How Large Language Models Enhance Low-Resource Mongolian-Chinese Machine Translation?

Zhenjie Gao[a,b,c], Feilong Bao[†a,b,c], Yuan Li[a,b,c], Ruichen Hou[a,b,c], Yibo Han[a,b,c]

[a]*College of Computer Science, Inner Mongolia University*
[b]*National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian*
[c]*Inner Mongolia Key Laboratory of Multilingual Artificial Intelligence Technology*

Abstract

The rapid development of Inner Mongolia has led to a growing demand for Mongolian-Chinese translation. However, Mongolian presents significant challenges for machine translation, such as data scarcity and its complex syntactic structures. Consequently, traditional Mongolian-Chinese machine translation methods often produce outputs with poor fluency and suffer from the loss of critical semantic information. In this paper, we propose a Mongolian-Chinese machine translation method based on large language models (LLM-CPT-SymFT). Specifically, LLM-CPT-SymFT involves continual pre-training (CPT) of the base model using Mongolian and Chinese corpora, followed by symmetrical fine-tuning (SymFT), which leverages a mix of original Mongolian-Chinese parallel data and its synthetically reversed counterpart, primarily to enhance Mongolian-to-Chinese translation performance. We evaluated our method on two datasets, achieving an average BLEU score improvement of 28.89 compared to baseline models. Results confirm our method's strong potential for improving Mongolian-Chinese machine translation.

Keywords: Large Language Model;Low-Resource;Mongolian-Chinese Translation;Continual Pre-training;Symmetrical Fine-tuning

1. Introduction

Accelerated economic and cultural expansion in Inner Mongolia has fostered increasingly frequent interactions with other domestic regions. Given these interactions and the linguistic differences between Mongolian (one of Inner Mongolia's official languages) and Chinese, ensuring effective cross-linguistic communication through high-quality Mongolian-Chinese translation has become critically important[25]. However, most existing machine translation research focuses on resource-rich language pairs such as English, Chinese, and French[10], where models like mT5[23] and NLLB[4] have demonstrated remarkable performance. These models have not been specifically optimized for the Mongolian-Chinese language pair, meaning its specific linguistic challenges and translation demands are largely unaddressed by these advanced approaches.

To bridge this gap, Zhi and Wang [30] proposed a data augmentation method based on BERT semantic similarity. This method translates Mongolian sentences from the Mongolian-Chinese

[†] Corresponding author: Feilong Bao (Email: csfeilong@imu.edu.cn)

dataset into Chinese, calculates the semantic similarity between these translations and the original Chinese sentences, and then selects high-similarity pairs to expand the dataset. Although this method improved performance to some extent, the synthesized data highly overlapped semantically with the original dataset, which limited the model's generalization ability. To address this issue, Yin et al. [26] introduced the CSGAN to generate pseudo-parallel data and combined it with co-training for optimization. However, CSGAN exhibited significant instability during training, and the pseudo-parallel data contained noise, negatively impacting translation quality. To further address these challenges, Zhang et al. [28] proposed a method that combined CNN and DESCAT. This approach leveraged phrase structure and dependency structure information to improve performance. However, this approach suffers from drawbacks such as the difficulty of Mongolian syntactic analysis and inadequate handling of Out-Of-Vocabulary (OOV) words.

In this paper, we propose a Mongolian-Chinese machine translation method based on large language models (LLM-CPT-SymFT). This method utilizes Gemma2-9B-it [19] as its base model, given that its robust multilingual processing capabilities and efficient tokenization strategy provide a solid foundation for the Mongolian-Chinese machine translation task. It comprises two stages: continual pre-training and symmetrical fine-tuning. First, we continually pre-train the base model using vast Mongolian and Chinese monolingual corpora to deepen its understanding of Mongolian syntactic structures and cultural context, while reinforcing its Chinese comprehension and generation capabilities. Second, we symmetrically fine-tune the pre-trained model using Mongolian-Chinese parallel dataset to enhance the model's Mongolian-Chinese alignment and generation capabilities. To evaluate the effectiveness of our method, we conducted extensive experiments on two datasets, comparing it with three baseline models. The experimental results demonstrate that our method achieves significantly superior performance to the baseline models in terms of the BLEU score. Our contributions are as follows:

- To the best of our knowledge, this is the first work to propose a Mongolian-Chinese machine translation method based on Large Language Models. It provides a novel and effective approach for low-resource translation challenges.
- Our method achieves competitive performance on both datasets, demonstrably outperforming several baseline models. This validates the effectiveness of the proposed approach.

