HKCanto-Eval: A Benchmark for Evaluating Cantonese Language Understanding and Cultural Comprehension in LLMs

Tsz Chung Cheng[1], Chung Shing Cheng[2], Chaak Ming Lau[3], Eugene Tin-Ho Lam[4], Chun Yat Wong[5], Hoi On Yu[6], Cheuk Hei Chong[7,8]

[1] Kyushu University, [2] hon9kon9ize, [3] The Education University of Hong Kong, [4] Independent Researcher, [5] The University of Hong Kong, [6] Independent Researcher, [7] Votee AI, [8] Beever AI

Correspondence: Tsz Chung Cheng: jed.cheng@mag.ed.kyushu-u.ac.jp, Chung Shing Cheng joseph.cheng@hon9kon9ize.com Chaak Ming Lau: lchaakming@eduhk.hk

Abstract

The ability of language models to comprehend and interact in diverse linguistic and cultural landscapes is crucial. The Cantonese language used in Hong Kong presents unique challenges for natural language processing due to its rich cultural nuances and lack of dedicated evaluation datasets. The HKCanto-Eval benchmark addresses this gap by evaluating the performance of large language models (LLMs) on Cantonese language understanding tasks, extending to English and Written Chinese for cross-lingual evaluation. HKCanto-Eval integrates cultural and linguistic nuances intrinsic to Hong Kong, providing a robust framework for assessing language models in realistic scenarios. Additionally, the benchmark includes questions designed to tap into the underlying linguistic metaknowledge of the models. Our findings indicate that while proprietary models generally outperform open-weight models, significant limitations remain in handling Cantonese-specific linguistic and cultural knowledge, highlighting the need for more targeted training data and evaluation methods. The code can be accessed at https://github.com/hon9kon9ize/hkeval2025

1 Introduction

Recent advancements in large language models (LLMs) such as GPT-4, Gemini, and various open-weight models have demonstrated remarkable capabilities in natural language understanding across multiple languages (Xu et al., 2024). However, the performance of most models significantly declines when applied to languages other than English, yielding particularly poor outcomes for low-resource languages (LRLs). These languages are under-represented lingua francas that play a crucial role in certain communities, and it is imperative to improve multilingual support for LRLs by creating benchmarks to guide the future development of multilingual LLMs. Since they are poorly supported due to the lack of training data, if there is a close language with more resources, this problem can potentially be mitigated through few-shot learning. A notable example of this strategy is the use of Bahasa Indonesian to handle regional languages in Indonesia (Aji et al., 2022; Winata et al., 2022). This strategy aligns with the spirit of language sustainability and AI support for marginalised communities (Du et al., 2020), which is also applicable to Cantonese.

This paper investigates the status of LLM support for Cantonese (ISO 639-3: *yue*), a member of the *Sinitic* ("Chinese") branch of the Sino-Tibetan language family, and a distinct variety unintelligible to users of Mandarin, the standard variety of Chinese used in Mainland China (Pǔtōnghuà) and Taiwan (Guóyǔ). Cantonese, spoken by over 85 million people according to *Ethnologue* (Eberhard et al., 2024), serves as the most common and *de facto* official language of Hong Kong and Macau, and is also widely used in parts of Guangdong, Guangxi, Malaysia, and Singapore. Additionally, it is used as a diasporic language in countries such as Canada (Sachdevl et al., 1987), the United States (Leung and Uchikoshi, 2012), Australia (Zhang et al., 2023), and the United Kingdom (Bauer, 2016; Tsapali and Wong, 2023). Despite its widespread use, Cantonese is still considered a low-resource language (Xiang et al., 2024) due to the lack of quality written resources. This scarcity results from a "diglossia" that requires Written Chinese (which resembles Mandarin) to be used in formal settings[1], and a longstanding, ideologically-driven stigmatisation of Cantonese as an informal/vulgar language (Lau, 2024), further confines written Cantonese to informal con-

[1] Even in Mandarin-like Written Chinese, there are persistent lexical differences with other regions due to vastly different governmental, legal and education systems. For instance, the word "taxi" is rendered as "出租車" in mainland China, "計程車" in Taiwan, and "的士" in Hong Kong and Macau.

Proceedings of the 29th Conference on Computational Natural Language Learning, pages 1–11

texts like social media and texting.

Cantonese is partially supported by certain LLMs, with models like GPT-4 and Gemini capable of comprehending and responding in Cantonese (Fu et al., 2024; Hong et al., 2024; Jiang et al., 2024). There are models dedicated to better supporting Chinese languages and dialects: The Hong Kong government is developing an internal tool based on locally developed LLMs for administrative use (Yiu, 2024); SenseTime released SenseChat (Cantonese), a model trained on 6 billion tokens of Hong Kong-specific data (SenseTime, 2024). However, the current support level is mostly contributed to by small pockets of Cantonese presented in the sheer volume of Written Chinese training data. There have been comparisons between Chinese and Western models on how well languages spoken in China are handled (Wen-Yi et al., 2025), showing that Chinese models outperformed Western ones on Mandarin, but the same cannot be said for Cantonese or other languages in China. The following section outlines how current benchmarking studies have yet to provide a comprehensive evaluation for Cantonese and Hong Kong-related tasks that tap into the in-depth representation of underlying aspects of the language, which we believe is the prerequisite for accurate comprehension in uncommon scenarios.

