.65	78.84	63.85	71.66
Llama-3.1-70b	67.32	76.57	76.93	60.96	73.56	72.23	78.13	78.23	64.16	74.90
Phi-3-medium	45.26	61.42	58.40	45.65	51.33	49.88	59.33	59.35	45.49	53.02
Gemma-2-27b	48.50	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	78.20	80.30
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	82.00	79.88	73.52	74.77
Claude-3.5	66.47	76.84	78.04	60.60	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	84.20	80.07	76.00	78.06
ChatGPT	49.78	58.13	58.74	45.46	52.42	60.28	59.81	60.61	47.50	54.54
GPT-4	67.68	75.29	77.26	60.12	74.46	71.19	76.75	77.56	63.50	74.57
YueTung-7b	**93.01**	**92.54**	**89.84**	**90.81**	**91.55**	**93.36**	**93.27**	**91.04**	**91.77**	**91.85**

Table 7: Results of the comparison between texts generated by YueTung-7b and baselines in Yue-MMLU based on 0-shot and 5-shot settings and the correct texts. S.S. stands for Social Sciences, C.S. stands for China Specific, Hum. stands for Humanities and Oth. stands for Other. **Bold face** indicates the best results for the metric.

YueTung-7b achieves SOTA performance across all Cantonese benchmarks. Its superior results demonstrate strong Cantonese language proficiency, excelling in context understanding, reasoning, and knowledge retrieval. The significant performance gap suggests that the high-quality Cantonese dataset (YueData) and tailored training strategies contribute greatly to its success. YueTung-7b's ability to outperform larger models like GPT-4 emphasizes the importance of language-specific data in low-resource languages.

4.2 Mainstream Language Benchmarks

In English-TruthfulQA (Table 11), YueTung-7b achieves Rouge-l scores of 37.41% (zero-shot) and 63.50% (five-shot), competitive with larger models like ChatGPT and GPT-4. Its high BERTScore indicates effective cross-lingual knowledge transfer. On English-GSM8K (Table 12), YueTung-7b attains accuracies of 84.32% (zero-shot) and 86.26% (five-shot), indicating robust mathematical reasoning in English. For English-ARC Challenge (Table 13), YueTung-7b achieves accuracies of 89.15% (zero-shot) and 95.25% (five-shot), surpassing several larger models, demonstrating effective handling of English multiple-choice questions. On CMMLU in Standard Chinese (Table 14), YueTung-7b achieves overall accuracies of 92.63% (zero-shot) and 94.49% (five-shot), outperforming all other models, including large-scale ones like Qwen-2.5-72b and GPT-4, indicating enhanced capabilities in closely related languages.

YueTung-7b not only excels in Cantonese but also demonstrates strong cross-lingual abilities in English and Standard Chinese. Its performance suggests that high-quality Cantonese data contributes to robust language understanding that generalizes beyond Cantonese. These findings highlight the potential of leveraging low-resource language data to improve overall LLM performance.

5 Conclusion

In this study, we successfully developed and utilized a large-scale Cantonese dataset (YueData) specifically for training and testing large language model. By collecting over 2 billion tokens of Cantonese text from multiple sources, we constructed a high-quality corpus and trained the YueTung model on this foundation. Through rigorous data processing and refined training, the YueTung demonstrated excellent performance in four Cantonese benchmark tests. This not only showcases the quality of the YueData dataset but also validates the effectiveness of our data processing and training strategies.

Limitations

While YueTung-7b exhibits exceptional performance, there are limitations to our current work. For instance, the YueData corpus, though extensive, predominantly comprises text from Hong Kong-specific sources. As a result, the model may be biased toward the linguistic styles, idioms, and colloquialisms prevalent in Hong Kong Cantonese, potentially limiting its generalizability to other Cantonese dialects spoken in different regions.

In addition, despite our rigorous data processing efforts, including language filtering, quality filtering, content filtering, and deduplication, some noise and biases may persist in the dataset. The complexities of Cantonese, such as code-switching with English and the use of non-standard characters, pose challenges that may affect the model's performance in certain contexts or with highly informal language.

We only use AI tools to polish the language of our paper.

Ethics Statement

This paper does not involve ethics-related issues.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *CoRR*, abs/1912.06670.

Tuomas Aura, Thomas A. Kuhn, and Michael Roe. 2006. Scanning electronic documents for personally identifiable information. In *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society*, WPES '06, page 41–50, New York, NY, USA. Association for Computing Machinery.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Ma teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Andy Chin. 2015. A linguistics corpus of mid-20th century hong kong cantonese. *Department of Linguistics and Modern Language Studies, The Hong Kong Institute of Education, Retrieved*, 23(3):2015.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hanna Hajishirzi, Noah A. Smith, and Jesse Dodge. 2024. What's in my big data? *Preprint*, arXiv:2310.20707.

ffreemt. 2023. fasttext-langdetect.

