Context-aware and Style-related Incremental Decoding framework for Discourse-Level Literary Translation

Yuanchang Luo, Jiaxin Guo, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu , Zhiqiang Rao, Shaojun Li, Jinlong Yang, Hao Yang
Huawei Translation Service Center, Beijing, China
{luoyuanchang1,guojiaxin1,weidaimeng,shanghengchao,lizongyao,
wuzhanglin2,raozhiqiang,lishaojun18,yangjinlong7,yanghao30}@huawei.com

Abstract

This report outlines our approach for the WMT24 Discourse-Level Literary Translation Task, focusing on the Chinese-English language pair in the Constrained Track. Translating literary texts poses significant challenges due to the nuanced meanings, idiomatic expressions, and intricate narrative structures inherent in such works. To address these challenges, we leveraged the Chinese-Llama2 model, specifically enhanced for this task through a combination of Continual Pre-training (CPT) and Supervised Fine-Tuning (SFT). Our methodology includes a novel Incremental Decoding framework, which ensures that each sentence is translated with consideration of its broader context, maintaining coherence and consistency throughout the text. This approach allows the model to capture long-range dependencies and stylistic elements, producing translations that faithfully preserve the original literary quality. Our experiments demonstrate significant improvements in both sentence-level and document-level BLEU scores, underscoring the effectiveness of our proposed framework in addressing the complexities of document-level literary translation.

1 Introduction

Machine Translation (MT) (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023) has become an essential tool in breaking language barriers, enabling the automatic translation of text from one language to another. While significant advancements (Vaswani et al., 2017; Sennrich et al., 2016; Wei et al., 2023; Gu et al., 2018; Ghazvininejad et al., 2019; Wang et al., 2021; Guo et al., 2021; Yu et al., 2021) have been made in MT for various text genres, translating literary texts remains a formidable challenge. Literary texts are rich in complex linguistic phenomena, such as nuanced meanings, idiomatic expressions, and intricate narrative structures. Unlike technical or news-related texts, literary works demand a deeper understanding of context, tone, and style, making them particularly challenging for MT systems. This difficulty is compounded by the scarcity of high-quality parallel datasets in the literary domain, limiting the ability of MT models to learn from extensive, diverse examples.

Document-level translation (Sun et al., 2020; Du et al., 2024; Wu et al., 2024) introduces another layer of complexity to MT, especially when dealing with longer texts such as novels. Unlike sentence-level translation, where context is limited to a single sentence, document-level translation requires the model to consider the broader discourse context to maintain coherence and consistency throughout the entire text. This is particularly crucial in literary translation, where the narrative thread, character development, and thematic elements must be preserved across sentences and paragraphs. Long-range dependencies, where information introduced early in a text influences later parts, pose a significant challenge for MT systems, which often struggle to retain and apply such context effectively over extended texts.

In this system report, we describe our participation in the WMT24 Discourse-Level Literary Translation Task, focusing on the Chinese-English language pair under the Constrained Track. Our approach leverages the Chinese-Llama2 model, specifically designed for this task, through a combination of Continual Pre-training (CPT) and Supervised Fine-Tuning (SFT). This methodology allows us to refine the model's understanding of literary texts while adapting it to the specific nuances of Chinese-English translation. Additionally, we employ an Incremental Decoding framework, which enables the model to translate documents sentence by sentence, ensuring that each translation is informed by the broader context. This approach is designed to tackle the challenges of document-level literary translation, aiming to produce translations

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 973–979
November 15–16, 2024.