2 Related Benchmarks

The development of LLMs has spurred significant research into evaluating their performance and comparing their capabilities to human reasoning across general and domain-specific tasks. A prominent benchmark in this area is the MMLU dataset (Hendrycks et al., 2020), which comprises 57 tasks ranging from elementary to university-level multiple-choice questions. Despite its widespread use, MMLU has been criticised for containing flawed questions and answers (Gema et al., 2024; Gupta et al., 2024). To address these shortcomings, alternative benchmarks such as BIG-Bench (Srivastava et al., 2022), MMLU-Pro (Taghanaki et al., 2024), and MMLU-Pro+ (Wang et al., 2024) have been introduced, aiming to improve accuracy while presenting more diverse and challenging questions.

In addition to comprehensive benchmarks, researchers have developed domain-specific, expert-curated datasets to evaluate the reasoning capabilities of LLMs in specialised fields such as programming (HumanEval (Chen et al., 2021); NL2Code (Zan et al., 2022)) and mathematical reasoning (GSM8K (Cobbe et al., 2021); MATH (Hendrycks et al., 2021); MATH 401 (Yuan et al., 2023); Omni-MATH (Gao et al., 2024)).

Although most existing LLM benchmarks focus on English-language tasks, culturally-aware datasets integrating machine-translated questions, native datasets, and exam questions have been developed in other languages, including Arabic (Koto et al., 2024), Basque (Etxaniz et al., 2024a,b), Spanish (Plaza et al., 2024), Indic languages (Verma et al., 2024), and Korean (Son et al., 2024). Similar benchmarks have been published for Chinese, such as CMMLU (Li et al., 2023) and C-Eval (Huang et al., 2024) that gathered questions from various academic and professional exams in mainland China, and TMLU (Chen et al., 2024) and TMMLU+ (Tam et al., 2024) that evaluate knowledge in Traditional Chinese in the context of Taiwan.

These benchmarks are not applicable to the Hong Kong context due to the aforementioned diglossia and regional lexical differences. Recently, Jiang et al. (2024) introduced a Cantonese evaluation benchmark that combines four datasets translated from other languages (ARC, GSM8K, CMMLU, and Truthful-QA)[2], resulting in a dataset that is heavily biased towards American culture (16.9% entries in the Truthful-QA dataset reference the United States) or mainland Chinese exams (CMMLU) (see Appendix A).

3 Methodology

HKCanto-Eval introduces a specialised benchmark to address the lack of systematic tests for evaluating the Cantonese capabilities and Hong Kong knowledge of an LLM in these aspects: (1) **Language Proficiency**, the capability in an accurate and nuanced understanding of Cantonese and local-flavoured Written Chinese, as well as generating fluent, idiomatic, genre-appropriate Cantonese text in question and answering, translation, and summarisation tasks; (2) **Cultural Knowledge**, in-depth knowledge about not only general historical and geographical facts related to Hong Kong, but also everyday practices, local customs, beliefs and values, and cultural references from

[2]It also contains a translation evaluation component for English-Cantonese and Simplified-to-Traditional Chinese translations, but its data sources and evaluation methods are not fully transparent.

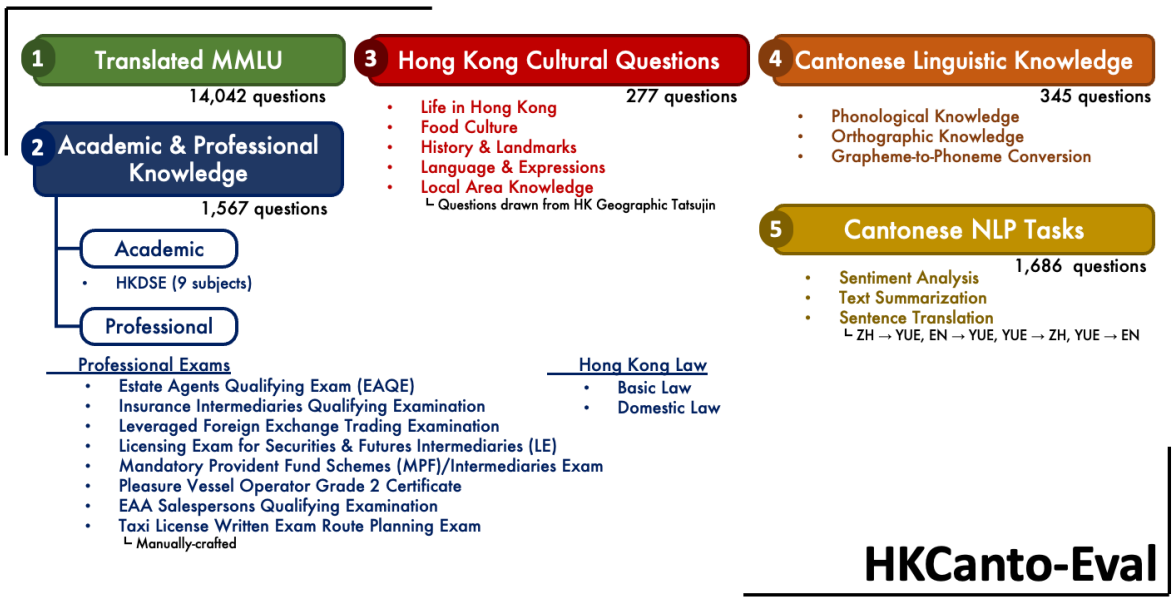

Figure 1: Diagram showing the tasks of the HKCanto-Eval Benchmark

movies, music, literature, and internet culture; (3) **Reasoning and Problem-Solving**, reasoning and problem-solving skills within a Cantonese and/or Hong Kong-based context, including reasoning about the sound and written forms of the language.

