LinChance×NTU for Unconstrained WMT2024 Literary Translation

Kechen Li[1], Yaotian Tao[1], Hongyi Huang[1], Tianbo Ji[2]
[1]Jiangsu Linchance Technology Co., Ltd. (LinChance)
[2]School of Transportation and Civil Engineering, Nantong University
{likechen,taoyaotian,huanghongyi}@linchance.com, jitianbo@ntu.edu.cn

Abstract

The rapid growth of deep learning has spurred significant advancements across industries, particularly in machine translation through large language models (LLMs). However, translating literary still presents challenges, including cross-cultural nuances, complex language structures, metaphorical expressions, and cultural differences. To address these issues, this study utilizes the Llama and Phi models using both LoRA and full-parameter techniques, alongside a prompt-based translation system. Full-parameter tuning of the Llama-3-Chinese-8B-Instruct model was unsuccessful due to memory constraints. In terms of the WMT task, the fully fine-tuned Phi 3 model was selected for submission due to its more natural and fluent translations. Nonetheless, results showed that LoRA and the prompt-based system significantly improved the Llama3 model's performance, surpassing other models in BLEU and ROUGE evaluations.

1 Introduction

In recent years, the development of deep learning has spread across various industries (Ji et al., 2024), and the impact of large language models (LLMs) on these industries has been particularly significant (Lyu et al., 2023). Despite the fact that many challenges in machine translation (MT) have been overcome (Wang et al., 2023), literary translation still encounters cross-cultural issues, including such as processing complex languages, understanding metaphorical expressions, and addressing cultural differences (Lyu et al., 2020). Meanwhile, choosing the right model has become a key topic in terms of neural network-based (NN-based) MT (Xia, 2020), as models based on the source language typically have advantages in handling tasks regarding that language. Two main models are involved in this research: Llama3-Chinese-8B-Instruct[1] and Phi-3-mini-128k-instruct-Chinese[2]. The former fine-tuned the Llama3 model (Dubey et al., 2024) with 5 million instruction data points from the community, which significantly enhances its performance in Chinese-language tasks with a better ability of understanding Chinese contexts. The latter is with less than half the size (3.8B parameters) of the Llama3 8B version, which can surpass the performance of Llama3 with less computational resources.

LoRA (Sundaram et al., 2019) is a lightweight fine-tuning technique mainly used for efficiently training large models. Compared to traditional full-parameter fine-tuning, LoRA decomposes the trained parameter matrices into low-rank forms, resulting in a reducing number of parameters and less computational cost of training. It is especially appropriate for fine-tuning LLMs with constrained resources while maintaining high performance. In this research, we utilize Llama-Factory[3] (Zheng et al., 2024), an optimized framework designed specifically for fine-tuning LLMs like Llama. It supports various advanced training techniques, including mixed precision training and gradient accumulation, to improve training efficiency and reduce computational resource requirements. By integrating lightweight methods like LoRA, it can achieve efficient and stable model fine-tuning in resource-constrained environments, helping to quick adjustment and deployment of LLMs.

In general, our contributions can be summarized as follows:

- We conduct a comprehensive experiment of two major large language models, Llama-3-Chinese-8B-Instruct and Phi-3-mini-128k-instruct-Chinese, for the task of WMT2024

[1]https://huggingface.co/hfl/Llama-3-chinese-8b-instruct
[2]https://huggingface.co/shareAI/Phi-3-mini-128k-instruct-Chinese
[3]https://github.com/hiyouga/Llama-Factory

Proceedings of the Ninth Conference on Machine Translation, pages 987–992
November 15-16, 2024

Literary Translation. Our results demonstrate that both models perform excellently in handling Chinese tasks, especially when facing cross-cultural challenges in literary translation.

- We applied the LoRA technique to efficiently fine-tune the Llama model, significantly reducing computational costs while maintaining high translation quality. We additionally optimized the fine-tuning process by leveraging the Llama-Factory framework. Our experimental results demonstrate that the combination of LoRA and Llama-Factory can effectively support the adaptation and deployment of large-scale models in resource-constrained environments.

- We investigate the strengths and weaknesses of each model, particularly in terms of fluency and diversity (captured by ROUGE) as well as accuracy (captured by BLEU).

2 Related Work

Large Language Models (LLMs) for Machine Translation The application of Large Language Models (LLMs) in machine translation (MT) has seen significant advancements, particularly in general domain translation (Wang et al., 2023). Pretrained models such as Llama and Phi3 have been increasingly employed for tasks requiring semantic understanding across languages. Studies have highlighted how instruction-tuned LLMs can improve translation quality by adapting to the syntactic structures and cultural nuances of target languages. This is particularly relevant for our work as we evaluate the Llama-3 Chinese model (Cui et al., 2023), which is fine-tuned for literary translation.

