Next-Level Cantonese-to-Mandarin Translation: Fine-Tuning and Post-Processing with LLMs

Yuqian Dai, Chun Fai Chan, Ying Ki Wong, Tsz Ho Pun
Logistics and Supply Chain MultiTech R&D Centre
{yuqian.dai,cfchan,skwong,thpun}@lscm.hk

Abstract

Large Language Models (LLMs) have improved performance across various natural language processing tasks. Despite these improvements, LLMs continue to face significant challenges, such as grammatical issues and code-switching to English, when applied to low-resource languages like Cantonese in Machine Translation (MT) scenarios. By addressing the unique linguistic and contextual challenges of Cantonese, we present a novel strategy to improve the understanding and translation capabilities of LLMs for Cantonese-to-Mandarin MT. Our strategy comprises three key components: (1) Syntax and Part-of-Speech (POS) fine-tuning, where we use the Universal Dependencies (UD) corpus to fine-tune LLM, focusing on the linguistic structures of Cantonese; (2) Specialized Cantonese to Mandarin sentence pairs, collected from diverse sources such as Cantonese grammar textbooks and manually translated sentences across various domains, to expose the model to a wide range of linguistic contexts; (3) Post-processing with additional LLMs, where we introduce additional LLMs to improve the initial translations, correcting Mandarin grammar and punctuation. Empirical evaluations on human-created test sets show that our proposed strategy improves translation performance and outperforms existing commercial translation models with at least 3 BLEU scores. Additionally, our strategy also benefits other LLMs and a reversed translation direction, demonstrating its generalization and effectiveness.

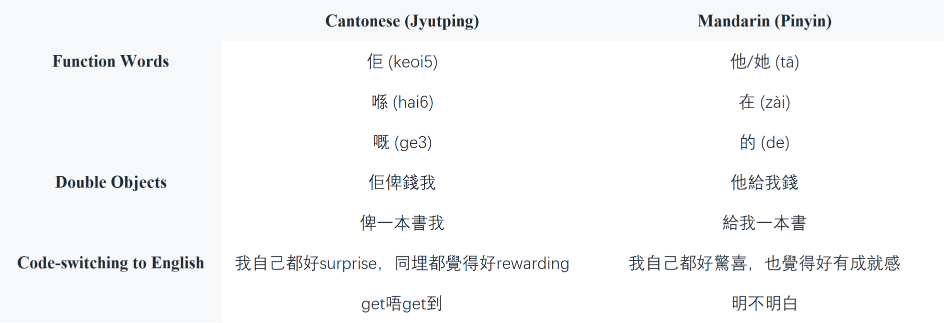
1 Introduction

The rapid advancement of Large Language Models (LLMs) has impacted various Natural Language Processing (NLP) tasks, including Machine Translation (MT), where LLMs leverage extensive pretraining to capture a wide range of linguistic patterns and contextual information to improve translation quality (Feng et al., 2024; Enis and Hopkins, 2024). Cantonese, a major Chinese dialect spoken primarily in Hong Kong, Macau, and parts of southern China, has unique linguistic characteristics that differ from standard Mandarin (Matthews and Yip, 2013).

As Figure 1 shows, lexical divergence represents a cardinal disparity between Cantonese and Mandarin, revealing not solely in lexical choice but also encompassing frequent usage, notably function words. At the syntactic level, although both dialects exhibit a broad alignment, they diverge markedly in specific facets. An example is the inversion of double object ordering, wherein Mandarin adheres to a [human object] + [thing object] configuration, while Cantonese reverses this to a [thing object] + [human object] construct (Snow, 2004; Matthews and Yip, 2013). Code-switching to English is a common phenomenon in Cantonese, which is not typically observed in standard Mandarin. However, the English used in these code-switching instances often deviates from formal English grammar and instead follows Cantonese grammatical structures in most cases (Li, 2000). This unique form of code-switching often results in a blend of Cantonese syntax and English vocabulary. These differences, including distinct phonology, syntax, and vocabulary, pose challenges for existing LLMs in MT scenarios, which are often trained on Mandarin-centric datasets and lack the necessary linguistic knowledge to handle Cantonese effectively (Jiang et al., 2024; Wen-Yi et al., 2024; Hong et al., 2024a).

To address these challenges, we propose a novel strategy to improve the performance of LLMs in Cantonese-to-Mandarin MT. Our strategy incorporates syntax and Parts-of-Speech (POS) fine-tuning, specialized Cantonese-to-Mandarin sentence pairs, and post-processing with additional LLMs within the translation pipeline. Experimental results show that our strategy significantly improves LLM performance in Cantonese-to-Mandarin translation

Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025), pages 427–436
January 20, 2025. ©2025 Association for Computational Linguistics



	Cantonese (Jyutping)	Mandarin (Pinyin)
Function Words	佢 (keoi5)	他/她 (tā)
	喺 (hai6)	在 (zài)
	嘅 (ge3)	的 (de)
Double Objects	佢俾錢我	他給我錢
	俾一本書我	給我一本書
Code-switching to English	我自己都好surprise，同埋都覺得好rewarding	我自己都好驚喜，也覺得好有成就感
	get唔get到	明不明白

Figure 1: Examples of lexical, syntax and English code-switching differences between Cantonese and Mandarin.

tasks, yielding higher BLEU scores on a human-created test set. Moreover, other LLMs also benefit from our proposed strategy to improve translation performance and are better able to handle reverse translation from Mandarin to Cantonese. Our main contributions are as follows.

