Appendix

A Questionable Practices in an Existing Work

Jiang et al. (2024) recently proposed a Cantonese evaluation dataset consisting of 5 datasets. The authors cited a HuggingFace organisation homepage in a footnote for their translation dataset but offered no further specifics. Yet, the dataset's entries were not identifiable within the linked account. Nonetheless, the translation pairs between Cantonese and English are dubious. During the machine translation process, the less advanced models before GPT3.5 tend to break English sentences into phases by punctuation marks or connecting words such as "and" and "but". Then the phases were translated individually and finally put together into one sentence in Cantonese. As a result, the translated texts were full of wrong wordings such as the following translation pair:

English:

> Once upon a time, there was a three year old girl named Gwen. One day, she was walking with her mom when she saw something unusual. She wondered out loud, â€œWhat is that?â€ Her mom explained, â€œThatâ€™s an old license. It shows that the person is allowed to drive.â€ Gwen then asked, â€œCan I get one?â€ Her mom smiled, shaking her head. â€œNo, not now.

Cantonese translation:

> 以前有個叫 Gwen 嘅小朋友，三年大。有一日，佢同媽媽行緊街，見到一件好特別嘅嘢。佢好奇地問：「呢個係咩嚟㗎？」媽媽解釋：「呢個係舊嘅駕駛執照，顯示呢個人可以駕駛。」Gwen 好興奮地問：「我可以有份㗎？！」媽媽笑住搖頭話：「唔得，而家仲細啦。

The age of a person was counted using 年 (year) instead of the correct word 歲 (years of age).

Another example is incomprehensible:

English:

> Once upon a time, there was a big dinosaur. He was very fast and could run really quickly. One day, the dinosaur went for a walk and saw a little boy. The boy was sad because he lost his toy car. The dinosaur felt sorry for the boy and decided to help him. He ran very fast to search for the toy car. After a while, the dinosaur found the toy car and returned it to the little boy. The boy was happy and thanked the dinosaur for being so kind. From that day on, the boy and the dinosaur became good friends and went on many adventures together.

Cantonese translation:

> 從前有隻好大嘅恐龍，佢好快，跑得好快。有一日，恐龍去散步，見到一個小男孩。個男孩好唔開心，因為佢唔見咗架玩具車。恐龍好同情個男孩，決定幫手。佢好快就跑去搵玩具車。一陣，恐龍搵到玩具車，交返俾個小男孩。個男孩好開心，多謝恐龍咁好心。由嗰日開始，個男孩同恐龍成為好朋友，一齊去好多冒險。

The translation service broke down the original English sentences by punctuation and connectives (such as "and"). Then, it translated the broken-down phases individually and joined them into a Cantonese/Chinese sentence, resulting in awkward and incorrect punctuation usage. The heavy use of poor machine translation of existing work offered little in terms of novel contributions or insights into the language.

The use of the BLEU score is inappropriate for Cantonese translation. Please see Appendix F for our rationale.

B Translated MMLU Dataset

The original MMLU dataset was translated into Cantonese by Gemini 1.5 Flash using the following prompt:

> 翻譯下面嘅選擇題做廣東話：
>
> *** Input: (Example question 1)
> ** A: (Example 1 option A)
> ** B: (Example 1 option B)
> ** C: (Example 1 option C)
> ** D: (Example 1 option D)
> ** Target: (Example 1 Answer)
>
> 廣東話翻譯：
>
> *** 輸入：(Manually translated Example Question 1)

** 選項 A：(Manually translated Example 1 option A)
** 選項 B：(Manually translated Example 1 option B)
** 選項 C：(Manually translated Example 1 option C)
** 選項 D：(Manually translated Example 1 option D)
** 目標：(Example 1 Answer)

==========

(Another four examples)

==========

翻譯下面嘅選擇題做廣東話：
** Input: (Input Question)
** A: (Option A)
** B: (Option B)
** C: (Option C)
** D: (Option D)
** Target: (Answer)

English translation of the prompt:

Translate the following multiple choice question to Cantonese：

*** Input: (Example question 1)
** A: (Example 1 option A)
** B: (Example 1 option B)
** C: (Example 1 option C)
** D: (Example 1 option D)
** Target: (Example 1 Answer)

Cantonese Translation：

*** Input：(Manually translated Example Question 1)
** Option A：(Manually translated Example 1 option A)
** Option B：(Manually translated Example 1 option B)
** Option C：(Manually translated Example 1 option C)
** Option D：(Manually translated Example 1 option D)
** Target：(Example 1 Answer)

==========

(Another four examples)

==========

Translate the following multiple choice question to Cantonese：
** Input: (Input Question)
** A: (Option A)
** B: (Option B)
** C: (Option C)
** D: (Option D)
** Target: (Answer)

C Academic and Professional Dataset

The Professional Dataset subset consists of 7 professional qualification exams:

1. **Estate Agents Qualifying Exam (EAQE)**: Exam from the Estate Agents Authority (EAA) which grants an estate agent's license for a person to become a director or a partner of an estate agency.

2. **Insurance Intermediaries Qualifying Examination**: The Insurance Authority in Hong Kong regulates and licenses all insurance intermediaries in Hong Kong and offers this exam for any person who wishes to carry out insurance activity.

3. **Leveraged Foreign Exchange Trading Examination**: An industry qualification exam offered by the Vocational Training Council (VTC) and approved by the Securities and Future Commission.