Challenges in Literary Translation While MT systems have progressed in many domains (Du et al., 2024), the translation of literary texts remains particularly challenging due to the need to capture nuanced expressions, idioms, and stylistic elements. Literary translation (Jones, 2019) is often considered the "last frontier" for MT. Prior work has explored how traditional sentence-based MT systems struggle with long, complex passages found in literary texts (Aliguliyev, 2009). This is in line with the focus of our experiments, which attempt to handle these unique challenges using Llama-3 Chinese models and fine-tuning techniques.

Fine-Tuning Techniques for MT In order to address the computational limitations and language understanding challenges associated with large models, various fine-tuning approaches have been proposed (Nicholas and Bhatia, 2023). Recent studies on Low-Rank Adaptation (LoRA) have demonstrated that memory-efficient fine-tuning methods allow for high-quality performance on GPUs with limited memory. LoRA's success in reducing memory consumption while maintaining model accuracy has been a key technique in our experiments, particularly with the Llama-3 Chinese model.

Evaluation of Translation Quality The evaluation of literary translations poses its own challenges, as traditional metrics like BLEU may not fully capture the nuances of a good translation. (Pang et al., 2024) Newer approaches, such as Monolingual Human Preference (MHP) and Bilingual LLM Preference (BLP), have been proposed to better assess translation quality in a literary context. (Wu et al., 2024) Our experiments draw on these evaluation techniques, comparing model outputs through both automated metrics and human preference assessments to gauge the effectiveness of different fine-tuning strategies.

3 Experiment

3.1 Evaluation Metrics

To achieve accurate evaluation of MT result (Chang et al., 2024), two prevailing evaluation metrics are utilized: BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). BLEU is an automated evaluation metric based on n-gram matching, primarily used to measure the similarity between machine translation outputs and reference translations. By calculating the overlap of n-grams of different lengths between the model's output and the reference translation, BLEU can reflect the accuracy of the translation to a certain extent. ROUGE is a widely used automatic text evaluation metric mainly used to compare the similarity between generated text and reference text. ROUGE-L, in particular, is based on the Longest Common Subsequence (LCS) (Bergroth et al., 2000) and measures the similarity between generated text and reference text in terms of length matching and word order. Compared to BLEU, ROUGE captures both text diversity and fluency.

3.2 Prompt Engineering

The prompt design focuses on the task of translating Guofeng (traditional Chinese-style) novels,

aiming to ensure that the translated text faithfully conveys the literary and cultural nuances of the original, while maintaining translation efficiency and accuracy. By establishing clear guidelines, the prompt emphasizes fidelity to the original text, concise output, and quality control to ensure that the translation remains fluent while preserving the original style and tone. The prompt is designed following the CoT (Chain of Thoughts) framework (Wei et al., 2022), with the specific approach outlined below:

First, the prompt introduces automatic language detection and translation features, enabling efficient Chinese-to-English translation of Guofeng novels to enhance processing speed and coherence. Secondly, fidelity to the original is critical, requiring the preservation of the original tone, style, and expression. Special attention is given to details such as pronouns, with a focus on word-for-word translation to avoid distorting the literary essence due to cultural or linguistic differences. The prompt further emphasizes objectivity in translation, avoiding any omissions or commentary, ensuring the completeness and authenticity of the translated text.

Additionally, the translated text must be concise, with no added annotations, ensuring that the classical charm and cultural context of the novel are naturally conveyed, enhancing the reader's experience. To ensure quality, the prompt requires thorough review and correction of the translated output, avoiding mistranslations or omissions, and ensuring that the text aligns with the target language's fluency and conventions. The prompt is task-oriented, providing only the final, revised translation, avoiding irrelevant information or excessive explanation, which improves processing efficiency and suits large-scale Guofeng novel translation projects.

3.3 Experiment 1: Training with Llama-3-Chinese-8B-Instruct

An initial attempt was made to fully train the Llama-3-Chinese-8B-Instruct model on the dataset. However, the process failed due to insufficient memory. The model's large size and the memory requirements exceeded the capabilities of the available hardware, necessitating use a smaller model or shift to a more memory-efficient fine-tuning method.

3.4 Experiment 2: Fine-Tuning Llama3 8B Model with LoRA

Given the memory constraints, the Llama3 8B model was fine-tuned using the LoRA technique on a dataset of 10,000 samples. The fine-tuning was performed with several key hyperparameters to optimize model performance and manage computational resources effectively. The learning rate was set to 1e-5 (Jin et al., 2023), using a cosine learning rate scheduler to gradually reduce the learning rate and improve training efficiency (Kim et al., 2021). A per-device train batch size of 2 was chosen to balance between memory usage and model update frequency, with gradient accumulation over 16 steps to simulate a larger batch size without requiring additional GPU memory. The training process was conducted over 10 epochs to ensure sufficient learning from the data, utilizing a maximum of 10,000 samples. Additionally, the model was trained with mixed precision (fp16) (Le Gallo et al., 2018) to reduce memory usage and accelerate computation. The evaluation strategy was set to evaluate the model performance at regular steps to monitor its progress closely. The results of this fine-tuning experiment are summarized in Table 1.