2. Related Work

2.1. Neural Machine Translation

Neural Machine Translation (NMT) refers to modeling the translation from a source language to a target language directly using neural networks[14]. Kalchbrenner and Blunsom [9] proposed the concept of an end-to-end translation model based on neural networks, which laid the foundation for modern NMT. Subsequently, Sutskever et al. [15] introduced the Encoder-Decoder architecture, which involves the encoder compressing a source sentence into a fixed-length vector representation, and the decoder generating the target sentence word by word based on this vector. However, this information compression can lead to information loss. To address this, Bahdanau et al. [2] introduced the attention mechanism, which allows the decoder to dynamically focus on relevant parts of the source sentence when generating each target word. Following this, Vaswani et al. [22] proposed the Transformer, which constructs both the Encoder and Decoder based on self-attention mechanisms. Subsequently, with breakthroughs in large-scale pre-trained language models (PLMs) such as BERT[5] and GPT[13], researchers began to explore integrating these models into NMT tasks. The latest trend involves directly utilizing Large Language Models for translation via prompting[29], without the need for task-specific fine-tuning. Neural Machine Translation has undergone rapid evolution in recent years. These advancements have not only driven improvements in translation quality but have also laid a solid foundation for the future development of machine translation technology.

2.2. Large Language Models

Large Language Models (LLMs) represent a significant development direction in Natural Language Processing. Their prominence surged with the release of OpenAI's ChatGPT[11] at the end of 2022, and they have gradually bifurcated into closed-source and open-source models. In the closed-source domain, OpenAI has released models such as GPT-4 Turbo[1], GPT-4o[7], and GPT-4.5. These models have undergone continuous optimizations in areas such as context window size, multimodal capabilities, and overall performance. Google introduced Gemini series[16]. Gemini 1.0 includes three sizes: Ultra, Pro, and Nano. Gemini 1.5 Pro[17] introduced a 1 million token context window and a Mixture-of-Experts (MoE) architecture. Gemini 2.0 and 2.5 series offer versions like Flash and Pro, further enhancing inference and tool-use capabilities. In the open-source domain, Meta AI's Llama series[21, 6], aimed at the research community, has spurred the development of numerous community-fine-tuned models. Alibaba developed the Qwen series, encompassing Qwen1.5[3], Qwen2, Qwen2.5[24], and Qwen3, with parameter sizes ranging from 0.5B to 72B, supporting context windows of up to 128k tokens and multiple languages. Additionally, Google launched the Gemma series models[18, 19, 20], based on Gemini technology. This series is characterized by its lightweight and versatile nature. The continuous evolution of these models demonstrates the immense potential of LLMs in the field of Natural Language Processing.

3. Methodology

This section details the LLM-CPT-SymFT. As shown in Figure 1, the method comprises two stages: *continual pre-training* and *symmetrical fine-tuning*. To facilitate understanding, we first introduce the task definition, followed by a detailed description of the specific implementation of each stage.

3.1. Task Definition

Mongolian-Chinese machine translation refers to automatically converting Mongolian text into semantically equivalent and fluently expressed Chinese text. Let the Mongolian-Chinese parallel dataset be denoted as $\mathcal{D}_{trans} = \{(\mathbf{x}^k, \mathbf{y}^k)\}_{k=1}^{N}$, where N represents the total number of parallel sentence pairs. Each sentence pair $(\mathbf{x}^k, \mathbf{y}^k)$ consists of a Mongolian source sentence $\mathbf{x}^k = (x_1, x_2, \ldots, x_m)$ and its corresponding Chinese target sentence $\mathbf{y}^k = (y_1, y_2, \ldots, y_n)$, where x_i represents the i-th token in the source sentence, m is the length of the source sentence, y_j represents the j-th token in the target sentence, and n is the length of the target sentence. Then, the training process of a Mongolian-Chinese machine translation model can be formulated as:

$$\hat{\theta} = \arg\max_{\theta} \sum_{k=1}^{N} \log P(\mathbf{y}^k \mid \mathbf{x}^k; \theta) \tag{1}$$