4. **Licensing Exam for Securities & Futures Intermediaries (LE)**: An exam offered by the Hong Kong Securities and Investment Institute and recognised by the Securities and Futures Commission (SFC). It is a professional qualification exam for people who would like to work in the securities and investment industry in Hong Kong.

5. **Mandatory Provident Fund Schemes (MPF)/Intermediaries Exam**: The MPF is the pension scheme in Hong Kong and any intermediaries involved in providing the service are required to pass this exam.

6. **Pleasure Vessel Operator Grade 2 Certificate**: The Marine Department of Hong Kong offers this exam for anyone who wishes to operate a boat less than 15 meters in length for pleasure purposes.

7. **EAA Salespersons Qualifying Examination**: Exam from the EAA which grants a salesperson's license for a salesperson to carry out estate agency work.

8. **Taxi License Written Exam Route Planning Exam**: Unique questions created in the form of the Taxi License Written Exam Route Planning Exam which are multiple-choice questions about selecting the shortest route between any two places or landmarks without considering toll fees.

All the exams suggested above were sourced from PDF files provided by the exam provider and processed by Google Gemini 1.5 Flash. The pages containing the model answers were separated from the questions pages, which could mitigate the data contamination issue. The following sentence is added to the front of the prompt during the evaluation:

> Follow the given examples and answer the question. The question is about professional knowledge in Hong Kong. You should only return the answer: A, B, C, or D.

The next two law-related categories were sourced from the Internet:

1. **Hong Kong Law**: Questions across 15 categories such as child abuse, domestic violence, the Employment Ordinance, Equal Opportunities, family law, etc. were gathered across the Internet.

2. **Basic Law**: Constitution-type question sourced from a website. The questions were updated according to the current and actual practice of the Basic Law and options were also appended to 4 options per question by the authors.

A sentence describing the task is added at the front of the prompt:

> Follow the given examples and answer the question. The question is about Hong Kong law. You should only return the answer: A, B, C, or D.

Name	Source	No.	Eng.
Estate Agent	PDF	50	Y
Insurance Interm.	PDF	130	Y
Leveraged FX	PDF	20	Y
Securities & Futures	PDF	400	Y
MPF Interm.	PDF	26	Y
Pleasure Vessel	PDF	66	Y
Salesperson	PDF	50	Y
Taxi	New	30	N
HK Law	Web	360	N
Basic Law	Web	64	N
Total		1,196	

Table 1: Professional Dataset Summary

The number of questions and the source of each professional exam can be found in Table 1.

The Academic Dataset subset was sourced from scanned copies of the HKDSE exam from the last 2 to 4 years. The details can be found in Table 2. The scanned PDFs were processed to extract text with the Gemini 1.5 Pro API. The extracted text was then filtered by the authors to remove questions that required information from other questions. Those requiring references to visual materials like pictures and maps were also removed but reserved for future evaluation of vision-enabled LLMs.

Although only multiple-choice questions were included in the dataset, more complicated questions that require complex reasoning and human judgment, such as Chinese listening and Liberal Studies exams were studied. Initial testing showed proprietary LLMs can easily achieve perfect scores in the Chinese listening exam when given with the transcript, while the length of the transcript exceeded some open source LLMs' context length. Further experiments were conducted with Google Gemini 1.5 Flash and Pro to take the audio file and written questions as input. Although the two models again achieved a perfect score, no other model supported audio input at the time of the experiment, so the task was excluded. For the now-defunct Liberal Studies, examinees were required to write a passage taking into account the provided reference materials and then develop their own point of view supported by examples from the candidates' own knowledge. However, initial testing showed that none of the LLMs were able to include their

Subject	No.	Eng.
Biology	52	Y
Business, Accounting & Financial Studies	50	Y
Chemistry	42	Y
Economics	46	Y
Geography	54	Y
Information & Comm. Tech.	73	Y
Mathematics	63	Y
Physics	30	Y
Tourism & Hospitality Studies	56	N
Total	466	

Table 2: Academic Dataset Summary

own examples despite explicitly requested in the prompt.

During the evaluation, a short sentence is added to the front to explain the nature of the dataset:

> Follow the given examples and answer the question. The question is about Hong Kong DSE. You should only return the answer: A, B, C, or D.

D Hong Kong Cultural Questions Dataset

The Hong Kong Cultural Questions Dataset contains four categories of newly created questions and one existing question set, which the breakdown can be found in Table 3.

1. **Food Culture**: Hong Kong has a unique culinary and food culture due to cultural influences from East and West. The questions were designed to highlight local dishes, street food, high-end dining, and "茶餐廳" Cha chaan teng (Hong Kong-styled cafe) culture.

2. **History and Landmarks**: Questions in this category assess the knowledge of significant historical events, figures, and iconic landmarks that have shaped Hong Kong's identity.

3. **Language and Expressions**: This category focuses on Cantonese-specific vocabulary, colloquialisms, slang, and common expressions used in daily life and social media networks.

Subject	No. of Q.	New Q.
History & Landmarks	61	Y
Food Culture	59	Y
Lang. & Expressions	49	Y
Life in HK	75	Y
Local Area Knowledge	33	N
Total	277	

Table 3: Summary of the Hong Kong Cultural Questions Dataset

4. **Life in Hong Kong**: This category covers the customs, traditions, social norms, popular culture (TV and cinema), recent events and everyday life experiences unique to Hong Kong.