Table 1: Results of Fine-Tuning Llama3 8B Model with LoRA

Metric	Value	Description
BLEU-4	55.28	A metric which evaluates the quality of a candidate by computing the n-gram ($n = 4$) precision with references.
ROUGE-1	60.18	A variation of ROUGE where 1 means unigrams.
ROUGE-2	37.40	A variation of ROUGE where 2 means bigrams.
ROUGE-L	55.91	A variation of ROUGE where L means longest common subsequences (LCS).
Runtime	3m8s	Total runtime
Sample/s	1.594	Samples processed per second
Step/s	1.594	Training steps per second

Analysis: The results from this experiment were quite promising, with high BLEU and ROUGE scores. The LoRA technique allowed the model to be fine-tuned without running into memory issues, demonstrating that it is an effective method for working with large models on limited hardware. The high ROUGE scores suggest that the model

was able to generate translations that were both accurate and fluent.

3.5 Experiment 3: Full Fine-Tuning with Phi Chinese Model

The Phi Chinese model was fully fine-tuned on a smaller dataset of 2,000 samples.

Table 2: Results of Full Fine-Tuning with Phi Chinese Model

Metric	Value
BLEU-4	50.93
ROUGE-1	51.90
ROUGE-2	27.19
ROUGE-L	46.41
Runtime	16m34s
Sample/s	0.503
Step/s	0.503

Analysis: The performance of the Phi Chinese model, while adequate, was noticeably lower than that of the Llama3 8B model fine-tuned with LoRA. The lower BLEU and ROUGE scores could be attributed to the smaller model size and the limited dataset, which may not have provided enough data for the model to generalize well.

3.6 Experiment 4: Full Fine-Tuning with Phi3 Chinese 3.5B Model

The Phi3 Chinese 3.5B model was fully fine-tuned on a dataset of 1,500,000 samples. The fine-tuning process was carefully configured with a set of key hyperparameters to optimize the model's performance while efficiently managing computational resources. We set the learning rate to 1e-5, which is low enough to ensure stable training and prevent the model from overshooting optimal weights but sufficient to allow for meaningful updates to the model parameters. A batch size of 128 was chosen to strike a balance between training speed and memory constraints. To further accommodate large batch sizes, a gradient accumulation step of 16 was used, effectively increasing the batch size without exceeding GPU memory limits. The model was trained using mixed-precision floating-point, allowing for faster computation and reduced memory usage, which is crucial when dealing with large-scale models. We set the number of epochs to 3.0 to provide sufficient training cycles while minimizing the risk of overfitting. A temperature parameter of 0.4 was employed during the generation phase to control the randomness and diversity of the model's output, balancing between creativity and coherence. The results of this fine-tuning experiment are summarized in Table 3.

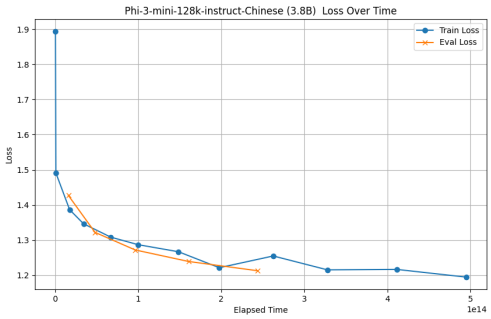

Figure 1: Loss Over Time for the Phi-3-mini-128k-instruct-Chinese (3.8B) model.

Table 3: Results of Full Fine-Tuning with Phi3 Chinese 3.5B Model

Metric	Value
BLEU-4	49.14
ROUGE-1	49.80
ROUGE-2	25.09
ROUGE-L	45.24
Runtime	2m36s
Sample/s	1.278
Step/s	1.278

Training and Evaluation Loss: To evaluate the fine-tuning process of the Phi-3-mini-128k-instruct-Chinese model, we tracked the training and evaluation loss over time, as illustrated in Figure 1. The model, which consists of 3.8 billion parameters, was fine-tuned on a diverse dataset using LoRA and full-parameter tuning techniques. Both the training and evaluation losses were monitored to assess model convergence and stability during the fine-tuning process.