These aspects are incorporated into the five datasets outlined below.

3.1 Translated MMLU Dataset

The first dataset comprises 14,042 questions from the original MMLU dataset in English (Hendrycks et al., 2020) and their Cantonese translation[3]. This allows us to compare how LLMs perform when handling knowledge in a wide range of subjects in Cantonese rather than in English (See Appendix B).

3.2 Academic and Professional Dataset

The Academic and Professional Dataset is a set of multiple-choice questions curated to measure LLMs' reasoning and problem-solving abilities in domain-specific knowledge. The dataset contains multiple-choice questions from 3 sub-categories: (1) **Academic**: Questions sourced from Hong Kong Diploma of Secondary Education (HKDSE), a territory-wide high-school graduate-level exam; extracted and manually corrected from scanned PDFs and are believed to have never appeared online in a plain-text form; (2) **Professional**: Questions from seven professional qualification exams, extracted from text PDF files found on the corresponding official sites (in which the model answers were not on the same page as the questions, avoiding data contamination concerns), and an additional set of Taxi Licensing Exam Styled Route Planning questions on Hong Kong roads and geographical features; (3) **Law**: Questions about law in Hong Kong across 15 categories sourced from the Internet, and an additional subset of the Basic Law edited by the authors.

All questions are in Written Chinese (in the Traditional script). We also included an English version if it is available. The details of this dataset can be found in Appendix C.

3.3 Hong Kong Cultural Questions Dataset

This dataset contains 277 manually crafted questions divided into five categories that capture cultural knowledge common to people who have lived or grown up in Hong Kong, that are often not learned in schools. The categories are **Food Culture**, **History and Landmarks**, **Language and Expressions**, **Life in Hong Kong** and **Local Area Knowledge**. The questions were collected in a way to capture knowledge from all walks of life. 244 questions were developed by the authors and volunteers for the first four categories, and the last category comes from an online quiz. Questions were created so that they were non-trivial and at

[3]The translation was done by the Google Gemini 1.5 Flash API, which offers a balance between top performance and cost as one would find in the later section. To address concerns regarding the accuracy of LLM translation, we have selected 4 questions from each category for human checking. 202 out of 228 sentences were judged to be good by the raters.

the same time not too obscure, and have been verified by all the authors. Details can be found in Appendix D.

3.4 Linguistic Knowledge Dataset

This is an assessment of the linguistic knowledge represented in the models, inspired by the approach of PhonologyBench (Suvarna et al., 2024) for English. To our knowledge, this innovative approach has never been incorporated into existing Cantonese or Chinese benchmarks in general.

3.4.1 Phonological Knowledge

The dataset contains 100 questions that evaluate phonological knowledge about characters and words of an LLM, including the judgment of homophones and rhyming and other non-trivial reasoning tasks based on word pronunciation. These are particularly important in the Cantonese context, as the writing system does not provide reliable cues about the pronunciation of words, and Cantonese materials are not accompanied by sound transcription. This knowledge needs to be present in the training data for tasks that require sound-related operations or reasoning (See Appendix E.1).

3.4.2 Orthographic Knowledge

The Orthographic Knowledge Dataset evaluates the character meta-knowledge of an LLM. Cantonese users from Hong Kong need to know around 4,000 characters by the age of 12 and will have built sound knowledge about the representation of the characters. This subset contains 100 questions about the strokes, structure, arrangement, and radical and constituent components of common characters. Cantonese uses the Traditional Chinese script (ISO 15924: Hant) in Hong Kong and Macau, and the script is also used in Taiwan. There could be influence from Mandarin data or Taiwan usage not shared by Cantonese. It is also expected that certain models may produce incorrect answers due to the over-reliance on simplified Chinese data (See Appendix E.2).

3.4.3 Grapheme-to-Phoneme (G2P) Conversion

This dataset addresses the task of converting a string of written text represented in Traditional Chinese characters into Jyutping, a widely adopted romanisation standard of Cantonese[4]. This is similar to typical G2P tasks except that Jyutping is used instead of the International Phonetic Alphabet (IPA) as the output. G2P functionalities have been implemented by PyCantonese (Lee et al., 2022), a Cantonese NLP package, Hambaanglaang Converter[5] and Visual Fonts[6]. As the task is non-deterministic, rule-based conversions are bound to be unreliable (although Visual Fonts have achieved very high accuracy now). There is also no reliable non-rule-based G2P system to our best knowledge. This part of the dataset contains 150 pairs of Character-Jyutping sentences from both Standard Written Chinese and Cantonese and in a range of formality levels, manually checked by professional linguists from the Linguistic Society of Hong Kong, the organisation that established and maintains the Jyutping system. The score calculation method is discussed in Appendix E.3.

3.5 NLP Tasks Dataset

Multiple-choice questions offer a structured approach to assess LLM factual knowledge and reasoning, but they are insufficient for evaluating real-world language understanding and generation. Open-ended tasks, including translation and summarisation, were incorporated.