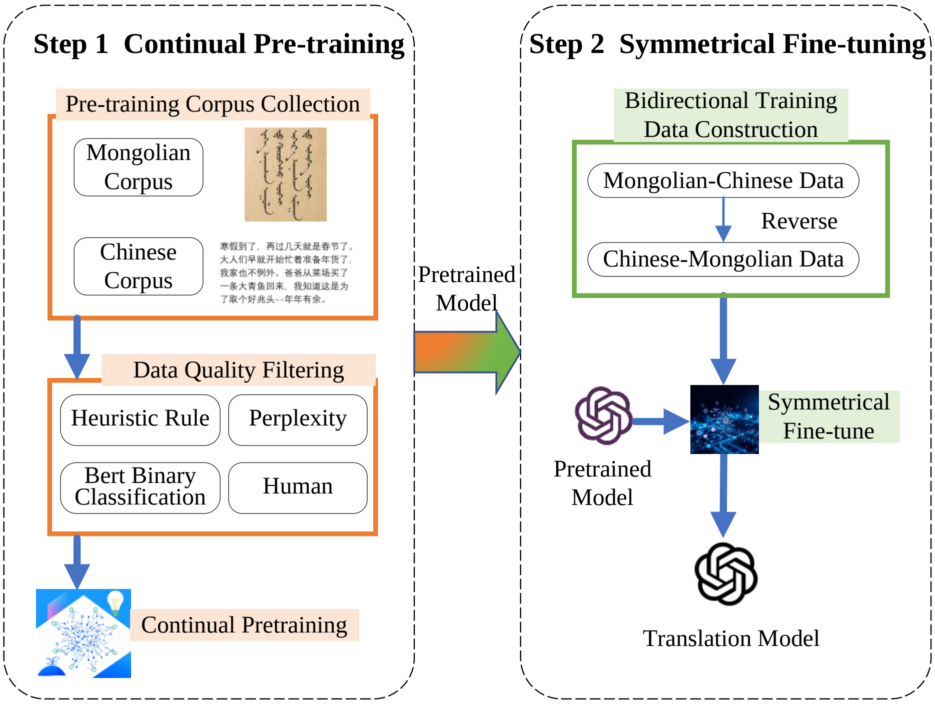

Step 1 Continual Pre-training
Pre-training Corpus Collection
Mongolian Corpus
Chinese Corpus
寒假到了，再过几天就是春节了。
大人们早就开始忙着准备年货了，
我家也不例外。爸爸从菜场买了
一条大青鱼回来，我知道这是为
了取个好兆头--年年有余。
Data Quality Filtering
Heuristic Rule
Perplexity
Bert Binary Classification
Human
Continual Pretraining
Pretrained Model
Step 2 Symmetrical Fine-tuning
Bidirectional Training Data Construction
Mongolian-Chinese Data
Reverse
Chinese-Mongolian Data
Pretrained Model
Symmetrical Fine-tune
Translation Model

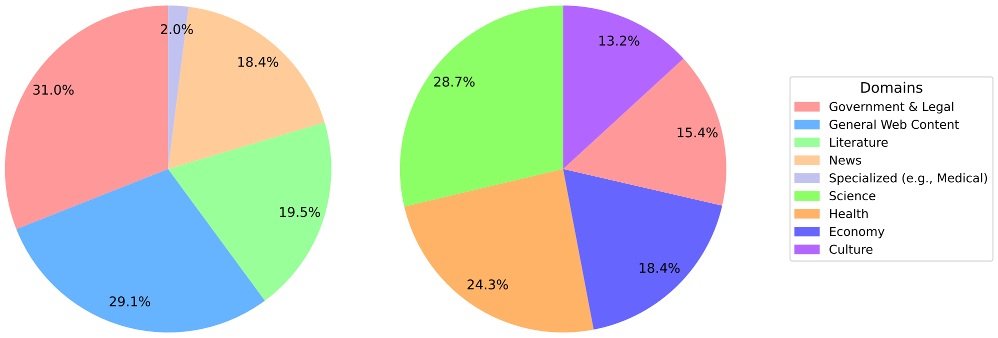
Figure 1: Illustration of the proposed method LLM-CPT-SymFT, detailing the Continual Pre-training (Step 1) and Symmetrical Fine-tuning (Step 2).

where θ represents the model parameters, and $P(\mathbf{y}^k \mid \mathbf{x}^k; \theta)$ denotes the conditional probability of generating the Chinese sentence $\mathbf{y}^k$ given the Mongolian sentence $\mathbf{x}^k$. The objective is to find parameters that maximize this conditional probability across the dataset. This is achieved through Maximum Likelihood Estimation, where the optimization objective is to find the parameters $\hat{\theta}$ that maximize the joint log-likelihood of the dataset. However, due to the low-resource nature of Mongolian, achieving high-quality Mongolian-Chinese machine translation faces numerous challenges. This motivates our exploration of solutions that leverage the capabilities of Large Language Models.

3.2. Continual Pre-training

3.2.1. Continual Pre-training Data

To enable the Large Language Model to deeply understand the linguistic characteristics of Mongolian and to consolidate Chinese processing capabilities, we constructed large-scale Mongolian and Chinese monolingual pre-training corpora. Given the relative scarcity of publicly available Mongolian monolingual corpora, we collected a 3 billion tokens corpus from the internet using web scraping techniques. However, this corpus contained substantial noise, such as HTML tags, non-Mongolian content, and low-quality text segments. Therefore, we designed a multi-stage Mongolian corpus filtering pipeline:



(a) Mongolian Corpus

(b) Chinese Corpus

Figure 2: Domain composition of Mongolian (a) and Chinese (b) corpora for Continual Pre-training.