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.

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

Naman Jain, Skanda Vaidyanath, Arun Iyer, Nagarajan Natarajan, Suresh Parthasarathy, Sriram Rajamani, and Rahul Sharma. 2022. Jigsaw: Large language models meet program synthesis. In *ICSE 2022*.

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

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Hun-tak Thomas Lee and Colleen H. Wong. 1998. Cancorp: The hong kong cantonese child language corpus. *Cahiers de Linguistique Asie Orientale*, 27:211–228.

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.

Man Leung and Sam Po Law. 2002. Hkcac: The hong kong cantonese adult language corpus. *International Journal of Corpus Linguistics*, 6:305–325.

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.

Andreas Liesenfeld. 2018. Mycancor: A video corpus of spoken malaysian cantonese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Loïc Paulevé, Hervé Jégou, and Laurent Amsaleg. 2010. Locality sensitive hashing: A comparison of hash function types and querying mechanisms. *Pattern recognition letters*, 31(11):1348–1358.

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.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling language models: Methods, analysis & insights from training gopher. *Preprint*, arXiv:2112.11446.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2019. Drcd: a chinese machine reading comprehension dataset. *Preprint*, arXiv:1806.00920.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *arXiv preprint*.

Nishant Subramani, Sasha Luccioni, Jesse Dodge, and Margaret Mitchell. 2023. Detecting personal information in training corpora: an analysis. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 208–220, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Alfred Kar Yin Truong. 2024a. cc_cached_downloader.

Alfred Kar Yin Truong. 2024b. LIHKG_scraper.

Alfred Kar Yin Truong. 2024c. openrice_scraper.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Grégoire Winterstein, Carmen Tang, and Regine Lai. 2020. CantoMap: a Hong Kong Cantonese MapTask corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2906–2913, Marseille, France. European Language Resources Association.

Lym Wong and Kk Luke. 2015. The hong kong cantonese corpus: design and uses.

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

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.

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

Tiezheng Yu, Rita Frieske, Peng Xu, Samuel Cahyawijaya, Cheuk Tung Yiu, Holy Lovenia, Wenliang Dai, Elham J. Barezi, Qifeng Chen, Xiaojuan Ma, Bertram Shi, and Pascale Fung. 2022a. Automatic speech recognition datasets in Cantonese: A survey and new dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6487–6494, Marseille, France. European Language Resources Association.

Tiezheng Yu, Rita Frieske, Peng Xu, Samuel Cahyawijaya, Cheuk Tung Shadow Yiu, Holy Lovenia, Wenliang Dai, Elham J Barezi, Qifeng Chen, Xiaojuan Ma, et al. 2022b. Automatic speech recognition datasets in cantonese: A survey and new dataset. *arXiv preprint arXiv:2201.02419*.

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*.

Eric Zhu, Vadim Markovtsev, Aleksey Astafiev, Arham Khan, Chris Ha, Wojciech Łukasiewicz, Adam Foster, Sinusoidal36, Spandan Thakur, Stefano Ortolani, Titusz, Vojtech Letal, Zac Bentley, fpug, hguhlich, long2ice, oisincar, Ron Assa, Senad Ibraimoski, Rupesh Kumar, Qin TianHuan, Michael Joseph Rosenthal, Keyur Joshi, Kevin Mann, JonR, Joe Halliwell, and Andrii Oriekhov. 2024. ekzhu/datasketch: v1.6.5.

A Appendix

A.1 Cantonese LLM: YueTung

When training the YueTung-7b model, the continuous pre-training Cantonese data contain some noisy entries, which can adversely affect the training process. To mitigate the impact of these noisy data and facilitate faster convergence, we appropriately decrease the β_2 parameter in the AdamW optimizer (Loshchilov and Hutter, 2017). By reducing β_2, the optimizer places more emphasis on recent gradients, allowing the model to adapt more quickly and minimize the influence of noisy data. The following algorithm outlines the training procedure using the modified AdamW optimizer:

Algorithm 1 Training YueTung model

```
Init params θ, moments m = 0, v = 0, and time step t = 0
for epochs do
    for minibatch (X, Y) do
        t ← t + 1
        Compute grad g = ∇_θ L(θ; X, Y)
        m ← β1 m + (1 − β1) g
        v ← β2 v + (1 − β2) g^2
        m̂ ← m/(1 − β1^t)
        v̂ ← v/(1 − β2^t)
        Update θ ← θ − α ( m̂/(√v̂ + ε) + λθ )
    end for
end for
```

where $L(\theta; X, Y)$ is the loss function (during continuous pre-training, Y may be omitted); decreasing β_2 allows the optimizer to adapt more quickly to recent gradients, mitigating the impact of noisy data; initial moments m and v are zero with bias correction applied; hyper-parameters include α, β_1, β_2, λ, and ϵ; weight decay $\lambda\theta$ is included directly in the update; the algorithm applies to both continuous pre-training and supervised fine-tuning stages.