A translation dataset comprising 20 Cantonese sentences with complex linguistic nuances was created, with each sentence manually translated into English and written Chinese (resulting in 4 translation pairs per sentence) (See Appendix F). For summarisation, 10 Cantonese articles and 10 TED talk subtitles were used. The importance of transcription-based summarisation, reflecting Cantonese's prevalence in oral communication, is emphasised by the inclusion of TED talks (See Appendix G).

Performance on traditional NLP tasks like sentiment analysis was also evaluated. Leveraging the OpenRice dataset (toastynews, 2020) (restaurant reviews categorised as positive, neutral, or negative), 1200 reviews (avg. 309 characters) with a balanced sentiment distribution were included. Additionally, a new dataset of 399 Facebook comments (avg. 24 characters), labelled by paid interns, was created (See Appendix H).

3.6 Evaluation Method

The evaluation process of multiple-choice questions follows the standard 5-shot evaluation procedures in MMLU formulation. However, for the

[4] https://lshk.org/jyutping-scheme

[5] https://test.hambaanglaang.hk

[6] https://visual-fonts.com

Hong Kong Cultural Questions Dataset, a zero-shot evaluation was also conducted to emulate actual usage. The translated MMLU dataset used the same system prompt as the original MMLU dataset. For other multiple-choice questions, a short sentence with the name of the exam or question subcategory is added.

For the G2P dataset, character error rates (CER) and Levenshtein distance were both used to calculate the discrepancy between the model output and the gold standard in a five-shot evaluation. The summarisation tasks were evaluated without any example to avoid exceeding the context length of any model, while zero and three-shot evaluations were carried out for the translation task.

The outputs of both translation and summarisation evaluation were evaluated and graded by paid undergraduate students and teaching assistants. The rubric can be found in Appendix F and G. As technology improves, future LLMs can perform the task to offer scalability. Nonetheless, the results from this human evaluation will be useful for verifying the validity and consistency of LLM-as-a-judge in the future.

3.7 Model Selection

13 model families were selected for evaluation. Proprietary models including OpenAI GPT4o (Hurst et al., 2024) and GPT4-mini (OpenAI, 2024), Google Gemini 1.5 Flash and Gemini 1.5 Pro (Gemini Team et al., 2024) and Anthropic Claude 3.5 Sonnet (Anthropic, 2024) were selected for their reported superior performance across different languages.

Three proprietary models from Chinese companies, including Doubao Pro from ByteDance (Doubao, 2024), Erne 4.0 from Baidu (Baidu Inc., 2023) and SenseChat (Cantonese) from SenseTime (SenseTime, 2024), were also incorporated. All proprietary models were accessed through their API, except SenseChat, which was accessed via the web interface due to a failure to get verified to use their API.

Popular multilingual open-weight models including Aya 23 8B (Aryabumi et al., 2024), Gemma 2 2B, 9B and 27B (Gemma Team et al., 2024), Llama 3.1 8B, 70B and 405B (Dubey et al., 2024), and Mistral Nemo Instruct 2407 12B (Mistral, 2024) were included to assess their cross-lingual ability. The collection also included two open-weight multilingual models from Chinese companies, Yi 1.5 6B, 9B and 34B (Young et al., 2024) and Qwen2 7B and 72B (Yang et al., 2024). In addition, CantoneseLLM (CLLM) v0.5 6B and 34B[7] are two of the few open-weight models trained specifically on Cantonese data. They were trained by fine-tuning Yi 1.5 6B and 34B models with around 400 million tokens of Hong Kong-related content. Open-weight instructions fine-tuned models smaller than 70B parameters were evaluated using Nvidia H100 GPUs. The 70B and 405B models were evaluated using the API of SiliconFlow[8].

4 Results

4.1 MMLU

Table 1 shows the results of the multiple-choice questions. Proprietary models and open-weight models like Llama 3.1 70B, 405B, and Qwen 2 72B performed well in MMLU, but experienced an average of 7.46 percentage point drop when questions were in Cantonese. Considering potential errors from machine translations, this is evidence of *Cantonese reasoning and problem-solving ability*.

4.2 Academic and Professional Questions

The results of this dataset showed expected problem-solving abilities across models in different subject areas, in particular, general weaknesses in handling secondary school-level mathematics and strong performance in legal questions. Proprietary models generally performed better than open-weight models. The sub-scores in the individual tasks show that most models struggled with academic questions that were never posted online. It is worth noting that some open-weight models (e.g. CLLM v0.5 34B and Qwen2 72B) outperformed most models, and we can conduct further investigation on what additional training data was used to achieve this performance. Written Chinese yielded better overall results, and this is attributed to the Law dataset, which only came in Chinese. Discounting this set, Written Chinese caused a slight drop in performance. This indicates that *multi-lingual open-weight LLMs showed cross-lingual capabilities*, maintaining similar performance across both languages.