Heuristic-based Filtering: This stage primarily involves coarse-grained cleaning. First, we used regular expressions to remove HTML tags and other irrelevant web markup from the text, and performed initial text format normalization. Subsequently, we employed fastText[8] to eliminate text segments in Chinese, English, and other non-Mongolian languages mixed within the corpus. Finally, to ensure each text segment contained sufficient contextual information for effective learning, we set a length threshold and removed all overly short sentences.

BERT-based Quality Classification Filtering: To further enhance corpus quality, we built a Mongolian text quality classifier based on BERT[5]. Specifically, we first manually annotated a set of high-quality (e.g., grammatically correct, fluently expressed) and low-quality (e.g., inappropriate word choice, disorganized structure) samples to train the classification model. After training, we applied this classifier to the heuristically filtered Mongolian corpus, discarding sentences predicted as low-quality and retaining the high-quality portion.

Perplexity-based Filtering: To ensure the fluency and naturalness of the corpus, we trained an N-gram language model using the high-quality Mongolian corpus obtained from the previous BERT-based filtering stage. Subsequently, this model was used to evaluate the perplexity score of each sentence in the corpus. Generally, sentences with higher perplexity scores are more likely to have unnatural or incoherent linguistic expressions. Therefore, we set an empirical threshold, discarding sentences with scores above this threshold and retaining those with lower perplexity, which the language model deemed more fluent.

After these filtering stages, we selected a 1 billion tokens high-quality corpus from the initial 3 billion tokens Mongolian corpus. Concurrently, to reinforce the LLM's Chinese capabilities, we extracted 1 billion tokens high-quality Chinese text directly from an existing open-source corpus, covering multiple domains such as news, encyclopedias, and literature. For a clearer understanding of the Mongolian corpus, we present its specific domain distribution in Figure 2(a), and for the Chinese corpus in Figure 2(b). Ultimately, the total corpus scale for continual pre-training reached approximately 2 billion tokens. This processed and proportioned pre-training corpus will serve as a collective input for the knowledge-enhancing continual pre-training of the base model.

3.2.2. Continual Pre-training Process

We selected Gemma2-9B-it[19] as the base model, primarily considering its strong multilingual processing capabilities and its use of BPE for tokenization. These features provide a solid foundation for enhancing the model's understanding and generation capabilities for Mongolian. Compared to the BBPE tokenization method commonly used by other mainstream large language models, BPE produces tokenization results that more effectively represent the rich morphological forms and linguistic structural characteristics of agglutinative languages like Mongolian. The core difference lies in their operational units: BPE operates on characters, allowing it to learn splits that align with the language's natural morphemic boundaries. It can, for example, learn to separate a word's root from its various grammatical suffixes (such as those indicating case or plurality), thus preserving the word's inherent linguistic structure. In contrast, BBPE operates on raw bytes, which can result in arbitrary splits that cut across these natural boundaries, sometimes even breaking apart the bytes that constitute a single traditional Mongolian character. Such fragmentation creates artificial, linguistically meaningless tokens. This segmentation into more meaningful sub-words provides a superior structural foundation for the model to learn the language's grammatical patterns during the continual pre-training phase.

The primary objective of the continual pre-training stage is autoregressive language modeling. Specifically, given the pre-training corpus $D_{pre} = D_M \cup D_C$, where D_M is the Mongolian corpus, and D_C is the Chinese corpus. For any text sequence $\mathbf{s} = (x_1, x_2, \ldots, x_l)$ in D_{pre}, where x_i represents the i-th token in the sequence, the model's objective is to predict the token x_t in the sequence based on $x_{<t} = (x_1, \ldots, x_{t-1})$. This process can be represented by the following formula:

$$\hat{\theta} = \arg\max_{\theta} \sum_{\mathbf{s} \in D_{pre}} \sum_{t=1}^{|\mathbf{s}|} \log P(x_t | x_{<t}; \theta) \tag{2}$$

where θ represents the trainable parameters of the model, and $P(x_t|x_{<t};\theta)$ is the probability of the model predicting token x_t given $x_{<t}$. The optimization objective is to obtain the optimized model parameters $\hat{\theta}$ by training on D_{pre}, thereby enabling the model to more accurately capture the syntactic structures and deep semantic patterns of both languages. This serves to enhance the model's understanding and generation capabilities for Mongolian and to consolidate its Chinese processing abilities.