5. **Local Area Knowledge**: Contains 33 questions on non-trivial local area knowledge selected from the Hong Kong Geography Tatsujin Challenge, an online questionnaire participated by over 100,000 participants.

A question bank of 244 questions, written in pure Cantonese, was created for the first four categories. To ensure the questions were based on common knowledge of the population, these questions were sent to 30 Hong Kong volunteers aged 20-50 years, and they were instructed not to use a search engine when attempting the questions. Consent was sought for using their data to inform the design of an LLM benchmark for Cantonese. Items with an accuracy under 40.0% were excluded. For the last category, the questions were randomly selected from the Hong Kong Geography Tatsujin Challenge, which we have access to question-based correctness measures. The average correct rate of the selected questions is 53.0% for this subset. The questions in this subcategory were in Written Chinese.

The following sentence is added to the front of the prompt:

> Follow the given examples and answer the question. The question is about Hong Kong. Only return the answer: A, B, C, or D. DO NOT EXPLAIN.

Given the unique and novel challenges presented by this dataset, a breakdown of model performance at five-shot evaluation is included in Table 4. The average score of the human evaluation was also included as a reference, where no model

outperformed humans in the Food Culture and Language and Expressions category. SenseChat (Cantonese) refused to answer any questions in the History & Landmark due to the appearance of the name of the Hong Kong Chief Executive John Lee in the 5-shot examples.

E Linguistic Knowledge Dataset

This dataset comprises three sub-datasets with questions carefully designed with expert input from researchers specialising in different areas of Cantonese linguistics. The structure and questions of each dataset are outlined below. The results of all three can be found in Table 5.

E.1 Phonological Knowledge Dataset

The Phonological Knowledge Dataset include three groups of questions: Homophone Judgment (25 questions), Rhyme Judgment (25 questions), and a group of Phonological Reasoning Tasks (MultiPron Resolution, Tone Matching, Poetry Rhyme, Shared Feature Judgment, and Couplet Reasoning, 50 questions in total).

1. **Homophone Judgment**: The task is to determine which character from a list shares the same pronunciation in Cantonese as the given character, if any. For example, if the given character is 一 and the list of characters is A 二, B 逸, C 醫, D 倚, the answer would be none of the above, as none of the characters (ji6, jat6, ji1, ji2) is a homophone of the given character (jat1).

2. **Rhyme Judgment**: The task is to determine which character from a list rhymes with a given character in Cantonese. For example, if the given character is 民 and the list is A 森, B 敢, C 笨, D 主, the answer will be C, as it is pronounced b**an**6 in Cantonese, which rhymes with the given character m**an**4.

3a. **MultiPron Resolution**: A word that contains a character that can be pronounced in multiple ways (like the English word read and be pronounced as red and ri:d), and the task is to decide how this character should be pronounced. For example, if the given word is 過河 and the question asks for a homophone for the 2nd character, given the list A 可 B 賀 C 呵 D 和, the answer is none of the above. The given character should be pronounced as ho4 and none of the answer (ho2, ho6, ho1, wo4) matches this pronunciation.

3b. **Tone Matching**: Given a polysyllabic word, choose a word from a list that matches the tonal (pitch) pattern based on tone-melodic requirement. For example, given the word 痴線, and the list A 低 B, B 多次, C 擔心, D 磁線, the answer is B. Its pronunciation is do**1**ci**3**, which is a high tone followed by a mid tone, same as the given word ci**1**sin**3**.

3c. **Poetry Rhyme**: Given a poem, choose the correct combination of rhyming characters, e.g. in the poem 白日依山盡，黃河入海流，欲窮千里目，更上一層樓。The last syllables are z**eon**6, l**au**4, m**uk**6, l**au**4 in Cantonese, 流樓 are the two characters that rhyme in the poem when recited in Cantonese.

3d. **Shared Feature Judgment**: The task is to choose an odd character from a list that does not share a common phonological feature. For example, the list A 誤 B 牙 C 打 D 沙 is ng6, ng**aa**4, d**aa**2, s**aa**1 in Jyutping. The last three characters share the same rhyme aa, so A is the odd one.

3e. **Couplet Reasoning**: One line of a couplet will be given. Characters in a couplet can be divided into two rhythmic categories, 平 (ping4, **L**evel) and 仄 (zak1, **O**blique). This distinction is taught at school and is a long-standing poetry tradition. The task is to determine the category for each of the characters, and decide whether it is the upper (上聯) or lower (下聯) line. For example, the line 天增歲月人增壽 is tin**1** zang**1** seoi3 jyut6 jan**4** zang**1** sau6 in Jyutping, and its pattern is LLOOLLO, and as it ends with an oblique sound, it is the upper line. Therefore it is described as 上聯、平平仄仄平平仄.

The questions were evaluated using the following system prompt:

> You are a speaker of Cantonese from Hong Kong. Please answer these questions about the sounds of the language. Do not include any further explanation.