- We propose a fine-tuning strategy for LLMs that enhances Cantonese-to-Mandarin MT task. This includes syntax and POS prediction, reordering, and random masking. We also compile a diverse data set from various sources[1] and introduce a post-processing framework using additional LLMs for better grammar and punctuation correction.

- Our proposed strategy significantly improves the performance of Yi-1.5-34B in Cantonese-to-Mandarin translation. Experimental results reveal that across five domain-specific gold test sets, BLEU scores improved by at least 3 points. Additionally, our strategy is applicable to LLMs of various sizes and types. Most models show an average BLEU score increase of 3 points, with smaller models displaying even more significant performance gains.

- Our strategy extends beyond Cantonese-to-Mandarin MT, it is equally effective for Mandarin-to-Cantonese translation. It highlights the flexibility of our strategy and the effective capture of linguistic knowledge for low-resource language.

2 Related Work

Existing studies on Cantonese-to-Mandarin MT primarily focus on translating from Mandarin to Cantonese. Unlike widely studied language pairs such as English-to-Mandarin, Cantonese is a low-resource language, and large-scale, high-quality parallel data for Cantonese is limited. The scarcity has prompted the exploration of diverse methods for corpus construction and translation improvement. (Liu, 2022) conduct parallel sentence mining to generate a substantial number of sentence pairs, significantly improving translation quality. Additionally, (Dare et al., 2023) compare different model architectures, tokenization schemes, and embedding structures to investigate the linguistic differences between Mandarin and Cantonese. (Zhang et al., 2022) propose a non-autoregressive MT model for Mandarin to Cantonese translation where it improves the intelligibility and naturalness of synthesised speech.

In recent years, traditional neural MT models have increasingly turned to LLMs to handle Cantonese sentences. (Hong et al., 2024b) perform a CANTONMT pipeline using LLMs to process Cantonese sentences and fine-tuning translation targets. (Jiang et al., 2024) discuss LLM's factual generation, mathematical logic, complex reasoning, and general knowledge in Cantonese and translation scenarios. (Guo et al., 2024) propose a strategy called TALEN, where it shows how to translate source sentence to target sentence via Cantonese syntax patterns. These studies highlight the growing importance of LLMs in advancing the state-of-the-art in low-resource language MT, particularly for Cantonese.

However, most studies have concentrated on Mandarin-to-Cantonese translation, leaving the Cantonese-to-Mandarin direction underexplored. This directional bias is likely driven by practical needs, such as converting standard Mandarin text into Cantonese for use in regions where Cantonese is spoken, including Hong Kong and Guangdong Province in China. Despite the progress made in Mandarin-to-Cantonese translation, there are

[1] Our dataset can be found at https://huggingface.co/datasets/HKAllen/cantonese-chinese-parallel-corpus

still gaps in the performance and discussion of Cantonese-to-Mandarin translation that need to be further investigated.

3 Methodology

In this section, we provide detailed descriptions of each module in our Cantonese-to-Mandarin translation strategy. Figure 2 shows the overall pipeline of the proposed translation strategy.

3.1 Syntax and POS Fine-Tuning

Syntax and POS fine-tuning comprises three prediction tasks designed to fine-tune LLM in one fine-tuning step, as shown in Figure 3.

First, we use 1,004 sentences from the PUD Cantonese corpus[2] (Wong et al., 2017), which are annotated with gold-standard syntax and POS tags. From these, 30% of the annotated Cantonese sentences are randomly selected for the single-word syntax and POS prediction task. One or more Cantonese words are randomly selected in each Cantonese sentence, and the LLM is required to predict the syntactic role or POS tag of these selected words. The purpose of this task is to enable the LLM to initially learn precise Cantonese syntax and POS tagging knowledge via annotated sentences, which differs from the unsupervised learning during pre-training and is more specific and accurate.

Second, another 30% of the annotated Cantonese sentences from the PUD Cantonese corpus are used for the syntax and POS reordering task. Given a gold annotated Cantonese sentence, the syntactic structure and POS tags of this sentence are completely shuffled (each Cantonese word contains its unique tag) and provided as input to the LLM. The LLM must reorder these tags based on the content of the input Cantonese sentence and the shuffled syntactic and POS information, ensuring they conform to the correct grammatical and lexical structure of the sentence. The goal of this task is to enhance the LLM's ability to understand the order of syntactic and POS tags in input Cantonese sentences, handling the complex linguistic structures and varied contexts of Cantonese.