Loss Over Time: As shown in Figure 1, the initial training loss starts relatively high, around 1.9, and decreases sharply during the early stages of training. By the end of the first epoch, the loss drops to approximately 1.3, indicating that the model quickly learns to generalize to the underlying patterns in the training data. The evaluation loss follows a similar trend, closely mirroring the training loss, which suggests that the model gener-

alizes well without overfitting during the training process. By the second epoch, the loss stabilizes around 1.2 for both training and evaluation, demonstrating the model's ability to maintain consistent performance throughout the training process. The convergence of the loss indicates that the model is reaching its optimal capacity under the current fine-tuning setup.

Observations and Insights: The relatively close alignment of training and evaluation losses suggests that the fine-tuning process successfully mitigated the risk of overfitting, which is often a concern when dealing with large models and smaller, task-specific datasets. Moreover, the overall reduction in loss suggests that the Phi-3-mini-128k-instruct-Chinese model was able to effectively capture the nuances of the Chinese language and the intricate nature of literary translation tasks, as intended in this study.

4 Conclusion and Future Work

In this paper, we conduct experiments to provide valuable insights into the performance of different models and fine-tuning techniques for the WMT2024 Literary Translation Task. The Llama3 8B model, when fine-tuned using the LoRA technique, demonstrated the best performance, highlighting the importance of memory-efficient training methods in dealing with large models on limited hardware. The results from the Phi and Phi3 models suggest that model size alone may not guarantee the better performance, and the choices of fine-tuning method and dataset size are critical factors in achieving high-quality translations. In the future, we plan to investigate the performance of even larger models (e.g., Llama3 70B) to explore the trade-offs between model size, computational resources, and translation quality. In addition, since the metrics we used may correlate negatively with human judgements (Ji et al., 2022), developing task-specific evaluation metrics would be valuable for the accurate assessment of model performance.

References

Ramiz M Aliguliyev. 2009. A new sentence similarity measure and sentence based extractive technique for automatic text summarization. *Expert Systems with Applications*, 36(4):7764–7772.

Lasse Bergroth, Harri Hakonen, and Timo Raita. 2000. A survey of longest common subsequence algorithms. In *Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000*, pages 39–48. IEEE.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Zefeng Du, Wenxiang Jiao, Longyue Wang, Chenyang Lyu, Jianhui Pang, Leyang Cui, Kaiqiang Song, Derek F Wong, Shuming Shi, and Zhaopeng Tu. 2024. On extrapolation of long-text translation with large language models. In *Findings of the Association for Computational Linguistics*.

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

Tianbo Ji, Yvette Graham, Gareth Jones, Chenyang Lyu, and Qun Liu. 2022. Achieving reliable human assessment of open-domain dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6416–6437, Dublin, Ireland. Association for Computational Linguistics.

Tianbo Ji, Kechen Li, Quanwei Sun, and Zexia Duan. 2024. Urban transport emission prediction analysis through machine learning and deep learning techniques. *Transportation Research Part D: Transport and Environment*, 135:104389.

Hongpeng Jin, Wenqi Wei, Xuyu Wang, Wenbin Zhang, and Yanzhao Wu. 2023. Rethinking learning rate tuning in the era of large language models. In *2023 IEEE 5th International Conference on Cognitive Machine Intelligence (CogMI)*, pages 112–121. IEEE.

Francis R Jones. 2019. Literary translation. In *Routledge encyclopedia of translation studies*, pages 294–299. Routledge.

Chiheon Kim, Saehoon Kim, Jongmin Kim, Donghoon Lee, and Sungwoong Kim. 2021. Automated learning rate scheduler for large-batch training. *arXiv preprint arXiv:2107.05855*.

Manuel Le Gallo, Abu Sebastian, Roland Mathis, Matteo Manica, Heiner Giefers, Tomas Tuma, Costas Bekas, Alessandro Curioni, and Evangelos Eleftheriou. 2018. Mixed-precision in-memory computing. *Nature Electronics*, 1(4):246–253.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chenyang Lyu, Tianbo Ji, and Yvette Graham. 2020. Incorporating context and knowledge for better sentiment analysis of narrative text. In *Text2Story@ ECIR*, pages 39–45.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*.

Gabriel Nicholas and Aliya Bhatia. 2023. Lost in translation: large language models in non-english content analysis. *arXiv preprint arXiv:2306.07377*.

Jianhui Pang, Fanghua Ye, Derek F Wong, and Longyue Wang. 2024. Anchor-based large language models. In *Findings of the Association for Computational Linguistics*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jothi Prasanna Shanmuga Sundaram, Wan Du, and Zhiwei Zhao. 2019. A survey on lora networking: Research problems, current solutions, and open issues. *IEEE Communications Surveys & Tutorials*, 22(1):371–388.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Minghao Wu, Yulin Yuan, Gholamreza Haffari, and Longyue Wang. 2024. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *arXiv preprint arXiv:2405.11804*.

Ying Xia. 2020. Research on statistical machine translation model based on deep neural network. *Computing*, 102(3):643–661.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.