Example questions for each task:

Model	Food Culture	History & Landmarks	Language & Expressions	Life in HK	Local Area Knowledge	Average
Claude 3.5 Sonnet	74.1%	80.4%	70.5%	**85.7%**	64.3%	75.0%
Doubao Pro	81.5%	82.1%	77.3%	77.1%	57.1%	75.0%
Ernie 4.0	74.1%	75.0%	75.0%	84.3%	67.9%	75.2%
Gemini 1.5 Flash	59.3%	69.6%	56.8%	77.1%	57.1%	64.0%
Gemini 1.5 Pro	75.9%	82.1%	65.9%	84.3%	60.7%	73.8%
GPT4o	83.3%	**83.9%**	68.2%	82.9%	**67.9%**	77.2%
GPT4o-mini	57.4%	67.9%	50.0%	74.3%	53.6%	60.6%
SenseChat	**87.0%**	N.A.	**79.6%**	78.6%	64.3%	**77.4%**
Aya 23 8B	35.2%	46.4%	34.1%	44.3%	28.6%	37.7%
CLLM 6B	51.9%	67.9%	63.6%	54.3%	42.9%	56.1%
CLLM 34B	**77.8%**	**87.5%**	**72.7%**	74.3%	71.4%	76.7%
Yi 1.5 6B	40.7%	64.3%	50.0%	55.7%	42.9%	50.7%
Yi 1.5 9B	53.7%	60.7%	52.3%	70.0%	50.0%	57.3%
Yi 1.5 34B	74.1%	82.1%	61.4%	75.7%	71.4%	72.9%
Gemma 2 2B	29.6%	37.5%	34.1%	42.9%	32.1%	35.2%
Gemma 2 9B	48.2%	57.1%	50.0%	60.0%	42.9%	51.6%
Gemma 2 27B	61.1%	69.6%	56.8%	67.1%	50.0%	60.9%
Llama 3.1 8B	51.9%	57.1%	50.0%	54.3%	50.0%	52.7%
Llama 3.1 70B	61.1%	78.6%	52.3%	72.9%	57.1%	64.4%
Llama 3.1 405B	64.8%	75.0%	59.1%	75.7%	75.0%	69.9%
Mistral Nemo 12B	31.5%	53.6%	45.5%	54.3%	28.6%	42.7%
Qwen2 7B	55.6%	67.9%	50.0%	68.6%	32.1%	54.8%
Qwen2 72B	75.9%	82.1%	**72.7%**	**82.9%**	**75.0%**	**77.7%**
Random	35.2%	21.4%	27.3%	31.4%	25.0%	28.1%
Human	81.7%	73.4%	85.8%	83.9%	53.0%	75.6%

Table 4: Model performance on the Hong Kong Cultural Questions Dataset in 5-shots settings

Model	Phonological Knowledge				Orthographic Knowledge			
	Homophone	Rhyme	Misc.	Avg.	Visual Similarity	Canton. Char.	Misc.	Avg.
Claude 3.5 Sonnet	28.0%	64.0%	16.0%	31.0%	50.0%	76.9%	59.3%	64.5%
Doubao Pro	16.0%	44.0%	16.0%	23.0%	70.0%	80.8%	48.1%	64.5%
Ernie 4.0	28.0%	60.0%	18.0%	31.0%	70.0%	80.8%	53.7%	67.5%
Gemini 1.5 Flash	12.0%	20.0%	24.0%	20.0%	40.0%	73.1%	31.5%	46.0%
Gemini 1.5 Pro	16.0%	40.0%	24.0%	26.0%	50.0%	**88.5%**	46.3%	60.5%
GPT4o	**56.0%**	**96.0%**	**28.0%**	**52.0%**	50.0%	65.4%	**63.0%**	63.5%
GPT4o-mini	20.0%	60.0%	20.0%	30.0%	30.0%	57.7%	40.7%	44.5%
SenseChat	16.0%	36.0%	22.0%	24.0%	**75.0%**	76.9%	42.6%	61.8%
Aya 23 8B	12.0%	40.0%	14.0%	20.0%	15.0%	19.2%	31.5%	25.8%
CLLM 6B	24.0%	8.0%	18.0%	17.0%	20.0%	50.0%	27.8%	33.0%
CLLM 34B	28.0%	28.0%	14.0%	21.0%	35.0%	76.9%	37.0%	48.8%
Yi 1.5 6B	28.0%	12.0%	12.0%	16.0%	10.0%	50.0%	20.4%	26.5%
Yi 1.5 9B	**36.0%**	40.0%	24.0%	**31.0%**	30.0%	57.7%	18.5%	32.5%
Yi 1.5 34B	16.0%	32.0%	**26.0%**	25.0%	30.0%	61.5%	33.3%	41.5%
Gemma 2 2B	8.0%	24.0%	18.0%	17.0%	25.0%	53.8%	22.2%	32.3%
Gemma 2 9B	20.0%	28.0%	24.0%	24.0%	25.0%	50.0%	33.3%	37.3%
Gemma 2 27B	20.0%	12.0%	16.0%	16.0%	25.0%	65.4%	24.1%	36.3%
Llama 3.1 8B	12.0%	16.0%	18.0%	16.0%	25.0%	42.3%	38.9%	38.3%
Llama 3.1 70B	28.0%	40.0%	12.0%	23.0%	30.0%	61.5%	35.2%	42.5%
Llama 3.1 405B	20.0%	**44.0%**	18.0%	25.0%	35.0%	65.4%	**50.0%**	52.8%
Mistral Nemo 12B	12.0%	28.0%	10.0%	15.0%	25.0%	23.1%	37.0%	32.3%
Qwen2 7B	8.0%	40.0%	12.0%	18.0%	35.0%	46.2%	33.3%	38.8%
Qwen2 72B	12.0%	28.0%	16.0%	18.0%	**50.0%**	**76.9%**	48.1%	**58.5%**
Random	16.0%	28.0%	24.0%	23.0%	30.0%	11.5%	27.8%	25.5%