3.3. Symmetrical Fine-tuning

To further adapt it to the Mongolian-Chinese machine translation task, we perform task-specific fine-tuning. Based on the common understanding that an individual's proficiency in using a language for creation and expression typically indicates a more thorough understanding of that language, this study proposes a symmetrical fine-tuning strategy. This strategy, through bi-directional collaborative learning on Mongolian-Chinese translation tasks, jointly optimizes the two inverse translation tasks. This compels the model to simultaneously learn the bi-directional mapping between the two languages, thereby promoting deeper cross-lingual alignment and representation learning.

In the symmetrical fine-tuning stage, we utilize the Mongolian-Chinese dataset $\mathcal{D}_{trans}$ to construct bi-directional fine-tuning instances: Mongolian-Chinese (source = $\mathbf{x}^k$, target = $\mathbf{y}^k$) and Chinese-Mongolian (source = $\mathbf{y}^k$, target = $\mathbf{x}^k$). During the training process, we mix instances from both translation directions in each training batch for joint optimization. The optimization objective is designed as a multi-task learning loss, formulated as:

$$\mathcal{L}_{SFT}(\theta) = -\sum_{k=1}^{n} \left(\log P(\mathbf{y}^k \mid \mathbf{x}^k; \theta) + \log P(\mathbf{x}^k \mid \mathbf{y}^k; \theta) \right) \tag{3}$$

where, θ represents the model parameters. $\log P(\mathbf{y}^k \mid \mathbf{x}^k; \theta)$ denotes the log-likelihood for the Mongolian-Chinese direction; and $\log P(\mathbf{x}^k \mid \mathbf{y}^k; \theta)$ denotes the log-likelihood for the Chinese-to-Mongolian direction. By minimizing $\mathcal{L}_{\text{SFT}}(\theta)$, the model is driven to simultaneously optimize translation tasks in both directions. This compels it to learn more robust cross-lingual alignment knowledge and enhances its comprehensive understanding and generation capabilities for the linguistic structures and semantics of both languages.

4. Experiments

4.1. Datasets

We conducted experiments on CCMT2021 [27] and the in-house laboratory dataset MCMT. CCMT2021 covers multiple domains, including news and literature. It consists of 50,000 parallel sentence pairs, which we randomly divided into 40,000 training pairs, 5,000 validation pairs, and 5,000 test pairs. MCMT is collected by our laboratory, which includes diverse content such as news, legal texts, and fiction. This diverse composition ensures the dataset's breadth. Similar to CCMT2021, the MCMT dataset also contains 50,000 parallel sentence pairs. Furthermore, we adopted an identical random splitting strategy to that of CCMT2021, allocating them into 40,000 training pairs, 5,000 validation pairs, and 5,000 test pairs.

4.2. Baselines

We compared the performance of our method with the following baseline methods: (i) **mT5-large**[23] is a multilingual extension of T5-large, pre-trained on the mC4 corpus using a text-to-text denoising generative objective, enabling it to handle understanding and generation tasks across multiple languages. (ii) **LLaMA-3.1-8B-Instruct**[6] has amassed a wealth of linguistic knowledge through its large-scale pre-training process. This makes it a highly adaptable tool for diverse NLP applications, including machine translation via its strong instruction-following capabilities. (iii) **Qwen2.5-7B-Instruct**[24] leverages a predominantly English training dataset that is augmented with extensive multilingual corpora. This hybrid training approach not only solidifies its performance in English-language tasks but also significantly boosts its cross-lingual capabilities.

4.3. Experiment Settings

To ensure reproducibility, we detail the following experimental parameter settings. (1) Training: For fine-tuning Gemma2-9B-it (our base model), LLaMA-3.1-8B-Instruct, and Qwen2.5-7B-Instruct, we employed the Low-Rank Adaptation (LoRA) method for supervised fine-tuning. The LoRA_rank was set to 8, the number of epochs was 3, the initial learning rate was 1.0×10^{-4}, and the warmup_ratio was 0.2. For the mT5-large model, we performed full-parameter fine-tuning with a learning_rate of 1.0×10^{-4} and 3 epochs. Baseline models employed unidirectional Mongolian-Chinese fine-tuning. (2) Evaluation: To ensure stable evaluations during generation, we set the temperature to 0, frequency_penalty to 0, presence_penalty to 0, and max_tokens to 1024. (3) Metric: To reflect the fluency and accuracy of the model-generated text, this study

Table 1: Comparison of BLEU scores between our method and baseline models on the CCMT2021 and MCMT datasets. Baseline models employed unidirectional Mongolian-Chinese fine-tuning. Red highlights indicate the best metrics for each dataset.