Finally, 40% of the annotated Cantonese sentences from the PUD Cantonese corpus are used for the randomly masking POS to predict syntax task. Certain POS tags are randomly masked, and the LLM is required to infer the masked POS tags using the known POS information and further predict the syntactic roles of the corresponding words. This is to strengthen the LLM's ability to integrate lexical and syntactic knowledge, improving its reasoning capabilities when dealing with incomplete or partial information.

3.2 Specialized Cantonese-Mandarin Sentence Pairs

Given that Cantonese is a low-resource language and existing open-source Cantonese-Mandarin parallel corpora are extremely limited, some of these corpora even involve machine-translating Cantonese to Mandarin for MT training sets[3][4][5][6][7][8][9]. We have undertaken additional efforts to collect and expand the available Cantonese-Mandarin parallel corpora. Specifically, we select and collect a substantial number of Cantonese situational dialogues and their corresponding Mandarin translations from Cantonese language textbooks and websites. These dialogues cover a wide range of domains, providing rich contextual information. Additionally, we compile a list of Cantonese-Mandarin item correspondence vocabulary and collect Cantonese sentences from multiple domains, which are then translated into Mandarin manually. We integrate these newly collected Cantonese-Mandarin sentence pairs into a new dataset and combined it with existing open-source Cantonese-Mandarin parallel corpora to form a more comprehensive and diverse resource as a training set, as shown in Table 1.

3.3 Post-Processing with Additional LLMs

To further optimize the translation results, we have additionally trained two specialized LLMs for post-processing the initial translations. One LLM is designed to correct potential language errors in Mandarin sentences, while the other focuses on correcting punctuation errors, as shown in Figure 4. The output Mandarin translation first passes

[2] https://universaldependencies.org/treebanks/yue_hk/index.html

[3] opus.nlpl.eu/results/yue&cmn/corpus-result-table

[4] opus.nlpl.eu/wikimedia/yue&zh/v20230407/wikimedia

[5] https://github.com/kiking0501/Cantonese-Chinese-Translation

[6] https://github.com/meganndare/cantonese-nlp?tab=readme-ov-file

[7] https://opus.nlpl.eu/TED2020/zh&zh_cn/v1/TED2020

[8] https://huggingface.co/datasets/botisan-ai/cantonese-mandarin-translations

[9] https://huggingface.co/datasets/raptorkwok/cantonese-traditional-chinese-parallel-corpus

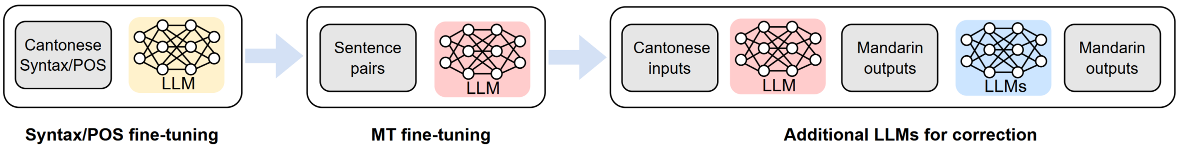


Figure 2: Overall pipeline of the proposed translation strategy, where the LLM undergoes syntax and POS fine-tuning, followed by MT fine-tuning, and additional LLMs improve and correct the initial Mandarin outputs.

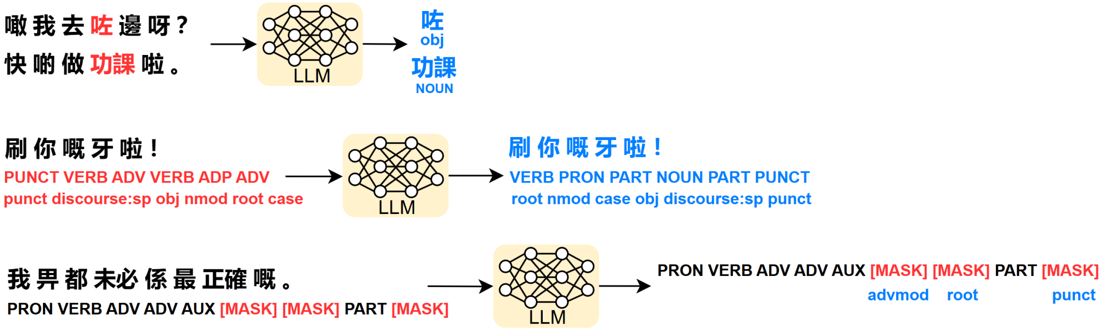

噉我去咗邊呀？
快啲做功課啦。
LLM
咗
obj
功課
NOUN
刷你嘅牙啦！
PUNCT VERB ADV VERB ADP ADV
punct discourse:sp obj nmod root case
LLM
刷你嘅牙啦！
VERB PRON PART NOUN PART PUNCT
root nmod case obj discourse:sp punct
我畀都未必係最正確嘅。
PRON VERB ADV ADV AUX [MASK] [MASK] PART [MASK]
LLM
PRON VERB ADV ADV AUX [MASK] [MASK] PART [MASK]
advmod root punct

Figure 3: Syntax and POS fine-Tuning on LLM includes three tasks: single word syntax and POS prediction, syntax and POS reordering, and randomly masking POS to predict syntax.