Table 5: Model performance on Linguistic Knowledge Dataset multiple-choice questions

1. Which of the following character is a homophone of ‘一’ in Cantonese? (A) 二 (B) 逸 (C) 醫 (D) 倚 **(E) None of the above**

2. Which of the following character rhymes with the character ‘民’ in Cantonese? (A) 森 (B) 敢 **(C)** 笨 (D) 主 (E) None of the above

3a. Which of the following character best represents the sound of ‘河’ in the word ‘過河’ in Cantonese? (A) 可 (B) 賀 (C) 呵 (D) 和 **(E) None of the above**

3b. Which of the following words have the nearest tone or pitch pattern with ‘痴線’ in Cantonese? (A) 低 B **(B)** 多次 (C) 擔心 (D) 磁線 (E) None of the above

3c. In the following poem
‘‘‘白日依山盡，
黃河入海流。
欲窮千里目，
更上一層樓。’’’
What are the last words of the rhyming lines if recited in Cantonese? ‘(A) 盡樓 (B) 盡流樓 (C) 流目 **(D)** 流樓 (E) None of the above’

3d. Three of the four characters below share a common phonological feature in Cantonese, and one does not. Which one is the odd one? **(A)** 誤 (B) 牙 (C) 打 (D) 沙 (E) None of the above

3e. This is a line of a couplet (對聯): ‘天增歲月人增壽’
Which of the following is a correct description of the line, according to its Cantonese pronunciation? **(A)** 上聯、平平仄仄平平仄 (B) 下聯、平平仄仄平平仄 (C) 上聯、平仄平仄仄仄平 (D) 下聯、平仄平仄仄仄平 (E) Cannot be determined.

E.2 Orthographic Knowledge Dataset

The Orthographic Knowledge Dataset consists of three sub-tasks: Visual Similarity Judgment

(25 questions), Cantonese Character Selection (26 questions), and a group of Orthographic Reasoning Tasks (Character Calculation, Radical Description, Character Structure, 54 questions in total).

1. **Visual Similarity Judgement**: The task is to determine which character from a list is the most visually similar to the given character. For example, given the character 未, and the list "A 尾 B 末 C 味 D 林", the answer will be B, as the character is only different from the given character in the length of the horizontal strokes.

2. **Cantonese Characters Selection**: Cantonese words do not necessarily have a standardised form. Some very common characters can be written in multiple ways. The task provides a sentence frame with a missing character, which should be clear from the context what the syllable is. This is followed by a list of possible characters for insertion. For example, given the sentence "我__得滯已經食唔落" (I am too *something* and I cannot finish eating this.), it should be filled with a word that means "full"in Cantonese. Given the list "A 撐 B 包 C 鮑 D 飽", D should be chosen as it is the most accepted word for the Cantonese word *baau2*. Answer A is marginally acceptable in Written Chinese, but it commonly means "to support" (caang3) in Cantonese, thus should not be selected.

3a. **Character Calculation**: The task is to determine the end result of operations that involve adding, removing or changing glyph components of a character, and choose the closest character from a list. For example, given the instruction to remove " 刂" from "别", and a list of "A 另 B 叻 C 刀 D 龜", the answer will be A as this is the closest character that matches with the instruction.

3b. **Radical Description**: A character will be given, followed by a list of Kangxi radicals. For example, the character "沖" is listed under the "水" (water) radical, and the Cantonese name for the radical is "三點水" (three drops of water).

3c. **Character Structure**: The task is to choose the best description of the structure of the given character. For example, the character "知" is formed by combining two characters in the left-right frame ("左右結構").

The questions were evaluated using the following system prompt:

> You are a speaker of Cantonese from Hong Kong. Please answer these questions about the properties of the language. Do not include any further explanation.

Example questions for each task:

1. Select the character that is visually similar to 、未、in the Hong Kong context (A) 尾 **(B)** 末 (C) 味 (D) 林 (E) None of the above

2. Consider this Cantonese sentence、我__得滯已經食唔落、. Choose one character below that is the most widely-accepted way to represent the missing word. (A) 撐 (B) 包 (C) 鮑 **(D)** 飽 (E) None of the above

3a. What character do you get by removing 、刂、from 、别、? **(A)** 另 (B) 叻 (C) 刀 (D) 龜 (E) None of the above

3b. What is the radical of the character 、沖、and how is it called in Cantonese? (A) 中（水中） (B) 行（彳亍行）(C) 氵（兩點水）**(D)** 水（三點水） (E) None of the above

3c. What is the best description of the character structure of 、知、? (A) 上下 (B) 包圍 (C) 左中右 (D) 前後 **(E) None of the above**

E.3 Grapheme-to-Phoneme (G2P) Conversion Dataset

Cantonese transliteration is not trivial because the characters and pronunciation form a many-to-many relation. The same syllable can be represented by different characters, which is a crucial feature of an ideographic writing system, whereas the same character may have multiple pronunciations due to the overloading of certain characters or multiple layers of pronunciation norms. These judgements are often not well-documented. Characters with multiple pronunciations are often semantically or lexically determined, for example: "行" can be pronounced as *hang4* (e.g. "行動" action, "流行" trend), *haang4* (e.g. "行路" to walk, "行街" to go shopping), *hong4* (e.g. "銀行" bank, "行業" occupation), *hong2* (e.g. "投行"

investment bank). For a smaller subset of characters, there can be different pronunciations due to a literary-colloquial distinction, e.g. “坐” (to sit) can be *zo6* or *co5*; “請” (to invite) can be *cing2* or *ceng2*.