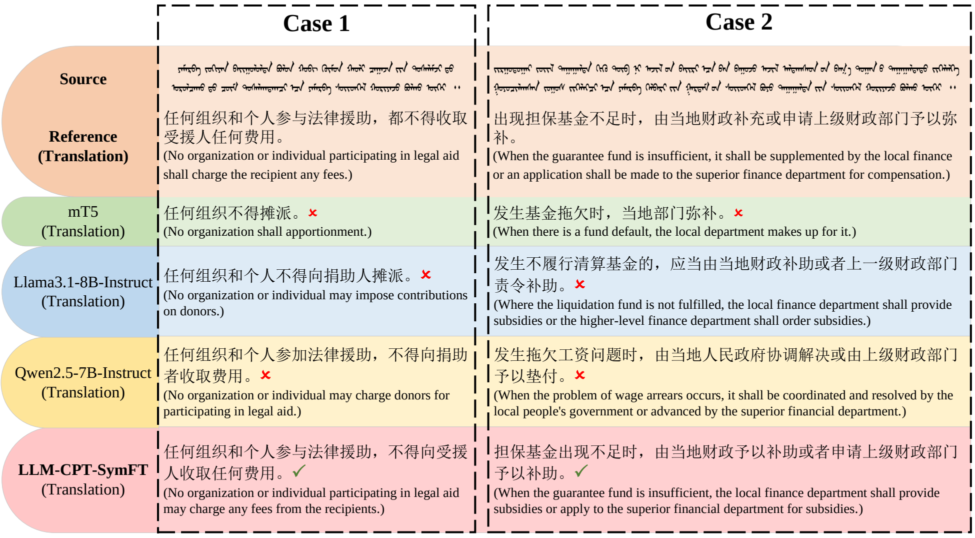
Method	**CCMT2021**	**MCMT**	**AVG**
mT5	34.15	16.25	25.20
LLaMA3.1-8B-Instruct	37.31	18.25	27.78
Qwen2.5-7B-Instruct	46.52	20.47	33.50
LLM-CPT-SymFT	72.50	42.93	57.72

uses BLEU[12] to evaluate the translation performance of the models. (4) Hardware: All model training and evaluation were conducted on 2 NVIDIA A100 GPUs. On this hardware, the symmetrical fine-tuning phase for each dataset took approximately 9 hours.

4.4. Results and Analysis

4.5. Main Results

As shown in Table 1, our proposed LLM-CPT-SymFT markedly outperforms all baseline models on both datasets. This confirms the overall superiority of the proposed method. Specifically, LLM-CPT-SymFT achieves a BLEU score of 72.50 on CCMT2021, which is 25.98 higher than that of the second-best model, Qwen2.5-7B-Instruct. On MCMT, it maintains an absolute lead with a BLEU score of 42.93, significantly outperforming Qwen2.5-7B-Instruct's score of 20.47. Overall, the proposed method achieves an average BLEU score of 57.72, representing an average improvement of 24.22 compared to Qwen2.5-7B-Instruct. This strongly demonstrates its stability and generalization ability across different test scenarios.



	Case 1	**Case 2**
Source	[illegible]	[illegible]
Reference (Translation)	任何组织和个人参与法律援助，都不得收取受援人任何费用。 (No organization or individual participating in legal aid shall charge the recipient any fees.)	出现担保基金不足时，由当地财政补充或申请上级财政部门予以弥补。 (When the guarantee fund is insufficient, it shall be supplemented by the local finance or an application shall be made to the superior finance department for compensation.)
mT5 (Translation)	任何组织不得摊派。✗ (No organization shall apportionment.)	发生基金拖欠时，当地部门弥补。✗ (When there is a fund default, the local department makes up for it.)
Llama3.1-8B-Instruct (Translation)	任何组织和个人不得向捐助人摊派。✗ (No organization or individual may impose contributions on donors.)	发生不履行清算基金的，应当由当地财政补助或者上一级财政部门责令补助。✗ (Where the liquidation fund is not fulfilled, the local finance department shall provide subsidies or the higher-level finance department shall order subsidies.)
Qwen2.5-7B-Instruct (Translation)	任何组织和个人参加法律援助，不得向捐助者收取费用。✗ (No organization or individual may charge donors for participating in legal aid.)	发生拖欠工资问题时，由当地人民政府协调解决或由上级财政部门予以垫付。✗ (When the problem of wage arrears occurs, it shall be coordinated and resolved by the local people's government or advanced by the superior financial department.)
LLM-CPT-SymFT (Translation)	任何组织和个人参与法律援助，不得向受援人收取任何费用。✓ (No organization or individual participating in legal aid may charge any fees from the recipients.)	担保基金出现不足时，由当地财政予以补助或者申请上级财政部门予以补助。✓ (When the guarantee fund is insufficient, the local finance department shall provide subsidies or apply to the superior financial department for subsidies.)