A.2 Results Analysis

A.2.1 Cantonese Benchmarks

We evaluate YueTung-7b on four Cantonese benchmarks: Yue-TruthfulQA, Yue-GSM8K, Yue-ARC-C, and Yue-MMLU. The results are summarized in Tables 3, 5, 6, and 7, respectively.

About Yue-TruthfulQA, As shown in Table 3, YueTung-7b achieves a Rouge-l score of 33.95% in the zero-shot setting and 35.12% in the five-shot setting, outperforming all baseline models of similar and larger scales. Notably, YueTung-7b substantially surpasses GPT-4 and ChatGPT, which achieve Rouge-l scores of 19.47% and 25.07% in the zero-shot setting, respectively. The BERTScore of YueTung-7b is also the highest among all models, indicating superior semantic similarity to the ground truth. These results demonstrate YueTung-7b's ability to generate truthful and coherent responses in Cantonese.

About Yue-GSM8K, Table 5 presents the accuracy results on Yue-GSM8K, a mathematical reasoning benchmark. YueTung-7b attains an accuracy of 84.65% in the zero-shot setting and 86.46% in the five-shot setting. This significantly exceeds the performance of other 7B to 8B scale models, such as Qwen-2.5-7b, which achieves 63.84% and 44.20% accuracy, respectively. YueTung-7b also outperforms larger models like GPT-4 and GPT-4o, highlighting its strong reasoning capabilities in Cantonese mathematical problem-solving.

About Yue-ARC-C, on the Yue-ARC-C benchmark, which tests knowledge and reasoning in multiple-choice questions, YueTung-7b achieves accuracies of 93.48% (zero-shot) and 94.65% (five-shot), as shown in Table 6. This places it ahead of all other models, including GPT-4o and GPT-4, which achieve accuracies around 92%. The substantial margin indicates YueTung-7b's proficiency in handling challenging Cantonese comprehension tasks.

About Yue-MMLU, YueTung-7b's performance on Yue-MMLU is detailed in Table 7. Across all categories—STEM, Humanities, Social Sciences, Computer Science, and Others—YueTung-7b consistently achieves accuracies above 89%, with the highest being 93.36% in the STEM category for the five-shot setting. Compared to other models, YueTung-7b exhibits a remarkable lead, outperforming larger models like Qwen-2.5-72b and GPT-4 by a significant margin. This consistent performance across diverse subjects underscores YueTung-7b's comprehensive knowledge base and understanding of Cantonese.

Analysis The experimental results on Cantonese benchmarks demonstrate that YueTung-7b achieves SOTA performance across all evaluated tasks. Its superior results in both zero-shot and five-shot settings indicate that the model not only has a strong grasp of the Cantonese language but also excels in understanding context, reasoning, and knowledge retrieval. The substantial performance gap between YueTung-7b and other models of similar scale suggests that the high-quality Cantonese dataset (YueData) and the tailored training strategies significantly contribute to its success.

Moreover, YueTung-7b's ability to outperform much larger models like GPT-4 emphasizes the importance of language-specific data in low-resource languages. The results validate our approach of focusing on data quality and appropriate training techniques to enhance model performance in Cantonese NLP tasks.

A.2.2 Mainstream Language Benchmarks

To assess the generalization capabilities of YueTung-7b beyond Cantonese, we evaluate the model on mainstream language benchmarks, including English and Standard Chinese tasks. The results are presented in Tables 11, 12, 13, 14.

About English-TruthfulQA, in Table 11, YueTung-7b achieves a Rouge-l score of 37.41% in the zero-shot setting and an impressive 63.50% in the five-shot setting on the English-TruthfulQA benchmark. These scores are competitive with larger models like ChatGPT, which scores 37.81% (zero-shot) and 50.43% (five-shot), and GPT-4, which achieves 19.58% (zero-shot) and 53.18% (five-shot). YueTung-7b's high BERTScore indicates strong semantic similarity to the ground truth, suggesting effective cross-lingual transfer of knowledge.