The five-shot evaluation prompt used for the G2P dataset evaluation:

> You are an expert in Cantonese linguistics. Please convert the given Cantonese sentence into Jyutping romanisation. You can ignore all punctuation marks, and normalise all numerals and English loanwords into Cantonese pronunciation. Do not include any further explanation.
>
> **Example 1**
> 今日天氣好好，不如出去散步？
> gam1 jat6 tin1 hei3 hou2 hou2 bat1 jyu4 ceot1 heoi3 saan3 bou6
>
> **Example 2**
> 我鍾意行商場唔鍾意行街市
> ngo5 zung1 ji3 haang4 soeng1 coeng4 m4 zung1 ji3 haang4 gaai1 si5
>
> **Example 3**
> 朱咪咪係唱歌好叻嘅歌手
> zyu1 mi1 mi4 hai6 coeng3 go1 hou2 lek1 ge3 go1 sau2
>
> **Example 4**
> 今天是學校正常上課日，請各位家長督促子女準時上學。
> gam1 tin1 si6 hok6 haau6 zing3 soeng4 soeng5 fo3 jat6 ceng2 gok3 wai2 gaa1 zoeng2 duk1 cuk1 zi2 neoi5 zeon2 si4 soeng5 hok6
>
> **Example 5**
> 學而時習之，不亦說乎
> hok6 ji4 si4 zaap6 zi1 bat1 jik6 jyut6 fu4

Score calculation: All acceptable variants have been listed as answers and all variants are considered equally good. This is to handle variant forms and ambiguous interpretations that may not be the standard, but native speakers of Cantonese accept. We will use the answer with the highest score for the subsequent calculation. Two metrics were used: character error rate (CER) and Levenshtein distance. A lower score means a better performance. The answers with the lowest Levenshtein distance (i.e. the best score) were used for the calculation. See Table 6 for the scores.

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 6B	94.1%	0.859
CLLM 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	5.0%	0.009

Table 6: Model performance in the Grapheme-to-Phoneme (G2P) dataset. Scores calculated based on character error rates (CER) and Levenshtein distance. (Lower is better)

F Translation Task

The translation dataset consists of 20 Cantonese sentences. These sentences were designed to check the models’ understanding of sentence nuances as they involve lexical or contextual ambiguities that require good linguistic reasoning to resolve. Each Cantonese sentence was translated into both English and written Chinese by a professional translator. This provides the source for four translation pairs per original Cantonese sentence: Cantonese-English, Cantonese-Written_Chinese, English-Cantonese, Written_Chinese-Cantonese. Here is an example sentence taken from the few shot examples:

Cantonese (avg. 28.6 characters):

> 返工返到得返半條人命，搞到自己身體嗷樣，賺埋都唔夠你睇醫生啦。

English (avg. 23.1 words):

> Your day job is so exhausting that you

Model	Translation (0-shot)	Translation (3-shot)	Summarisation	Sentiment
Claude 3.5 Sonnet	96.2%	91.9%	**92.7%**	76.0%
Doubao Pro	**97.5%**	97.4%	85.5%	67.7%
Ernie 4.0	86.2%	87.1%	83.3%	74.1%
Gemini 1.5 Flash	92.6%	95.5%	70.0%	74.9%
Gemini 1.5 Pro	95.3%	91.1%	90.0%	75.3%
GPT4o	96.7%	**98.3%**	83.7%	**79.7%**
GPT4o-mini	89.8%	95.2%	84.8%	74.7%
SenseChat	90.6%	93.8%	54.3%	76.5%
Aya 23 8B	81.0%	79.4%	52.8%	67.1%
CLLM 6B	**97.5%**	94.7%	29.0%	66.5%
CLLM 34B	86.6%	90.3%	43.2%	73.1%
Yi 1.5 6B	62.9%	56.7%	42.5%	64.3%
Yi 1.5 9B	69.5%	88.1%	62.2%	69.1%
Yi 1.5 34B	87.3%	95.8%	70.3%	78.2%
Gemma 2 2B	71.4%	64.4%	86.3%	71.6%
Gemma 2 9B	87.8%	88.2%	**91.3%**	72.6%
Gemma 2 27B	89.3%	78.4%	89.0%	76.1%
Llama 3.1 8B	76.3%	81.1%	15.3%	68.6%
Llama 3.1 70B	92.1%	86.6%	81.7%	77.5%
Llama 3.1 405B	77.5%	93.3%	8.0%	**78.8%**
Mistral Nemo 12B	77.0%	84.8%	40.7%	72.7%
Qwen2 7B	84.5%	90.1%	14.8%	77.9%
Qwen2 72B	97.1%	**96.6%**	62.2%	78.2%
Model Answer	99.1%	99.1%	-	-

Table 7: Model performance in various NLP tasks

are like half-dead after work. If you keep pushing yourself like this, the money you earn won't even cover your medical bills.