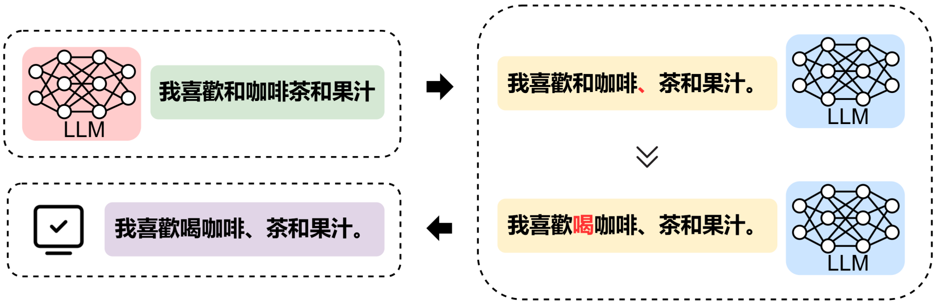


Figure 4: Post-processing with additional LLMs. In the Mandarin output sentence "I like and coffee tea and juice", some errors are present. First, LLMs detect and correct the punctuation: "I like and coffee, tea and juice." Subsequently, the sentence is further corrected: "I like drinking coffee, tea and juice." Due to the linguistic differences between Mandarin and English, this example may not fully capture the intended meaning of the original Mandarin sentence above.

Source	Number of Sentences
Open-source Data Set	200,843
Human Collection	27,433
Human Translation	78,474

Table 1: In addition to the existing open-source datasets, Cantonese-Mandarin parallel sentence pairs also come from two other sources: Human Collection, which involves manually collecting additional sentence pairs, and Human Translation, which involves manually translating Cantonese sentences

.

through an LLM that specializes in correcting punctuation errors in Mandarin sentences. This model is trained to identify and correct various punctuation errors, such as missing commas, incorrect periods, and misplaced quotation marks. By focusing on punctuation, this LLM ensures that the translated text adheres to the standard conventions of written Mandarin, enhancing its readability and clarity. Following this, the text is processed by the second LLM, which is designed to correct potential language errors in the Mandarin sentences. This model can identify and correct issues such as grammatical mistakes, lexical errors, and syntactic irregularities, ensuring that the translated text conforms to the grammatical rules and conventions of standard Mandarin.

4 What Happens to Translation Performance?

We demonstrate the effectiveness of our proposed strategy for LLM in Cantonese-to-Mandarin MT task. We also conduct ablation experiments to highlight the impact of specific components of our strategy on the model's performance.

4.1 Experiment Settings

We evaluate the effectiveness of the proposed strategy using the BLEU score. The training set consists of 303,682 sentence pairs, while the validation set contains 3,068 sentence pairs. We collect Cantonese sentences from social media and the Cantonese Wikipedia and have them manually translated into Mandarin to form our gold test set. The test set is divided into five categories: Conversation (Conv), Finance (Fin), History (Hist), Technology (Tech), and Biology (Bio) to provide a more detailed evaluation of the model's performance improvement. The conversation category contains 1,000 gold-standard Cantonese-Mandarin translations, while each of the other categories includes 200 sentence pairs. We also incorporated commonly used commercial translation engines (Google Translate[10], Microsoft Translator[11], and Baidu Translator[12]) for comparison to validate the effectiveness of our proposed strategy.

In the experiments, we fine-tune Yi-1.5-34B[13] (AI et al., 2024) using instruction tuning and Low-Rank Adaptation (LoRA) (Hu et al., 2021) with the following parameters: rank of the low-rank decomposition = 4, scaling factor for LoRA = 8, learning rate = 0.0005, training epochs = 2, optimizer = AdamW, quantization bit = 4, and per GPU training batch size = 6 for the first step of syntax/POS fine-tuning. For the MT fine-tuning step, we increase the rank of the low-rank decomposition to 8, training epochs = 2 and the scaling factor for LoRA to 16. Additionally, we use the THUCTC news dataset[14] to fine-tune Yi-1.5-9B[15] for punctuation correction. We manually remove or disorder punctuation in Mandarin sentences to serve as inputs, and the LLM detects and corrects these errors to produce correctly punctuated sentences. We also employ the NLPCC2023[16], MuCGEC[17], and Pycorrector[18] datasets to fine-tune the same type of LLM for addressing Mandarin grammatical errors. These LLMs are trained with LoRA where rank of the low-rank decomposition = 8, scaling factor for LoRA = 16, learning rate = 0.0005, training epochs = 3, AdamW optimizer, and a per GPU training batch size = 16. The experiments are conducted on 8 NVIDIA A100 40GB GPUs and 16 NVIDIA V100 32GB GPUs.