About English-GSM8K, Table 12 shows that YueTung-7b attains accuracies of 84.32% (zero-shot) and 86.26% (five-shot) on the English-GSM8K benchmark. While it slightly lags behind top-performing models like Qwen-2.5-72b and GPT-4o, which achieve accuracies above 93%, YueTung-7b's performance is notable given its smaller parameter size and focus on Cantonese data. The results indicate that YueTung-7b retains robust mathematical reasoning abilities in English.

About English-ARC Challenge, on the English-ARC Challenge benchmark (Table 13), YueTung-7b achieves accuracies of 89.15% (zero-shot) and 95.25% (five-shot). This performance is competitive with larger models and surpasses several, such as Qwen-2-72b and Llama-3-70b. YueTung-7b's strong results suggest that it can effectively handle English multiple-choice questions, demonstrating cross-lingual generalizability.

About CMMLU, Table 14 presents YueTung-7b's performance on the CMMLU benchmark in Standard Chinese. The model achieves high accuracies across all categories, with overall accuracies of 92.63% (zero-shot) and 94.49% (five-shot). YueTung-7b outperforms all other models, including large-scale models like Qwen-2.5-72b and GPT-

4. This indicates that training on comprehensive Cantonese data enhances the model's capabilities in closely related languages like Standard Chinese.

Analysis YueTung-7b's performance on mainstream language benchmarks reveals that the model not only excels in Cantonese but also demonstrates strong cross-lingual abilities in English and Standard Chinese. The model consistently performs well across different tasks and settings, suggesting that the high-quality Cantonese data contributes to a robust underlying language understanding that generalizes beyond Cantonese.

These findings highlight the potential of leveraging low-resource language data to improve overall model performance. YueTung-7b's ability to compete with or surpass larger models on mainstream benchmarks underscores the effectiveness of our data collection and training approach.

A.3 Related Work

A.3.1 Cantonese Datasets

At the end of the 16th century, Matteo Ricci compiled the first "Modern Bilingual Chinese Dictionary", significantly incorporating Cantonese terms and highlighting Cantonese's role in Sino-Western interactions. By the 19th century, most bilingual dictionaries focused on Cantonese (Xiang et al., 2024). Historically, Hong Kong and its institutions have led Cantonese data initiatives. Wu (Wu, 1994) created a bilingual parallel corpus from the Hong Kong Legislative Council records in both Standard Chinese and English. This effort was complemented by Hun (Hun-tak Lee, 1999), who pioneered a Cantonese-only corpus with one million characters from dialogues involving Hong Kong children, and by Yip (Yip and Matthews, 2007), who developed a bilingual corpus for Cantonese-speaking children. Additionally, a notable Cantonese corpus was derived from Hong Kong television and theatrical productions (Leung and Law, 2001). The University of Hong Kong further contributed by collecting and annotating spontaneous speech from dialogues and broadcasts, focusing on segmentation, part-of-speech tagging, and phonetic transcription (Ping-Wai, 2006). Lee (Lee, 2011) introduced 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, containing 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.4 Cantonese LLMs

Developing Cantonese LLMs presents significant challenges due to the scarcity of linguistic resources and the unique characteristics of the Cantonese language, which necessitate extensive high-quality datasets for effective continuous pre-training. Despite these hurdles, some closed-source Cantonese LLMs with undocumented training processes have demonstrated proficiency in processing Cantonese[38]. Aligning Cantonese LLMs on downstream tasks is generally less resource-intensive than continuous pre-training. Techniques such as prompting, supervised fine-tuning (SFT), and reinforcement learning from human feedback (RLHF) are employed to reduce biases and align model outputs with the specific cultural and contextual nuances of Cantonese usage.

Recent studies (Fu et al., 2024) have highlighted the effectiveness of ChatGPT in Cantonese dialogue and sentiment analysis. An analysis of over 6,000 messages from a Hong Kong-based web counseling service showed that ChatGPT achieved competitive results compared to traditional models. Additionally, the introduction of the CanChat bot aims to enhance counseling services in Hong Kong secondary schools by providing initial support to students facing academic and familial challenges, enabling human counselors to focus on more critical issues. CanChat offers personalized conversations and an alert system for timely interventions, improving students' emotional well-being during and beyond the COVID-19 pandemic (Fung et al., 2023).

A.5 Evaluation Tools

- Rouge-l: from rouge_metric import PyRouge
- Bleu-4: from nltk.translate.bleu_score import sentence_bleu, SmoothingFunction

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

- BERTScore: bert-base-multilingual-cased & roberta-large

A.6 SFT Data Construction

A.6.1 Cons. Prompts

We can directly extract readily available dialogue data from Huggingface and Common Crawl to construct it into the SFT data format.

Wikipedia data can be structured into the format of SFT dialogues based on the following prompt.