[7] https://huggingface.co/hon9kon9ize/CantoneseLLMChat-v0.5

[8] https://siliconflow.cn

Model	MMLU		Academic & Professional		Cultural		Average	
	EN	YUE	EN	ZH	0-shot	5-shot	EN	ZH/YUE
Claude 3.5 Sonnet	**85.0%**	**81.5%**	75.1%	75.2%	71.7%	75.0%	80.1%	75.8%
Doubao Pro	79.8%	74.2%	60.8%	70.5%	70.7%	75.0%	70.3%	72.6%
Ernie 4.0	81.0%	75.2%	70.4%	72.4%	68.2%	75.2%	75.7%	72.8%
Gemini 1.5 Flash	79.0%	73.1%	67.4%	68.3%	61.0%	64.0%	73.2%	66.6%
Gemini 1.5 Pro	83.2%	77.6%	71.0%	71.7%	74.0%	73.8%	77.1%	74.3%
GPT4o	84.8%	80.3%	**77.6%**	75.3%	**77.5%**	77.2%	**81.2%**	**77.6%**
GPT4o-mini	76.7%	69.4%	62.0%	65.6%	55.6%	60.6%	69.4%	62.8%
SenseChat	78.7%	70.1%	73.6%	**75.6%**	67.4%	**77.4%**	76.1%	68.8%
Aya 23 8B	56.6%	47.1%	44.8%	49.0%	39.5%	37.7%	50.7%	43.3%
CLLM v0.5 6B	58.6%	51.7%	50.9%	53.5%	52.0%	56.1%	54.7%	53.3%
CLLM v0.5 34B	75.9%	69.9%	66.8%	69.9%	72.5%	76.7%	71.3%	72.3%
Yi 1.5 6B	64.1%	54.0%	53.7%	58.3%	47.7%	50.7%	58.9%	52.7%
Yi 1.5 9B	70.9%	60.8%	59.2%	63.3%	48.7%	57.3%	65.0%	57.5%
Yi 1.5 34B	76.1%	68.5%	63.7%	68.7%	67.7%	72.9%	69.9%	69.5%
Gemma 2 2B	58.5%	46.5%	45.3%	48.5%	33.3%	35.2%	51.9%	40.9%
Gemma 2 9B	73.4%	64.3%	63.6%	64.0%	49.1%	51.6%	68.5%	57.3%
Gemma 2 27B	76.4%	68.4%	65.1%	68.1%	57.1%	60.9%	70.7%	63.6%
Llama 3.1 8B	69.0%	56.4%	51.4%	57.1%	45.6%	52.7%	60.2%	52.9%
Llama 3.1 70B	80.3%	74.9%	68.2%	70.0%	63.0%	64.4%	74.2%	68.1%
Llama 3.1 405B	**84.5%**	**78.4%**	70.9%	74.2%	67.9%	69.9%	77.7%	72.6%
Mistral Nemo 12B	68.8%	58.4%	54.6%	58.0%	40.1%	42.7%	61.7%	49.8%
Qwen2 7B	71.2%	64.8%	60.7%	65.4%	53.6%	54.8%	66.0%	59.6%
Qwen2 72B	82.9%	78.3%	**74.7%**	**76.3%**	**72.9%**	**77.7%**	**78.8%**	**76.3%**
Random	25.0%	25.5%	22.9%	24.6%	29.8%	28.1%	23.9%	27.0%

Table 1: Model performance on MMLU, Academic and Professional, and Cultural questions. Note that SenseChat refused to answer one subset of questions in Cultural Question 5-shot evaluation.

4.3 Hong Kong Cultural Questions

Proprietary models and Qwen 2 72B showed a good understanding of Hong Kong cultural knowledge, yet none of the models performed well across the subcategories. Looking into the sub-scores, models occasionally matched humans in most subtests (e.g. Food Culture and Life in HK). However, when inspecting the results, good performance by percentage *only reflects the size of existing Hong Kong knowledge represented in Wikipedia entries.* For example, only two models (Yi 1.5 6B and Qwen2 72B) correctly answered the origin of Demae Itcho noodles sold in Hong Kong, while 94% of humans did. The results for Language & Expressions also show that *most models did not have a nuanced understanding of Cantonese*. Compared to human performance at 85.8%, SenseChat scored the highest point out of all models in 5-shot (79.6%), but its performance dropped significantly in zero-shot (61.4%). In zero-shot evaluation, CLLM v0.5 34B delivered the best performance at 77.3%. Furthermore, model size affects the performance of geospatial tasks, with open-source models in the 6-9B parameter range achieving only about 50% of larger models' performance on Local Area Knowledge (e.g. Yi 1.5 34B 67.9%, 9B 35.7%). The overall results of this dataset suggest that Hong Kong cultural knowledge is underrepresented in LLM training. See Appendix C for details.

4.4 Linguistic and NLP Tasks

These two groups of tasks reveal the representation of Cantonese phonological, orthographic, lexical and grammatical knowledge in existing models. The overall results (Table 2) show a consistent trend where proprietary models outperformed open-weight models (but more pronounced in linguistic tasks). GPT-4o led with 76.7% and 89.6% in both *linguistic* and *NLP* tasks. Lower scores are often due to chance-level performance when