4.2 Results

As shown in Table 2, commercial MT engines display varying levels of effectiveness in translating Cantonese to Mandarin, with Microsoft Bing generally surpassing other commercial engines. However, Yi-1.5-34B-baseline model, fine-tuned using collected open-source dataset, demonstrates better translation performance compared to the commercial engines. When applied to our specialized Cantonese-Mandarin sentence pairs (Yi-1.5-34B-v1), we observe further improvement across all domains, where each domain increases at least 1 BLEU score. It confirms our specialized Cantonese-Mandarin sentence pairs is in enhancing translation quality, emphasizing the advantage of incorporating multi-domain training set to achieve higher accuracy in translations.

Despite utilizing only 1,004 Cantonese sentences for syntax/POS fine-tuning, the translation performance of Yi-1.5-34B-v2 shows a significant improvement, leading to at least a 2-point increase in BLEU scores compared to Yi-1.5-34B-baseline. This suggests that fine-tuning the grammatical structures of Cantonese sentences beforehand can provide a strong foundation for subsequent MT task fine-tuning. Targeted enhancements of specific linguistic knowledge in LLMs may yield better improvements than simply increasing the training set size. The difference between Yi-1.5-34B-v3 and Yi-1.5-34B-v4 lies in the use of additional LLMs for correcting Mandarin outputs. While Yi-1.5-34B-v4 achieves the highest BLEU scores, the improvement over Yi-1.5-34B-v3 is marginal. This is likely because the initial Cantonese fine-tuning with syntax/POS addressed most grammatical and structural corrections, leaving little room for further enhancement by post-processing.

[10] https://translate.google.com

[11] https://www.bing.com/translator

[12] https://fanyi.baidu.com/mtpe-individual/multimodal

[13] https://github.com/01-ai/Yi-1.5

[14] http://thuctc.thunlp.org/

[15] https://huggingface.co/01-ai/Yi-1.5-9B

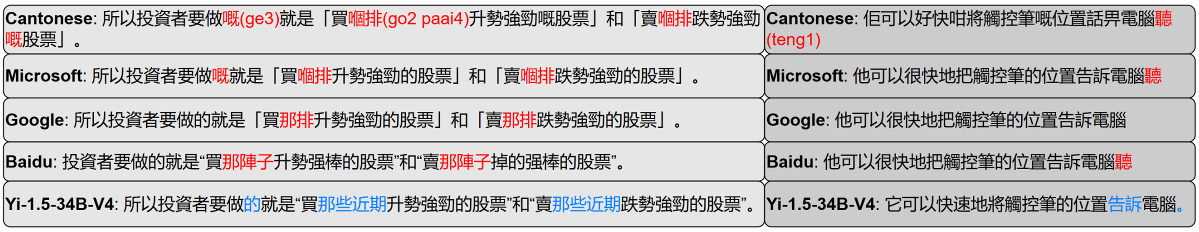
[16] http://tcci.ccf.org.cn/conference/2023/taskdata.php

[17] https://github.com/HillZhang1999/MuCGEC

[18] https://github.com/shibing624/pycorrector

	SSP	Syntax/POS	LLMs	Domains (Cantonese→Mandarin)				
				Conv	Fin	Hist	Tech	Bio
Baidu Translator				46.872	43.619	62.447	65.635	54.446
Google Translate				43.515	46.366	77.991	72.140	63.479
Microsoft Bing				40.776	47.865	83.769	74.259	63.603
Yi-1.5-34B-baseline	✗	✗	✗	48.367	50.101	83.676	75.713	67.107
Yi-1.5-34B-v1	✓	✗	✗	50.524	51.865	85.092	77.456	68.606
Yi-1.5-34B-v2	✗	✓	✗	51.317	52.590	85.723	79.079	70.649
Yi-1.5-34B-v3	✓	✓	✗	52.001	53.413	87.116	81.025	71.350
Yi-1.5-34B-v4	✓	✓	✓	**52.125**	**53.686**	**87.400**	**81.304**	**72.021**

Table 2: Cantonese-to-Mandarin MT BLEU scores on different domains. SSP denotes specialized Cantonese-Mandarin sentence pairs collected by us. If SSP is not checked, the model uses the original open-source dataset. Syntax/POS refers to syntax and POS fine-tuning, LLMs represents post-processing with additional LLMs.