Human: 請問<concept>係咩？ **\n** Assistant: <content>。

Translation:

Human: What is <concept>? **\n** Assistant: <content>.

A.6.2 SFT Data Translation Prompts

Our first round of prompt from Chinese to Cantonese:

你係一位專業中文翻譯粵語翻譯員。你任務係準確噉將所提供中文文本翻譯成粵語，同時保留原文意思、語氣同格式。嚴格遵守以下規則：**\n** 1. 只輸出翻譯結果，唔加入任何解釋、步驟或額外內容。**\n** 2. 保持原有段落結構同標點。**\n** 3. 唔重複、遺漏或改動原文任何部分。**\n** 4. 保持數字或公式不變，唔進行任何計算或修改。**\n\n** 中文文本：**\n**

Translation:

You are a professional translator specializing in translating from Chinese into Cantonese. Your task is to accurately translate the provided Chinese text into Cantonese while preserving the original meaning, tone, and format. Strictly adhere to the following rules: **\n** 1. Only output the translation result, without adding any explanations, steps, or additional content. **\n** 2. Maintain the original paragraph structure and punctuation. **\n** 3. Do not repeat, omit, or alter any part of the original text. **\n** 4. Keep numbers or formulas unchanged, without performing any calculations or modifications. **\n\n** Chinese text: **\n**

The second round of prompt from Chinese to Cantonese:

(System prompt) You are a professional translator specialized in translating Chinese into Cantonese. Your task is to refine and provide a more accurate Cantonese translation based on the original Chinese text and the previous translation result. Please strictly follow these guidelines: **\n\n** 1. Only output the corrected Cantonese translation. Do NOT adding any explanations, steps, calculations, inferences, or extra content. **\n** 2. Preserve the original paragraph structure and punctuation. **\n** 3. Do not repeat, omit, or alter any part of the original text. **\n** 4. Keep numbers and formulas unchanged, without performing any calculations or modifications. **\n\n**

(Human few shot 1) Example 1:**\n** Original Chinese Text: 目：’小明每天早上花10分走到校，如果小明家距离校2公里，那么他每分走多少米？’ **\n** Cantonese Translation: ’題目：小明每日朝早花10分鐘時間行去學校，如果小明屋企距離學校2公里，咁佢每分鐘行幾多米？’ **\n** Example 2: **\n** Original Chinese Text: ’目：今天小明自行家到校用了20分，回家用了25分。如果小明在上和回家的路上的速度一，那么他家到校的距离是校到家的距离的百分之几？’ **\n** Cantonese Translation: ’題目：今日小明踩單車由屋企去學校用20分鐘，返屋企用25分鐘。如果小明返學同返屋企路上速度係一樣，咁佢由屋企去學校距離係學校返屋企距離百分之幾？’ **\n** Example 3: **\n** Original Chinese Text: ’目：**\n** 鹿了24苹果，她想平均分她的3只小鹿吃，每只小鹿可以分到几苹果？’ **\n** Cantonese Translation: ’題目：**\n** 鹿媽媽買24個蘋果，佢想平均分俾佢3隻小鹿食，每隻小鹿可以分到幾多個蘋果？’ **\n**

(Human few shot 2) Example 1: **\n** Original Chinese Text: ’是一于速度、路程、的 。我可以通公式：速度＝路程÷ 解。 **\n** 因小明每天早上走2公里，所以他的路程2千米。而他每天早上要花10分 走到校，因此他的10分，即600秒。 **\n** 所以小明每分走的距离 2公里 / 600秒 = 0.0033公里/秒 或 3.3米/秒。 **\n** 答案：小明每分走3.3米。’ **\n** Cantonese Translation: ’呢個係一條關於速度、路程、時間數學題目。我可以通過公式：速度＝路程÷時間 來解決。 **\n** 因為小明每日朝早行2公里，所以佢路程為2公里。而佢每日朝早要花10分鐘時間行去學校，因此佢時間為10分鐘，即係600秒。 **\n** 所以小明每分鐘行距離係 2公里 / 600秒 = 0.0033公里/秒 或 3.3米/秒。 **\n** 答案：小明每分鐘行3.3米。’ **\n** Example 2: **\n** Original Chinese Text: ’假小明家到校的距离x千米，根据速度等于路程除以的公式，可以得出小明的速度：家到校的速度 = x / 20，校到家的速度 = x / 25。因小明在上和回家的路上的速度一，所以有：x / 20 = x / 25，解出 x = 5/4 千米。 **\n** 因此，家到校的距离是校到家的距离的百分之几，可以通求比值得到：x / (5/4)x = 4/5 = 0.8，即小明家到校的距离是校到家的距离的百分之80。’ **\n** Cantonese Translation: ’假設小明屋企去學校距離係x公里，根據速度等於路程除以時間公式，可以得出小明速度為：屋