Written Chinese (avg. 29.6 characters):

打工打到只剩下半條命，把自己身體弄成這樣，賺了的錢也不夠你去看醫生。

Evaluation prompt for the three-shot translation task (from Written Chinese to Cantonese) (Note: the prompt uses the word "Traditional Chinese" to force the model to return the results in the Traditional script.):

參考以下例子，將以下繁體中文句子翻譯成香港的廣東話：

例子 1：中文書面語: (e.g. 1) 廣東話: (e.g. 1)
例子 2：中文書面語: (e.g. 2) 廣東話: (e.g. 2)
例子 3：中文書面語: (e.g. 3) 廣東話: (e.g. 3)

請翻譯以下文本：
(Text to translate)

確保翻譯準確自然，並符合香港的廣東話語法及表達方式

只需回覆翻譯部份，不需要解釋

English translation of the prompt:

Translate the following Traditional Chinese sentence to the Cantonese used in Hong Kong, referring to the examples below:

Example 1: Written Chinese: (e.g. 1) Cantonese: (e.g. 1)
Example 2: Written Chinese: (e.g. 2) Cantonese: (e.g. 2)
Example 3: Written Chinese: (e.g. 3) Cantonese: (e.g. 3)

Text to Translate:
(Text to translate)

Ensure the translation is accurate and natural, preserving the original meaning and using concise and fluent English expression.
Only return the translation. Do not explain.

The other 2 translation pairs used a similar format, but an English prompt is used when translating from Cantonese to English:

Translate the following Cantonese text into English, referring to the examples below:

Example 1: Cantonese: (e.g. 1) English: (e.g. 1)
Example 2: Cantonese: (e.g. 2) English: (e.g. 2)
Example 3: Cantonese: (e.g. 3) English: (e.g. 3)

Text to Translate:
(Text to translate)

Ensure the translation is accurate and natural, preserving the original meaning and using concise and fluent English expression.
ONLY RETURN THE TRANSLATION. DO NOT EXPLAIN.

For zero-shot evaluation, similar prompts were used but without the few-shot examples and references to them.

F.1 Translation Results Evaluation

We used manual evaluation for the translation task. The BLEU score was not used in this work to evaluate the translation task because it relied on exact matches between the answer and the gold standard, which is often not ideal for pairs that involve Written Chinese. Translation between closely related varieties in a diglossic situation is similar to stylistic change, with a wide range of acceptable answers. One kind of translation strategy would be favoured by using a BLEU score, and models that could have been rated excellent by Hong Kong users would be penalised. Adding to this is the lack of existing libraries that handle orthographic variants well enough to conduct a fair string comparison with the gold standard. This is why manual

annotation was opted for by us. This is to address the problems of an earlier benchmark (Jiang et al., 2024), outlined in Appendix A.

A graphical user interface based on Label Studio was configured such that annotators (paid undergraduate students from Hong Kong) can highlight mistakes in the translated text for the following categories:

1. Additional words
2. Collocation
3. Inconsistent Terminology
4. Literal Translation
5. Mistranslation
6. Mixed up Cantonese-Mandarin
7. Orthography
8. Omission (Labelled in the source text)
9. Register Issues
10. Ungrammatical
11. Unintelligibility
12. Unnatural Code-mixing

The annotators were required to label the five most relevant issues in the translated text, and every sentence was evaluated by four annotators. It should be noted that translation performance can be highly subjective, especially for languages like Cantonese that do not have a well-defined standard norm for its written form.

Some models exhibited a tendency to repeat the last sentence excessively, resulting in extra words beyond the expected translation. To account for this, if the "Additional words" labelled constituted more than half of the generated output, the labelled additional words were removed and not counted as erroneous output. The accuracy of the translation was then calculated using the following formula:

$$accuracy = \frac{len_{target} - len_{labelled\ error_{target}}}{len_{target} + len_{labelled\ error_{source}}} \quad (1)$$

The average score across all translation pairs of all models can be found in Table 7.

G Summarisation

The following prompt was used in the summarisation task evaluation:

> 我哋會提供一段用廣東話寫成嘅文本。請你將文本概括成 200 字嘅廣東話，保留核心訊息同主題，確保內容準確、行文流暢連貫，並且忠於原意。
>
> ## 文本（原文）：*(Original Text)*
>
> ## 預期輸出：

English translation of the prompt:

> We will provide a text written in Cantonese. Please summarise the text into a 200 words Cantonese text while preserving the core message and theme. Ensure the content is accurate, the writing flows smoothly and coherently, and it remains faithful to the original meaning.
>
> ## Text（Original text）：*(Original Text)*
>
> ## Expected Output：

G.1 Summarisation Results Evaluation

For the summarisation outputs, annotators (paid undergraduate students and research assistants from Hong Kong) were given the following rubric and tasked to give a score for each output. Under each category, the annotators rated whether the summary passes (5 points) or fails (3 points for minor violation, 1 point for complete failure).

Task instructions given to annotators:

> 你會收到多篇 TED 演講嘅廣東話逐字稿或廣東話文章（原文）你嘅工作係閱讀原文，然後按照指定準則為不同版本嘅撮寫評分首先你需要閱讀原文，理解文章大意可以讀多過一次然後開始逐篇撮寫閱讀實際評分喺 Google Form 上進行

English Translation:

> You will receive multiple Cantonese transcripts of TED talks or passages (*Original Text*). Your task is to read the original text and then score different versions of the summaries according to the specified guidelines. First, you need to read the original text to understand the main idea of the article. You may read

it more than once. Then, begin reading the summaries one by one. The actual scoring will be conducted on a Google Form.