Model	Phonological Knowledge			Orthographic Knowledge			NLP
	Homo-phone	Rhyme	Misc.	Visual Sim.	Canton. Char.	Misc.	Avg.
Claude 3.5 Sonnet	28.0%	64.0%	16.0%	50.0%	76.9%	59.3%	89.2%
Doubao Pro	16.0%	44.0%	16.0%	70.0%	80.8%	48.1%	87.0%
Ernie 4.0	28.0%	60.0%	18.0%	70.0%	80.8%	53.7%	82.7%
Gemini 1.5 Flash	12.0%	20.0%	24.0%	40.0%	73.1%	31.5%	83.2%
Gemini 1.5 Pro	16.0%	40.0%	24.0%	50.0%	**88.5%**	46.3%	87.9%
GPT4o	**56.0%**	**96.0%**	**28.0%**	50.0%	65.4%	**63.0%**	**89.6%**
GPT4o-mini	20.0%	60.0%	20.0%	30.0%	57.7%	40.7%	86.1%
SenseChat	16.0%	36.0%	22.0%	**75.0%**	76.9%	42.6%	78.8%
Aya 23 8B	12.0%	40.0%	14.0%	15.0%	19.2%	31.5%	70.1%
CLLM v0.5 6B	24.0%	8.0%	18.0%	20.0%	50.0%	27.8%	71.9%
CLLM v0.5 34B	28.0%	28.0%	14.0%	35.0%	76.9%	37.0%	73.3%
Yi 1.5 6B	28.0%	12.0%	12.0%	10.0%	50.0%	20.4%	56.6%
Yi 1.5 9B	**36.0%**	40.0%	24.0%	30.0%	57.7%	18.5%	72.2%
Yi 1.5 34B	16.0%	32.0%	**26.0%**	30.0%	61.5%	33.3%	82.9%
Gemma 2 2B	8.0%	24.0%	18.0%	25.0%	53.8%	22.2%	73.4%
Gemma 2 9B	20.0%	28.0%	24.0%	25.0%	50.0%	33.3%	85.0%
Gemma 2 27B	20.0%	12.0%	16.0%	25.0%	65.4%	24.1%	83.2%
Llama 3.1 8B	12.0%	16.0%	18.0%	25.0%	42.3%	38.9%	60.3%
Llama 3.1 70B	28.0%	40.0%	12.0%	30.0%	61.5%	35.2%	**84.5%**
Llama 3.1 405B	20.0%	**44.0%**	18.0%	35.0%	65.4%	**50.0%**	64.4%
Mistral Nemo 12B	12.0%	28.0%	10.0%	25.0%	23.1%	37.0%	68.8%
Qwen2 7B	8.0%	40.0%	12.0%	35.0%	46.2%	33.3%	66.8%
Qwen2 72B	12.0%	28.0%	16.0%	**50.0%**	**76.9%**	48.1%	83.5%
Random/Control	16.0%	28.0%	24.0%	30.0%	11.5%	27.8%	76.8%

Table 2: Model performance on Linguistic Knowledge Dataset multiple-choice questions and NLP tasks. The bottom row indicates the expected correctness from random selection for the Phonological and Orthographic Knowledge tasks. For NLP, the reported figure is the average evaluation of professionally prepared translations for translation tasks serving as a control.

knowledge is absent, or below chance-level due to influence from Mandarin. Here are the key findings and observations:

Most LLMs understand Cantonese fine. Most models performed well in Sentiment Analysis (GPT4o 79.7%, Llama 3.1 405B 78.8%), Translation (3-shot: GPT4o 98.3%, Qwen2 72B 96.6%), and Summarisation (Claude 3.5 Sonnet 92.7%, Gemma 2 9B 91.3%). Models that obtained lower scores are often due to task completion problems, e.g. failure to handle long input and problems with low-frequency/mixed-language tokens.

Proprietary and large open-weight models have good Cantonese lexical knowledge. The performance in translation and sentiment analysis is closely tied to the ability to determine the meaning of Cantonese-specific words that are not found or used differently in Mandarin. Most models also performed well in the Cantoense Character Selection sub-task (Canton. Char. in Table 2) under Orthographic Knowledge. It is noteworthy that despite good performance with proprietary models (73.1% - 88.5%) and some open-weight models (CLLM v0.5 34B and Qwen2 72B, both 76.9%), GPT4o struggled with Cantonese orthography (65.4%).

LLMs in general lack knowledge about Cantonese pronunciation. In the Grapheme-to-Phoneme (G2P) conversion task, all models performed far worse than the rule-based control (Visual Fonts v3.3, CER 0.8%), with the closest being GPT-4o (5.4%) and Claude 3.5 Sonnet (7.9%) as shown in Table 3. The appalling results from all tested language models reveal how linguistic knowledge is seriously under-represented. While it is expected that the G2P tasks will be significantly improved in newer/future models, actual linguistic tasks that involve sounds require more ad-

Model	CER	Levenshtein
Claude 3.5 Sonnet	7.9%	0.018
Doubao Pro	20.9%	0.044
Ernie 4.0	34.4%	0.094
Gemini 1.5 Flash	34.7%	0.083
Gemini 1.5 Pro	15.3%	0.030
GPT4o	**5.4%**	**0.009**
GPT4o-mini	12.0%	0.023
SenseChat	54.4%	0.163
Aya 23 8B	96.6%	0.724
CLLM v0.5 6B	94.1%	0.859
CLLM v0.5 34B	**23.4%**	**0.058**
Yi 1.5 6B	99.0%	0.577
Yi 1.5 9B	97.2%	0.528
Yi 1.5 34B	79.6%	0.837
Gemma 2 2B	97.5%	0.524
Gemma 2 9B	73.0%	0.259
Gemma 2 27B	62.5%	0.201
Llama 3.1 8B	69.9%	0.270
Llama 3.1 70B	31.3%	0.086
Llama 3.1 405B	26.3%	0.074
Mistral Nemo 12B	59.8%	0.201
Qwen2 7B	97.3%	0.466
Qwen2 72B	74.0%	0.268
Rule Based	0.8%	0.001

Table 3: Model performance in the Grapheme-to-Phoneme (G2P) dataset. Scores calculated based on character error rates (CER) and Levenshtein distance. (Lower is better)

vanced knowledge about the language's sound system. Most models struggled with tasks like judging homophone or rhyme pairs in Table 2, with GPT-4o being a notable exception (Homophone: 56.0%; Rhyming: 96.0%). Poor (close to chance level) performance in other models is not only due to the lack of G2P ability, a prerequisite for phonological reasoning, but also due to how Mandarin homophones partially influence this task. This will continue to be challenging for Cantonese due to limited specialised data.