企去學校速度 = x / 20，學校返屋企速度 = x / 25。因為小明返學同返屋企路上速度係一樣，所以有：x / 20 = x / 25，解出 x = 5/4 公里。 **\n** 因此，屋企去學校距離係學校返屋企距離百分之幾，可以通過求比值得出：x / (5/4)x = 4/5 = 0.8，即小明由屋企去學校距離係學校返屋企距離百分之80。' **\n** Example 3: **\n** Original Chinese Text: '鹿了24苹果，平均分3只小鹿吃，那么每只小鹿可以分到的苹果就是苹果除以小鹿的只。 **\n** 24÷3=8 **\n** 每只小鹿可以分到8苹果。所以，答案是每只小鹿可以分到8苹果。' **\n** Cantonese Translation: '鹿媽媽買24個蘋果，平均分俾3隻小鹿食，咁每隻小鹿可以分到蘋果數就係總蘋果數除以小鹿隻數。 **\n** 24÷3=8 **\n** 每隻小鹿可以分到8個蘋果。所以，答案係每隻小鹿可以分到8個蘋果。' **\n**

Translation:

(Human few shot 1) Example 1: **\n** Original Chinese Text: 'Question: "Xiao Ming spends 10 minutes every morning walking to school. If his home is 2 kilometers away from school, how many meters does he walk per minute?"' **\n** Cantonese Translation: 'Question: Xiao Ming spends 10 minutes every morning walking to school. If his home is 2 kilometers away from the school, how many meters does he walk per minute?' **\n** Example 2: **\n** Original Chinese Text: 'Question: "Today Xiao Ming rode his bicycle from home to school in 20 minutes and it took him 25 minutes to return. If his speed was the same going to and coming from school, what percentage of the distance from home to school is the distance from school to home?"' **\n** Cantonese Translation: 'Question: Today Xiao Ming cycled from home to school in 20 minutes, and it took him 25 minutes to return. If his speed on both the trip to school and the return home was the same, then what percentage of the distance from home to school is the distance from school to home?' **\n** Example 3: **\n** Original Chinese Text: 'Question: **\n** "Deer Mother bought 24 apples, she wants to evenly distribute them to her 3 fawns, how many apples does each fawn get?"' **\n** Cantonese Translation: 'Question: **\n** Deer Mother bought 24 apples, she wants to distribute them evenly among her 3 fawns, how many apples does each fawn get?' **\n**

(Human few shot 2) Example 1: **\n** Original Chinese Text: 'This is a math question about speed, distance, and time. We can solve it using the formula: Speed = Distance ÷ Time. **\n** Since Xiao Ming walks 2 kilometers every morning, his distance is 2 kilometers. Since he spends 10 minutes walking to school each morning, his time is 10 minutes, which is 600 seconds. **\n** Therefore, the distance Xiao Ming walks per minute is 2 kilometers / 600 seconds = 0.0033 kilometers/second or 3.3 meters/second. **\n** Answer: Xiao Ming walks 3.3 meters per minute.' **\n** Cantonese Translation: 'This is a math problem about speed, distance, and time. We can solve it using the formula: Speed = Distance ÷ Time. **\n** Since Xiao Ming walks 2 kilometers every morning, his distance is 2 kilometers. Since he spends 10 minutes walking to school each morning, his time is 10 minutes, or 600 seconds. **\n** Thus, the distance Xiao Ming walks per minute is 2 kilometers / 600 seconds = 0.0033 kilometers/second or 3.3 meters/second. **\n** Answer: Xiao Ming walks 3.3 meters per minute.' **\n** Example 2: **\n** Original Chinese Text: 'Assuming the distance from Xiao Ming's home to school is x kilometers, based on the formula that speed equals distance divided by time, Xiao Ming's speed can be calculated as: speed from home to school = x / 20, speed from school to home = x / 25. Since Xiao Ming's speed to and from school is the same, we have: x / 20 = x / 25, solving x = 5/4 kilometers. **\n** Therefore, the percentage of the distance from home to school that is the distance from school to home can be found by calculating the ratio: x / (5/4)x = 4/5 = 0.80, meaning the distance from home to school is 80% of the distance from school to home.' **\n** Cantonese Translation: 'Assuming the distance from Xiao Ming's home to school is x kilometers, based on the formula that speed equals distance divided by time, we can calculate Xiao Ming's speed as: speed from home to school = x / 20, speed from school to home = x / 25. Since Xiao Ming's speed to and from school is the same, we have: x / 20 = x / 25, solving x = 5/4 kilometers. **\n** Therefore, the percentage of the distance from home to school that is the distance from school to home can be calculated by determining the ratio: x / (5/4)x = 4/5 = 0.80, meaning the distance from home to school is 80% of the distance from school to home.' **\n** Example 3: **\n** Original Chinese Text: 'Deer Mother bought 24 apples and wants to divide them equally among her 3 fawns, so the number of apples each fawn gets is the total number of apples divided by the number of fawns. **\n** 24÷3=8 **\n** Each fawn gets 8 apples. Therefore, the answer is each fawn gets 8 apples.' **\n** Cantonese Translation: 'Deer Mother bought 24 apples and wants to divide them equally among her 3 fawns, so the number of apples each fawn gets is the total number of apples divided by the number of fawns. **\n** 24÷3=8 **\n** Each fawn gets 8 apples. Therefore, the answer is each fawn gets 8 apples.' **\n**