Figure 3: Two Mongolian-to-Chinese translation examples comparing outputs from mT5, Llama3.1-8B-Instruct, Qwen2.5-7B-Instruct, and LLM-CPT-SymFT against a reference translation. English translations are provided for clarity. Correct (✓) and incorrect (✗) translations are marked.

Table 2: Ablation study of Continual Pre-training (CPT) and Symmetrical Fine-tuning (SymFT), where UniFT denotes unidirectional Mongolian-Chinese translation fine-tuning.

Method	**CCMT2021**	**MCMT**	**AVG**
LLM-UniFT	60.04	31.47	45.76
LLM-SymFT	61.24	33.11	47.18
LLM-CPT-UniFT	63.00	35.87	49.44
LLM-CPT-SymFT	72.50	42.93	57.72

To intuitively demonstrate the advantages of our method, Figure 3 presents two illustrative translation examples. In Case 1, for instance, our method successfully identifies and translates both the core legal concept and the key referential term. In contrast, baseline models often misinterpret or mistranslate these critical components, leading to severe semantic deviations. Case 2, which involves the exposition of a fiscal supplementary mechanism, further underscores our model's capabilities: the translation generated by LLM-CPT-SymFT is logically coherent and contextually complete. This level of accuracy and completeness is markedly superior to that of the baseline models, which frequently exhibit deficiencies such as the omission of key information or significant semantic distortion. These examples suggest that the continual pre-training phase establishes a robust foundation for the model's comprehension of Mongolian. Furthermore, symmetrical fine-tuning optimizes translation accuracy and fluency via bi-directional learning, thereby enhancing the model's capacity to process complex long sentences and discern subtle semantic nuances within specific contexts. Consequently, the translations produced by LLM-CPT-SymFT demonstrate superior overall performance in terms of both faithfulness and fluency.

Additionally, we can observe performance differences among the baseline models from the Table1. Among the baseline models that all employed unidirectional fine-tuning, Qwen2.5-7B-Instruct and LLaMA3.1-8B-Instruct demonstrated stronger performance than the mT5 model. Specifically, Qwen2.5-7B-Instruct achieved BLEU scores of 46.52 on CCMT2021 and 20.47 on MCMT, with an average of 33.5. This performance is significantly superior to that of mT5, exceeding it by approximately 8.3 on average. LLaMA3.1-8B-Instruct also outperformed mT5 on both CCMT2021 and MCMT. This indicates that Qwen2.5 and LLaMA3.1, being newer generation large language models than mT5, possess advantages such as larger parameter scales, and greater scale and diversity in pre-training data. These factors likely contribute to their higher performance even under identical unidirectional fine-tuning conditions. Meanwhile, Qwen2.5-7B-Instruct, also a large language model, significantly outperformed LLaMA3.1-8B-Instruct on both datasets and in terms of average score. This advantage exhibited by Qwen2.5 may stem from richer Chinese and multilingual coverage in its pre-training data, specific optimizations for Asian languages, differences in model architecture details, or its instruction fine-tuning strategies, all of which could potentially influence its final performance on the Mongolian-Chinese translation task.

4.6. Ablation Study

To further evaluate the effectiveness of each component, we conducted a series of ablation studies, focusing on two key aspects: the effectiveness of Continual Pre-training and the effectiveness of Symmetrical Fine-tuning. By systematically examining these components, we aim to understand the contribution of each component to the overall performance. The following details the findings from each ablation study.

Effectiveness of Continual Pre-training: As shown in Table 2, the impact of Continual Pre-training (CPT) on performance is significant. Compared to LLM-SymFT, LLM-CPT-SymFT demonstrates substantial improvements in BLEU scores: on CCMT2021, the score increased from 61.24 to 72.50; on MCMT, from 33.11 to 42.93. Overall, the average BLEU score improved by 10.54. Furthermore, comparing LLM-CPT-UniFT with LLM-UniFT, CPT also yielded significant gains, with the average BLEU score increasing by 3.68 points. This demonstrates that continual pre-training enables the large language model to learn richer semantic information and domain knowledge, thereby effectively alleviating the model's deficiency in understanding low-resource languages.

Effectiveness of Symmetrical Fine-tuning: The results in Table 2 show that employing Symmetrical Fine-tuning (LLM-CPT-SymFT) instead of Unidirectional Fine-tuning (LLM-CPT-UniFT) increases the BLEU score on CCMT2021 from 63.0 to 72.5, and on MCMT from 35.87 to 42.93, with an average BLEU improvement of 8.28. These results indicate that symmetrical fine-tuning, by synchronously optimizing both Mongolian-Chinese and Chinese-Mongolian translation tasks, can more fully leverage the bi-directional translation knowledge and constraints within the parallel dataset than traditional unidirectional fine-tuning. This promotes the model's learning of more robust and precise cross-lingual alignments. As we posited, exercising the generation capability in a language can, in turn, promote an enhanced understanding of it.