The annotation rubrics were explained in detail in face-to-face meetings and in online working chat. A manually-crafted marking scheme was prepared to ease the annotation task.

Annotation Rubrics (English Translation):

- **Category: Relevance**
  The summary successfully retains the key points of the original text, rather than extracting overly verbose content, judged by whether the summary contains all the key points listed in the reference marking scheme provided. Omitted content and problematic sentences should be added as a comment if a fail (3 or 1) rating is given.
- **Category: Accuracy**
  The summary correctly extracted information from the original text, judging by comparing sentence by sentence with the original text, to identify any completely opposite or fabricated content. Sentences that contain incorrect information should be added as a comment if a fail rating is given.
- **Category: Fluency**
  The words and sentence structures used in the summary are smooth and conforming to Cantonese usage. This score does not require referencing the original text. The text should be acceptable for reading on a news programme. If words that are exclusive to Written Chinese (e.g. 這, 的, 了) or grammatical issues were found, a fail rating should be given.
- **Category: Coherence**
  The sentences and paragraphs in the summary are logically coherent, with appropriate structure and adequate sentential connection. This score does not require referencing the original text.

Score Calculation: Scores from all raters on the four categories were aggregated and normalised to 100%, then a length penalty is applied, based on its length: if the summary exceeds 500 characters, a penalty factor is calculated by reducing the score proportionally to the excess length, ensuring the penalty factor remains between 0 and 1. The resultant score can be found in Table 7.

H Sentiment Analysis

As detailed in Section **??**, the sentiment analysis dataset consists of an existing dataset known as the OpenRice (toastynews, 2020) dataset and a newly created dataset. The newly created datasets were sourced from Facebook comments and filtered by CantoneseDetect (Lau et al., 2024). Paid interns (undergraduate students from Hong Kong) then labelled the comments with the following guidelines:

> 呢個工作嘅目標係要判斷一段文字表達嘅情感係正面 (positive)、負面 (negative)，定係中性 (neutral)。我哋會用三個標籤：正面 (positive), 負面 (negative), 同埋中性 (neutral)。請仔細閱讀以下嘅指引，確保標註嘅一致性同埋準確性。
>
> 正面 **(positive)**
>
> 1. 意思: 表達開心、滿意、讚賞、樂觀、或者其他正面嘅情感。
> 2. 解釋: 文字入面嘅內容令人感覺良好、開心、或者對事物有正面評價。
> 3. 例子: “呢間餐廳啲嘢食真係好好味！服務又好！”
> 4. 例子: “今日天氣好好，心情都靚晒！”
>
> 負面 **(negative)**
>
> 1. 意思: 表達唔開心、唔滿意、批評、嬲怒、悲傷、或者其他負面嘅情感。
> 2. 解釋: 文字入面嘅內容令人感覺唔好、唔開心、或者對事物有負面評價。
> 3. 例子: “間酒店嘅服務態度真係好差，以後都唔會再嚟！”
> 4. 例子: “份報告寫得一啲都唔清楚，睇到我好嬲！”
>
> 中性 **(neutral)**
>
> 1. 意思: 表達客觀事實、資訊、或者冇明顯情感嘅內容。
> 2. 解釋: 文字主要係陳述事實、提供資訊，冇明顯嘅正面或負面情感傾向。

3. 例子: “今日香港嘅最高氣溫係攝氏 32 度。”
4. 例子: “呢個產品嘅主要功能係幫助用戶管理時間。”

Translation in English:

This task aims to determine whether the sentiment expressed in a given text is positive, negative, or neutral. We will use three labels: positive, negative, and neutral. Please carefully read the following guidelines to ensure consistency and accuracy in labelling.

Positive

1. Meaning: Expresses happiness, satisfaction, appreciation, optimism, or other positive emotions.
2. Explanation: The content of the text makes people feel good and happy, or has a positive evaluation of things.
3. Examples:
   (a) “The food in this restaurant is really delicious! The service is also great!”
   (b) “The weather is so nice today, it makes me feel good!”

Negative

1. Meaning: Expresses unhappiness, dissatisfaction, criticism, anger, sadness, or other negative emotions.
2. Explanation: The content of the text makes people feel bad, unhappy, or has a negative evaluation of things.
3. Examples:
   (a) “The service attitude of this hotel is really bad, I will never come again!”
   (b) “This report is not clear at all, it makes me very angry!”

Neutral

1. Meaning: Expresses objective facts, information, or content without obvious emotions.
2. Explanation: The text mainly states facts, provides information, and has no obvious positive or negative emotional tendency.
3. Examples:
   (a) “The highest temperature in Hong Kong today is 32 degrees Celsius.”
   (b) “The main function of this product is to help users manage their time.”

The average score across the two datasets of each model can be found in Table 7.

References

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.

Chaak Ming Lau, Mingfei Lau, and Ann Wai Huen To. 2024. The extraction and fine-grained classification of written cantonese materials through linguistic feature detection. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI)@ LREC-COLING 2024*, pages 24–29.

toastynews. 2020. openrice-senti.