Cantonese: 所以投資者要做嘅(ge3)就是「買嗰排(go2 paai4)升勢強勁嘅股票」和「賣嗰排跌勢強勁嘅股票」。

Microsoft: 所以投資者要做嘅就是「買嗰排升勢強勁的股票」和「賣嗰排跌勢強勁的股票」。

Google: 所以投資者要做的就是「買那排升勢強勁的股票」和「賣那排跌勢強勁的股票」。

Baidu: 投資者要做的就是“買那陣子升勢强棒的股票”和“賣那陣子掉的强棒的股票”。

Yi-1.5-34B-V4: 所以投資者要做的就是“買那些近期升勢強勁的股票”和“賣那些近期跌勢強勁的股票”。

Cantonese: 佢可以好快咁將觸控筆嘅位置話畀電腦聽(teng1)

Microsoft: 他可以很快地把觸控筆的位置告訴電腦聽

Google: 他可以很快地把觸控筆的位置告訴電腦

Baidu: 他可以很快地把觸控筆的位置告訴電腦聽

Yi-1.5-34B-V4: 它可以快速地將觸控筆的位置告訴電腦。

Figure 5: Translations of the given Cantonese sentence from different translation engines. Red markings indicates key Cantonese components and incorrect translation results from commercial translation engines, while blue represents our translations.

As illustrated in Figure 5, the left example demonstrates that the Microsoft translation fails to correctly identify the Cantonese particles "ge3" (indicating possession or modification) and "go2 paai4" (indicating a specific period or phase). Both Google and Baidu Translate interpret "go2 paai4" as a general time reference, while only our model accurately translates these Cantonese particles. In the right example, the Cantonese sentence positions the verb "teng1" (to hear) at the end as a complement, emphasizing that the information has been conveyed or understood by the recipient. Microsoft and Baidu translations adopt a direct approach (thing + verb), which is not grammatically appropriate in Mandarin (verb + thing). Only Google Translate and our model considered the grammatical structure of Mandarin and translated correctly, with our model additionally incorporating punctuation into the translation.

5 What Happens to Other Models?

The experiments presented above demonstrate the effectiveness of our proposed strategy in improving Cantonese-to-Mandarin MT performance. However, a question remains: does this strategy also enhance the translation performance when applied to other LLMs? To investigate this, we conduct additional experiments using the latest state-of-the-art LLMs and traditional MT models to evaluate the generalizability of our strategy and its potential impact on the field of MT. These experiments involve fine-tuning multiple LLMs with the same LoRA configurations and datasets, comparing their performance on the same tasks.

5.1 Experiment Settings

We employ another LLM with the same parameter size, Qwen-2.5-32B[19] (Team, 2024), alongside Yi-1.5-9B, GLM-4-9B [20] (GLM et al., 2024), and the smaller MiniCPM3-4B[21] (Hu et al., 2024). The training settings and datasets used for these models are consistent with those utilized for Yi-1.5-34B on Cantonese-to-Mandarin MT task, where all models undergo the same syntax/POS fine-tuning process, the same training set and additional LLMs for post-processing to ensure a fair comparison. Only Qwen-2.5-34B uses 4 bits quantization.

Traditional MT models such as NLLB-1.3B (Costa-jussà et al., 2022), M2M100-1.2B (Fan et al., 2021), and mRASP (Lin et al., 2020) adopt different fine-tuning settings and do not support syntax/POS fine-tuning due to their different working mechanisms. These models use a learning rate

[19] https://huggingface.co/Qwen/Qwen2.5-32B
[20] https://huggingface.co/THUDM/glm-4-9b
[21] https://huggingface.co/openbmb/MiniCPM3-4B

= 0.0005, batch size = 32, and the optimizer = Adam. For the Adam optimizer, β_1 is set at 0.9 and β_2 at 0.999, with a weight decay of 0.01. Furthermore, all models utilize an early stopping strategy during the fine-tuning process to prevent overfitting. All experiments are conducted on 8 NVIDIA A100 40GB GPUs and 16 NVIDIA V100 32GB GPUs.

5.2 Results

Table 3 demonstrates that the proposed strategy provides significant benefits for LLMs of different scales, although the extent of BLEU score improvements varies. Overall, the Average Improvement Score (AvgIS) for models after applying our strategy increases by at least 3 points. Qwen-2.5-32B model achieved the highest BLEU scores across all domains. In contrast, GLM-4-9B and Yi-1.5-9B models also show significant improvements, particularly in the Conv domain. The smaller-scale MiniCPM3-4B model shows an improvement with an AvgIS of 3.733 points after applying the strategy, with this improvement in BLEU score surpassing that of the 32B and 9B LLMs. Smaller-scale models exhibit more substantial performance gains after fine-tuning, while larger-scale models achieve further improvements in absolute performance levels. This indicates that larger models, with their greater number of parameters, can capture more complex and general language features and structures, thus performing better in low-resource translation tasks.

Traditional MT models, such as NLLB-1.3B, M2M100-1.2B, and mRASP, show significant performance improvements even without syntax/POS fine-tuning strategies. Notably, the AvgIS of M2M100-1.2B and mRASP are substantially higher than those of other LLMs, with values of 5.941 and 4.576, respectively. However, their BLEU scores in each domain still fall short of those achieved by the latest LLMs, further confirming the advantage of LLMs in low-resource language translation tasks. The parameter size of the 32B LLM is over three times that of the 9B LLM, leading to significant differences in hardware requirements. Yet, the BLEU scores across various domains do not exhibit a threefold difference. This suggests that while increasing the parameter size can improve model performance, the marginal gains in translation quality diminish as the model size exceeds a certain point. Therefore, focusing solely on increasing model size may not be the most effective approach to achieving significant improvements in MT tasks, especially in low-resource scenarios.