A.6.3 Data Leakage Concerns

We also focus on data leakage, given that pretrained data primarily learns Cantonese language patterns and does not involve test data from the Yue-Benchmark.

Our concern is whether there is data in the SFT that resembles that in the Yue-Benchmark. Due to computational speed limitations, we used the Bleu metric to identify data most similar to the Yue-Benchmark test data and performed Bleu and BERTScore linguistic similarity calculations, ultimately observing a Bleu score of 0.08 and a BERTScore of 0.32.

A.7 Language Recognition Tool

The tool we use was specifically designed and optimized for Cantonese linguistic characteristics, encompassing support for traditional Chinese characters, colloquial forms, and Cantonese-English code-switching, all while demonstrating high robustness and accuracy in real-world applications.

By contrast, OpenLID, which is based on fastText, performs poorly for Cantonese, with official test results reporting F1 and FP scores of only 0.0059 and 0.0025, respectively[39].

In addition, the tool we use integrates a large-scale Cantonese corpus to ensure high compatibility with traditional characters, colloquial usage, and mixed Cantonese-English materials, thus meeting the requirements for data collection and cleaning.

A.8 All Results

A.8.1 Cantonese Benchmarks

A.8.2 Mainstream Language Benchmarks

Models (7-8b scale)	**Acc. (0-shot)**	**Acc. (5-shot)**
Qwen-7b	0.68	6.75
Qwen-1.5-7b	36.62	26.31
Qwen-2-7b	50.49	61.11
Qwen-2.5-7b	63.84	44.20
Llama-2-7b	0.83	1.82
Llama-3-8b	52.46	49.66
Llama-3.1-8b	63.91	61.64
Yi-6b	2.12	10.16
Yi-1.5-6b	3.94	3.49
Internlm-7b	4.55	9.48
Internlm-2-7b-chat	56.41	48.67
Internlm-2-7b	11.37	23.96
Internlm-2.5-7b-chat	65.96	64.67
Internlm-2.5-7b	56.79	42.99
YueTung-7b	**84.65**	**86.46**
Models (> 7-8b scale)	**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
Mistral-8x22b	65.20	66.19
Mistral-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-2-20b	12.81	8.87
Internlm-2-20b-chat	60.42	59.21
Internlm-2.5-20b-chat	71.87	72.33
Internlm-2.5-20b	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
YueTung-7b	**84.65**	**86.46**

Table 8: **All results** of the comparison between answer generated by YueTung-7b and baselines in Yue-GSM8K based on 0-shot and 5-shot settings and ground truth.

[39] (https://github.com/laurieburchell/open-lid-dataset/blob/main/languages.md

Models (7-8b scale)	Acc. (0-shot)	Acc. (5-shot)
Qwen-7b	11.02	14.60
Qwen-1.5-7b	65.24	67.55
Qwen-2-7b	79.08	78.39
Qwen-2.5-7b	81.64	83.35
Llama-2-7b	23.57	34.24
Llama-3-8b	70.11	53.80
Llama-3.1-8b	69.00	67.81
Yi-6b	31.00	66.01
Yi-1.5-6b	34.59	66.70
Internlm-7b	44.75	55.34
Internlm-2-7b	44.75	55.34
Internlm-2.5-7b-chat	81.21	79.85
Internlm-2.5-7b	77.37	77.37
YueTung-7b	**93.48**	**94.65**
Models (> 7-8b scale)	**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
Mistral-8x22b	76.09	76.09
Mistral-large-2	89.50	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
Internlm-2.5-20b	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	88.90	88.73
ChatGPT	69.68	70.71
GPT-4o	91.97	94.45
GPT-4	92.66	92.06
YueTung-7b	**93.48**	**94.65**

Table 9: **All results** of the comparison between answer generated by YueTung-7b and baselines in Yue-ARC-C based on 0-shot and 5-shot settings and ground truth.