LLMs in general do not have meta-linguistic knowledge represented in Cantonese. Although certain models, especially the Chinese proprietary models, performed well in the visual similarity task (SenseChat 70%, Doubao 70%, Ernie 75%) or orthographic reasoning (GPT4o 63.0%), the knowledge seems to have come from Simplified Chinese, thus their good performance is not transferred to Cantonese-specific items. This seems to be caused by insufficient descriptive knowledge about the structure and properties associated with the individual glyphs.

5 Conclusion

This paper presents HKCanto-Eval, the first comprehensive evaluation benchmark focusing on Hong Kong Cantonese, by comparing the Cantonese language support of 6 proprietary and 7 open-weight model families. Our findings indicate that while these models can understand Cantonese in various contexts, retrieve knowledge about Hong Kong, and address problems written in or about Cantonese to some extent, there are notable limitations. Most models, especially open-weight models in the 6–9B range, lack sufficient linguistic, cultural and professional knowledge in Cantonese and Hong Kong. Performance was particularly poor for questions requiring knowledge not commonly found in major online sources.

One area that we paid close attention to is the presence of metalinguistic knowledge in these models. There is concern that models showed Cantonese proficiency in linguistic and NLP tasks primarily through Mandarin. If their linguistic understanding is based solely on Mandarin, they may perform well on simpler tasks but struggle significantly with "false friends" between languages, as Mandarin knowledge becomes a hindrance. This benchmark introduces a novel perspective, focusing on Cantonese processing abilities beyond superficial slang and expressions. By requiring reasoning about sounds and characters specific to Cantonese, our benchmark provides a fairer judgement that credits models accurately capturing Cantonese phonology and orthography, while exposing those that appear competent in Cantonese but are heavily reliant on Mandarin.

This challenge in processing Cantonese is shared by other low-resource languages. As training data increases, models tend to favour high-resource languages like Mandarin Chinese. The apparent similarity between Cantonese and Written Chinese further affects the ability of even proprietary models to distinguish between these linguistic contexts accurately. Addressing the segregation of regional and linguistic knowledge is crucial for developing culturally and linguistically adaptive LLMs. This issue extends beyond Cantonese to other under-represented language communities.

6 Limitations & Future Directions

The current benchmark exhibits several limitations.

Inaccuracies in machine-translated materials: First, the use of machine translation introduces potential inaccuracies. While Gemini 1.5 Flash balances cost and quality, human-translated questions could provide a more reliable benchmark, albeit at a higher resource cost. The reliance on multiple-choice and text-based questions does not fully capture the capabilities required for practical LLM applications such as code generation and mathematical problem-solving, which demand coherent and contextual text generation. The dataset also lacks multi-modal data like image and audio, which is now supported by proprietary models and should be evaluated.

Biases in topic selection: The newly and manually created questions might contain biases and a lack of scalability and comprehensiveness. The cultural questions, predominantly created by colleagues and relatives of the authors, may introduce bias in cultural references and wordings, leading to an over-representation of certain perspectives while under-representing others, such as traditional practices. Political topics were also specifically excluded, due to political complications, limiting cultural representation. This can also be considered a reasonable compromise since many models (e.g. those from Chinese companies) are configured to censor these topics, and there is a risk that our accounts or IP addresses will be banned before we complete all the benchmarking tasks for this paper.

Lack of Crosslingual Evaluation: English translations for cross-lingual ability evaluation were also not included due to resource limitations. An additional comparison should be added to compare whether the same set of questions will be answered less satisfactorily when presented in English or Standard Written Chinese instead of Cantonese, in line with the evaluation done for Basque (Etxaniz et al., 2024a) and Mongolian and Tibetan (Zhang et al., 2025). We will leave this for future research.

Reliance on human evaluation: Human evaluation, while insightful, is not scalable. Automated and objective evaluation methods, such as LLM-as-a-judge or rule-based approaches, are necessary for efficient evaluation, but this is challenging due to the low-resource nature of Cantonese.

Future directions include developing benchmarks incorporating audio, images, and tables, and addressing the aforementioned limitations to create more comprehensive and representative evaluations.

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.

Anthropic. 2024. Claude 3 model card.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Baidu Inc. 2023. Baidu launches ernie 4.0 foundation model, leading a new wave of ai-native applications.

Robert S. Bauer. 2016. The hong kong cantonese language: Current features and future prospects. *Global Chinese*, 2(2):115–161.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Po-Heng Chen, Sijia Cheng, Wei-Lin Chen, Yen-Ting Lin, and Yun-Nung Chen. 2024. Measuring taiwanese mandarin language understanding. *arXiv preprint arXiv:2403.20180*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

ByteDance Doubao. 2024. Doubao models.