6 What Happens to a Reversed Direction?

To further validate the robustness of our proposed strategy, we extend our experiments to the reverse translation direction, from Mandarin to Cantonese. It allows us to examine whether the improvements observed in Cantonese-to-Mandarin translation are specific to that direction or generalize to the opposite direction as well. We aim to establish its broader applicability and reliability in real-world MT scenarios. This comprehensive validation not only enhances the credibility of our strategy but also contributes significantly to the broader field of MT, especially for low resource languages like Cantonese.

6.1 Experiment Settings

We continue to use Baidu Translate, Google Translate, and Microsoft Bing as reference for current commercial MT performance. Beyond the Yi-1.5-34B model, we have incorporated additional LLMs, namely GLM4-9B, and MiniCPM3-4B. The fine-tuning methods for these LLMs remain consistent with those used in our previous experiments. We reverse the translation direction of the dataset, while the number and division of sentence pairs in the training and test sets remain unchanged, with the source language now being Mandarin and the target language being Cantonese. All experiments are conducted on 8 NVIDIA A100 40GB GPUs and 16 NVIDIA V100 32GB GPUs.

6.2 Results

According to Table 4, when the translation direction is switched to Mandarin-to-Cantonese, Google Translate shows the best overall performance among the three commercial translation engines. In contrast, the other two engines experience significant declines, with BLEU scores in most domains dropping by at least 5 points. This highlights substantial differences in model adaptability across different language pairs in commercial translation engines.

Similar issues are observed in LLMs, where simply fine-tuning via collected open-source datasets as training sets does not effectively improve their translation performance. Neither small nor large parameter LLMs can surpass that of Google Translate. For instance, without our strategy, Yi-1.5-34B only achieves a BLEU score of 75.941 in the Hist domain, which is lower than that of Google Translate and fails to demonstrate the advantages of its

Models	Strategy			BLEU Scores (Cantonese→Mandarin)					AvgIS
	SSP	Syntax/POS	LLMs	Conv	Fin	Hist	Tech	Bio	
Qwen-2.5-32B	✗	✗	✗	48.038	51.224	84.130	77.872	69.088	-
Qwen-2.5-32B-ours	✓	✓	✓	52.169	54.838	86.440	80.603	71.660	+3.071
GLM-4-9B	✗	✗	✗	45.885	50.477	82.167	74.922	67.485	-
GLM-4-9B-ours	✓	✓	✓	51.966	52.616	84.839	79.556	69.533	+3.514
Yi-1.5-9B	✗	✗	✗	48.733	49.351	83.286	74.840	66.758	-
Yi-1.5-9B-ours	✓	✓	✓	51.720	52.304	85.236	79.538	70.269	+3.279
MiniCPM3-4B	✗	✗	✗	46.960	50.883	81.050	72.178	65.889	-
MiniCPM3-4B-ours	✓	✓	✓	51.219	52.569	84.435	79.317	68.085	+3.733
NLLB-1.3B	✗	✗	✗	39.498	40.930	59.407	64.108	54.944	-
NLLB-1.3B-improved	✓	✗	✓	43.789	44.234	61.456	67.789	56.082	+2.986
M2M100-1.2B	✗	✗	✗	42.480	45.541	58.194	65.173	55.924	-
M2M100-1.2B-improved	✓	✗	✓	45.725	51.628	65.969	72.604	61.093	+5.941
mRASP	✗	✗	✗	37.281	38.568	46.643	55.180	50.301	-
mRASP-improved	✓	✗	✓	40.463	42.675	54.322	59.036	54.361	+4.576

Table 3: The BLEU scores of different LLMs and traditional translation models after applying our strategy across different domains. SSP denotes specialized Cantonese-Mandarin sentence pairs collected by us. If SSP is not selected, the model utilizes the original open-source dataset. NLLB, M2M100, and mRASP are not suitable for syntax/POS fine-tuning, as they do not follow the same working mechanism as LLMs.

Models	Strategy	BLEU Scores (Mandarin→Cantonese)				
		Conv	Fin	Hist	Tech	Bio
Baidu Translator	-	45.095	39.112	63.254	55.160	56.557
Google Translate	-	43.017	53.051	77.441	72.251	70.969
Microsoft Bing	-	44.332	36.530	65.560	64.774	58.298
Yi-1.5-34B	✓	45.956	54.623	78.401	73.607	72.285
	✗	43.198	51.029	75.941	71.693	69.707
GLM-4-9B	✓	45.614	49.364	78.582	69.731	69.022
	✗	44.248	48.952	77.285	67.587	68.187
MiniCPM3-4B	✓	42.598	51.656	73.330	69.613	57.512
	✗	41.500	50.123	71.538	67.154	56.348

Table 4: BLEU scores of various models across domains in Chinese-to-Cantonese translation, where ✗ denotes training on open-source datasets without employing our specific strategy and ✓ indicates the application of our strategy by the model.