5. Conclusion

In this paper, we proposed LLM-CPT-SymFT, a Mongolian-Chinese machine translation method based on large language models. This method begins with continual pre-training on Mongolian and Chinese monolingual corpora. Subsequently, symmetrical fine-tuning is employed, using joint optimization on the complementary Mongolian-to-Chinese and Chinese-to-Mongolian translation tasks. Experimental results across two datasets show that our proposed method significantly surpasses multiple baseline models, validating its effectiveness in Mongolian-Chinese machine translation and offering a novel approach for this low-resource language pair. Future work could explore the applicability of this framework to other low-resource language pairs to further verify its generalizability.

Acknowledgements

This work was supported in part by the Inner Mongolia Autonomous Region Science and Technology Project (2025KYPT0041, 2025KYPT0064), Hohhot City Universities and Institutes Collaborative Innovation Project, Supporting the Reform and Development Funds for Local Universities (Discipline Construction), and Inner Mongolia Autonomous Region First-Class Discipline Scientific Research Special Project (No. YLXKZX-ND-036).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

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

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

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pages 4171–4186, 2019.

[6] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

[7] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. arXiv preprint arXiv:2410.21276, 2024.

[8] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759, 2016.

[9] Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In Proceedings of the 2013 conference on empirical methods in natural language processing, pages 1700–1709, 2013.

[10] Palanichamy Naveen and Pavel Trojovskỳ. Overview and challenges of machine translation for contextually appropriate translations. Iscience, 27(10), 2024.

[11] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35:27730–27744, 2022.

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL `https://aclanthology.org/P02-1040/`.

[13] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[14] Felix Stahlberg. Neural machine translation: A review. Journal of Artificial Intelligence Research, 69:343–418, 2020.

[15] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. Advances in neural information processing systems, 27, 2014.

[16] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023.

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

[18] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295, 2024.

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

[20] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. arXiv preprint arXiv:2503.19786, 2025.

[21] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.

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

[23] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934, 2020.

[24] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115, 2024.

[25] Ji Yatu, Zhang Huinuan, Wu Nier, Lu Min, Shi Bao, et al. A review of mongolian neural machine translation from the perspective of training. In 2024 International Joint Conference on Neural Networks (IJCNN), pages 1–10. IEEE, 2024.

[26] Yujuan Yin, Lixia Wen, Heya Sa, Bao Shi, et al. Research on mongolian-chinese neural machine translation based on csgan and co-training. In Journal of Physics: Conference Series, volume 2219, page 012061. IOP Publishing, 2022.

[27] Shen Yingli, Bao Wugedele, and Zhao Xiaobing. Mongolian-Chinese Machine Translation Correction Dataset, June 2022.

[28] Aijun Zhang, Weijian Hu, Mingxing Luo, and Lingfang Li. Chinese-mongolian machine translation combining sentence structure information. In 2023 8th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), pages 588–592. IEEE, 2023.

[29] Biao Zhang, Barry Haddow, and Alexandra Birch. Prompting large language model for machine translation: A case study. In International Conference on Machine Learning, pages 41092–41110. PMLR, 2023.

[30] Xiu Zhi and Siriguleng Wang. Research on the application of bert in mongolian-chinese neural machine translation. In Proceedings of the 2021 13th International Conference on Machine Learning and Computing, pages 404–409, 2021.

Author Biographies

Zhenjie Gao is currently a Ph.D. Candidate in Computer Science at Inner Mongolia University. His research work focuses on machine translation, large language models, and Mongolian information processing.
Email: gzj@mail.imu.edu.cn

Feilong Bao, PhD in Engineering, is the professor, and PhD supervisor at the College of Computer Science, Inner Mongolia University. His research work centers on Mongolian information technology, computational linguistics, and Mongolian information systems.
Email: csfeilong@imu.edu.cn

Yuan Li is currently a Ph.D. Candidate in Computer Science at Inner Mongolia University.Research Focus: Speech Recognition & Natural Language Processing.
Email: liyuan@mail.imu.edu.cn

Ruichen Hou is currently a doctoral candidate at Inner Mongolia University, enrolled in the 2024 cohort, with my primary research focus being on the low-resource domain of large language models and the evaluation of such models.
Email: guilty.kiss@qq.com

Yibo Han is currently pursuing a Master's degree in Computer Science and Technology at the School of Computer Science and Artificial Intelligence, Inner Mongolia University. His research interests include large language models and machine translation.
Email: hybnivk@163.com