Jia Tina Du, Iris Xie, and Jenny Waycott. 2020. Marginalized communities, emerging technologies, and social innovation in the digital age: Introduction to the special issue. *Information Processing & Management*, 57(3):102235.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. *Ethnologue: Languages of the World*, 27 edition. SIL International, Dallas.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 2024a. Bertaqa: How much do language models know about local culture? *Advances in Neural Information Processing Systems*, 37:34077–34097.

Julen Etxaniz, Oscar Sainz, Naiara Miguel, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024b. Latxa: An open language model and evaluation suite for basque. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14952–14972.

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.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, et al. 2024. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, et al. 2024. Are we done with mmlu? *arXiv preprint arXiv:2406.04127*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Vipul Gupta, David Pantoja, Candace Ross, Adina Williams, and Megan Ung. 2024. Changing answer order can decrease mmlu accuracy. *arXiv preprint arXiv:2406.19470*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Kung Yin Hong, Lifeng Han, Riza Theresa Batista-Navarro, and Goran Nenadic. 2024. Cantonmt: Cantonese-english neural machine translation looking into evaluations. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pages 133–144.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card.

Jiyue Jiang, Liheng Chen, Pengan Chen, Sheng Wang, Qinghang Bao, Lingpeng Kong, Yu Li, and Chuan Wu. 2024. How far can cantonese nlp go? benchmarking cantonese capabilities of large language models. *arXiv e-prints*, pages arXiv–2408.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, et al. 2024. Arabicmmlu: Assessing massive multitask language understanding in arabic. *arXiv preprint arXiv:2402.12840*.

Chaak Ming Lau. 2024. Ideologically driven divergence in cantonese vernacular writing practices. In J.-F. Dupré, editor, *Politics of Language in Hong Kong*. Routledge.

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, Marseille, France. European Language Resources Association.

Genevieve Leung and Yuuko Uchikoshi. 2012. Relationships among language ideologies, family language policies, and children's language achievement: A look at cantonese-english bilinguals in the us. *Bilingual Research Journal*, 35(3):294–313.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

Mistral. 2024. Mistral nemo.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence.

Irene Plaza, Nina Melero, Cristina del Pozo, Javier Conde, Pedro Reviriego, Marina Mayor-Rocher, and María Grandury. 2024. Spanish and llm benchmarks: is mmlu lost in translation? *arXiv preprint arXiv:2406.17789*.

Itesh Sachdevl, Richard Bourhis, Sue-wen Phang, and John D'Eye. 1987. Language attitudes and vitality perceptions: Intergenerational effects amongst chinese canadian communities. *Journal of Language and Social Psychology*, 6(3-4):287–307.

SenseTime. 2024. Sensetime introduces sensechat (cantonese) to hong kong users, delivering localised ai experiences free-of-charge.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024. Kmmlu: Measuring massive multitask language understanding in korean. *arXiv preprint arXiv:2402.11548*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Ashima Suvarna, Harshita Khandelwal, and Nanyun Peng. 2024. PhonologyBench: Evaluating phonological skills of large language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 1–14, Bangkok, Thailand. Association for Computational Linguistics.

Saeid Asgari Taghanaki, Aliasgahr Khani, and Amir Khasahmadi. 2024. Mmlu-pro+: Evaluating higher-order reasoning and shortcut learning in llms. *arXiv preprint arXiv:2409.02257*.

Zhi-Rui Tam, Ya-Ting Pai, Yen-Wei Lee, Jun-Da Chen, Wei-Min Chu, Sega Cheng, and Hong-Han Shuai. 2024. An improved traditional chinese evaluation suite for foundation model. *arXiv preprint arXiv:2403.01858*.

toastynews. 2020. openrice-senti.

Maria Tsapali and Hiu Ching Wong. 2023. The future of cantonese and traditional chinese among newly arrived hong kong immigrant children in the united kingdom–a study on parents'attitudes, challenges faced and support needed. *Cambridge Educational Research e-Journal*, 10:14–31.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2024. Milu: A multi-task indic language understanding benchmark. *arXiv preprint arXiv:2411.02538*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Andrea W Wen-Yi, Unso Eun Seo Jo, and David Mimno. 2025. Do chinese models speak chinese languages? *arXiv preprint arXiv:2504.00289*.

Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preotiuc-Pietro. 2022. Cross-lingual few-shot learning on unseen languages. 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 (Volume 1: Long Papers)*, pages 777–791, Online only. Association for Computational Linguistics.

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.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kevin Xu, Yuqi Ye, and Hanwen Gu. 2024. A survey on multilingual large language models: Corpora, alignment, and bias. *arXiv preprint arXiv:2404.00929*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

William Yiu. 2024. Hong kong government to adopt city's own chatgpt-style tool after openai further blocks access.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. 2023. How well do large language models perform in arithmetic tasks? *arXiv preprint arXiv:2304.02015*.

Daoguang Zan, Bei Chen, Fengji Zhang, Dianjie Lu, Bingchao Wu, Bei Guan, Yongji Wang, and Jian-Guang Lou. 2022. Large language models meet nl2code: A survey. *arXiv preprint arXiv:2212.09420*.

Chen Zhang, Zhiyuan Liao, and Yansong Feng. 2025. Cross-lingual transfer of cultural knowledge: An asymmetric phenomenon. *arXiv preprint arXiv:2506.01675*.

Lubei Zhang, Linda Tsung, and Xian Qi. 2023. Home language use and shift in australia: Trends in the new millennium. *Frontiers in Psychology*, 14:1096147.