34B parameter size. But after applying our proposed strategy, Yi-1.5-34B's BLEU scores in all domains surpass those of commercial translation engines, with each domain seeing an increase of approximately 2 BLEU scores. Similarly, GLM-4-9B and MiniCPM3-4B exhibited comparable results, with BLEU scores in each domain improving by at least 1 point. This suggests that although larger model parameters are beneficial for low-resource translation, directly fine-tuning LLMs with parallel corpus datasets may fail to fully develop their potential. While Cantonese has been the source language in previous experiments, the benefits of LLMs acquiring its linguistic knowledge can also extend to scenarios where Cantonese is the target language then. Additionally, the BLEU scores of LLMs in all domains do not increase as much as when Cantonese is the source language, indicating that the proposed strategy or translation direction may still be constrained by directionality effects in MT scenarios.

7 Conclusion

In this paper, we present a strategy to improve translation performance in low-resource language MT scenarios, focusing on Cantonese-to-Mandarin translation. Our approach enables Yi-1.5-34B to better understand Cantonese sentence structures through syntax/POS fine-tuning. By leveraging a custom-compiled dataset and additional LLMs for post-processing, we significantly improve Cantonese-to-Mandarin translation performance, with BLEU scores increasing by at least 5 points compared to current commercial MT engines. This strategy is effective not only for Yi-1.5-34B but also for other LLMs, particularly smaller parameter models. Furthermore, our experiments show that LLMs continue to benefit from this strategy in the reverse translation direction, achieving higher BLEU scores than commercial MT engines and baseline versions of LLMs.

8 Limitations

Due to time and GPU resource constraints, we adopt a more resource-friendly approach using LoRA for LLM fine-tuning, where full parameters fine-tuning has not been confirmed and discussed. Additionally, the BLEU score has some limitations as it primarily measures n-gram overlap and may not fully capture the fluency, coherence, and accuracy of the translations. Future work can explore the performance of the LLM with full parameter fine-tuning and additional evaluation metrics, such as METEOR, ROUGE, or human evaluations, to provide a more comprehensive evaluation of the model's performance.

9 Acknowledgments

We would like to acknowledge the valuable support provided by the Logistics and Supply Chain MultiTech R&D Centre (LSCM), which is essential in making this paper possible. The research is not only supported by the LSCM but also funded by the Innovation and Technology Fund of the Hong Kong Special Administrative Region under project ITP/051/23LP.

References

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

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

Maxim Enis and Mark Hopkins. 2024. From llm to nmt: Advancing low-resource machine translation with claude. *Preprint*, arXiv:2404.13813.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. Tear: Improving llm-based machine translation with systematic self-refinement. *Preprint*, arXiv:2402.16379.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Ping Guo, Yubing Ren, Yue Hu, Yunpeng Li, Jiarui Zhang, Xingsheng Zhang, and Heyan Huang. 2024. Teaching large language models to translate on low-resource languages with textbook prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15685–15697, Torino, Italia. ELRA and ICCL.

Kung Hong, Lifeng Han, Riza Batista-Navarro, and Goran Nenadic. 2024a. CantonMT: Cantonese to English NMT platform with fine-tuned models using real and synthetic back-translation data. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 590–599, Sheffield, UK. European Association for Machine Translation (EAMT).

Kung Yin Hong, Lifeng Han, Riza Batista-Navarro, and Goran Nenadic. 2024b. 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: Presentations)*, pages 133–144, Chicago, USA. Association for Machine Translation in the Americas.

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

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

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 preprint arXiv:2408.16756*.

David CS Li. 2000. Cantonese-english code-switching research in hong kong: A y2k review. *World Englishes*, 19(3):305–322.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2649–2663, Online. Association for Computational Linguistics.

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

Stephen Matthews and Virginia Yip. 2013. *Cantonese: A comprehensive grammar*. Routledge.

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

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

Andrea W Wen-Yi, Unso Eun Seo Jo, Lu Jia Lin, and David Mimno. 2024. How chinese are chinese language models? the puzzling lack of language policy in china's llms. *arXiv preprint arXiv:2407.09652*.

Tak-sum Wong, Kim Gerdes, Herman Leung, and John SY Lee. 2017. Quantitative comparative syntax on the cantonese-mandarin parallel dependency treebank. In *Proceedings of the fourth international conference on Dependency Linguistics (Depling 2017)*, pages 266–275.

Junhui Zhang, Wudi Bao, Junjie Pan, Xiang Yin, and Zejun Ma. 2022. A novel chinese dialect tts frontend with non-autoregressive neural machine translation. *Preprint*, arXiv:2206.04922.