Models (7-8b scale)	0-shot					5-shot				
	STEM	**Hum.**	**S.S.**	**C.S.**	**Oth.**	**STEM**	**Hum.**	**S.S.**	**C.S.**	**Oth.**
Qwen-7b	10.10	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	53.00	60.14	70.09	65.55	58.31	65.02
Qwen-2-7b	70.06	81.04	80.07	69.54	76.04	74.08	80.45	80.70	73.70	79.52
Qwen-2.5-7b	72.86	81.66	78.25	66.56	75.19	78.05	80.37	78.99	69.82	78.86
Llama-2-7b	23.34	23.84	23.76	22.78	24.52	27.48	30.40	31.76	28.90	24.38
Llama-3-8b	49.13	59.30	56.51	47.53	53.72	44.04	58.47	53.94	46.24	52.55
Llama-3.1-8b	45.96	58.27	56.08	44.86	53.70	53.45	58.06	58.31	45.86	53.65
Yi-6b	36.46	67.62	57.32	57.42	50.06	58.11	72.14	68.40	60.56	68.46
Yi-1.5-6b	17.34	35.98	38.77	32.90	25.00	58.53	67.89	66.56	60.00	62.05
Internlm-7b	31.90	48.79	44.03	41.14	39.82	39.84	51.74	50.06	43.60	42.32
Internlm-2-7b	51.69	70.92	64.71	59.31	58.93	53.11	68.51	62.68	59.77	58.14
Internlm-2.5-7b-chat	64.40	80.92	76.80	70.24	75.02	65.04	80.84	76.79	70.47	75.19
Internlm-2.5-7b	65.34	82.43	79.24	73.11	74.15	66.73	81.06	77.80	71.65	75.37
YueTung-7b	**93.01**	**92.54**	**89.84**	**90.81**	**91.55**	**93.36**	**93.27**	**91.04**	**91.77**	**91.85**

Models (> 7-8b scale)	0-shot					5-shot				
	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	84.80	89.31
Qwen-2-72b	81.68	89.93	88.47	81.90	87.48	85.70	89.54	88.12	83.72	87.73
Qwen-2.5-72b	83.72	87.88	87.20	80.68	85.36	83.89	89.70	88.75	82.34	87.42
Mistral-8x22b	50.40	57.08	59.28	44.02	48.76	58.94	59.72	62.44	49.78	57.83
Mistral-large-2	60.38	76.08	74.92	60.19	70.74	68.50	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	74.90
Phi-3-medium	45.26	61.42	58.40	45.65	51.33	49.88	59.33	59.35	45.49	53.02
Gemma-2-27b	48.50	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	78.20	80.30
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	72.86	86.10	82.14	79.06	74.70	69.65	78.79	76.56	70.28	77.20
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	82.00	79.88	73.52	74.77
Claude-3.5	66.47	76.84	78.04	60.60	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	84.20	80.07	76.00	78.06
ChatGPT	49.78	58.13	58.74	45.46	52.42	60.28	59.81	60.61	47.50	54.54
GPT-4o	74.16	83.28	84.12	71.60	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	63.50	74.57
YueTung-7b	**93.01**	**92.54**	**89.84**	**90.81**	**91.55**	**93.36**	**93.27**	**91.04**	**91.77**	**91.85**

Table 10: **All results** of the comparison between texts generated by YueTung-7b and baselines in Yue-MMLU based on 0-shot and 5-shot settings and the correct texts.

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
Mistral-8x22b	18.59	12.91	85.78	31.05	20.61	87.58
Mistral-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
Qwen-2.5-7b	9.46	11.70	82.95	15.47	10.93	84.33
YueTung-7b	37.41	22.54	89.15	63.50	36.13	93.14

Table 11: 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
Mistral-8x22b	91.51	91.58
Mistral-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	95.00	94.77
Qwen-2.5-7b	88.62	88.65
YueTung-7b	84.32	86.26

Table 12: 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
Mistral-8x22b	90.82	88.07
Mistral-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
Qwen-2.5-7b	83.95	78.20
YueTung-7b	89.15	95.25

Table 13: 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
Mistral-8x22b	57.88	63.27	64.51	49.18	57.28	62.38	62.97	63.7	51.52	58.26
Mistral-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
Qwen-2.5-7b	71.29	82.50	78.43	69.57	74.04	75.77	84.26	81.64	71.72	78.54
YueTung-7b	94.08	92.81	90.68	91.92	92.63	94.48	93.39	91.98	93.57	94.49

Table 14: Results of the comparison between texts generated by various LLMs in CMMLU based on 0-shot and 5-shot settings and the correct texts.