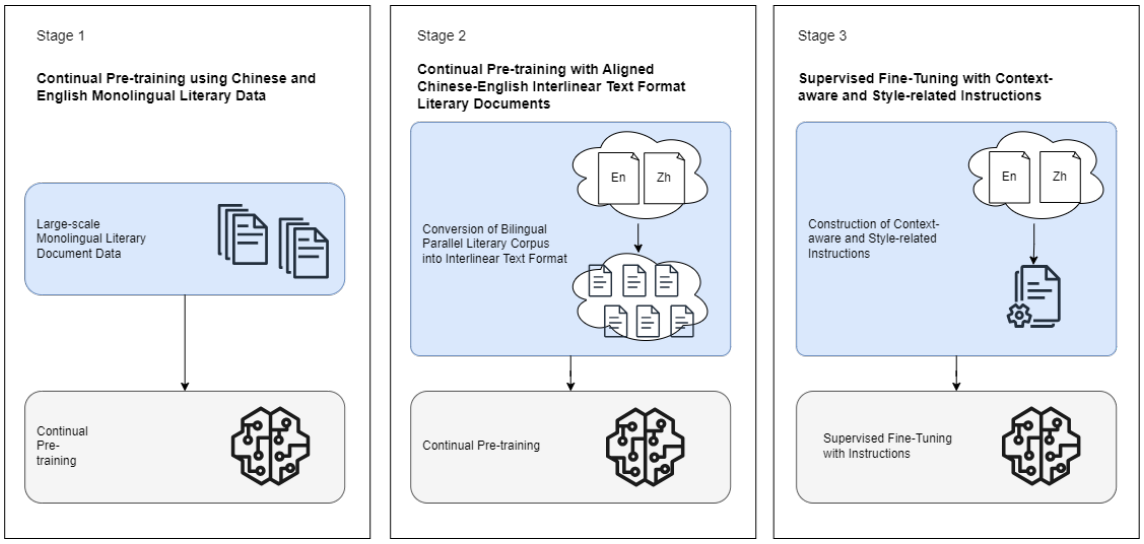

Figure 1: The overall of our approach.

that are not only accurate but also faithful to the original text's literary quality.

2 Background: TP3

Machine Translation (MT) is the automated process of converting text from one language to another using computational methods. Traditionally, MT relies on encoder-decoder models, where the encoder processes the source language and the decoder generates the translation, often requiring large bilingual datasets and data augmentation to improve performance. Recently, Large Language Models (LLMs) like GPT have become prominent in MT, enabling translation through zero-shot or few-shot learning by conditioning on a source sentence (Jiao et al., 2023; Zeng et al., 2023; Chen et al., 2023; Xu et al., 2023; Yang et al., 2023; Zhang et al., 2023). These models can also be fine-tuned with high-quality bilingual data and tailored instructions to enhance translation accuracy and robustness, offering new possibilities for MT with limited resources.

TP3 Guo et al. (2024) propose a novel training paradigm, consisting of Three-Stages Translation Pipeline (TP3), to boost the translation capabilities of LLMs. The training paradigm includes:

Stage 1: Continual Pre-training using Extensive Monolingual Data. This stage aims to expand the multilingual generation capabilities of LLMs. While it is inherently related to machine translation tasks, it is not essential.

Stage 2: Continual Pre-training with Interlinear Text Format Documents. They construct interlinear text format from sentence-aligned bilingual parallel data and utilize them for continual pre-training of LLMs. Experimental results demonstrate the critical importance of this stage, resulting in a significant improvement in translation quality, particularly for English-Other translations.

Stage 3: Leveraging Source-Language Consistent Instruction for Supervised Fine-Tuning. In this stage, they discover that setting instructions consistent with the source language benefits the supervised fine-tuning process.

3 Methods

3.1 TP3 for Discourse-Level Literary

We introduce the TP3 training paradigm into the literary translation task, with the entire training process illustrated in Figure 1.

Stage 1: Continual Pre-training using Chinese and English Monolingual Literary Data In this stage, we adapt a general-purpose large language model (LLM) into a specialized Literary LLM by using monolingual literary data in both Chinese and English. While existing LLMs like Llama perform well in English-centric tasks, their capabilities in other languages, especially in literary contexts, are often limited. To improve this, we employ continual pre-training with extensive monolingual literary texts, enhancing the model's understanding of nuanced language, stylistic elements, and narrative structures. This step is critical for enabling the model to generate more coherent and contextually appropriate translations.

For this task, continual pre-training is essential, transforming a general LLM into one tailored for

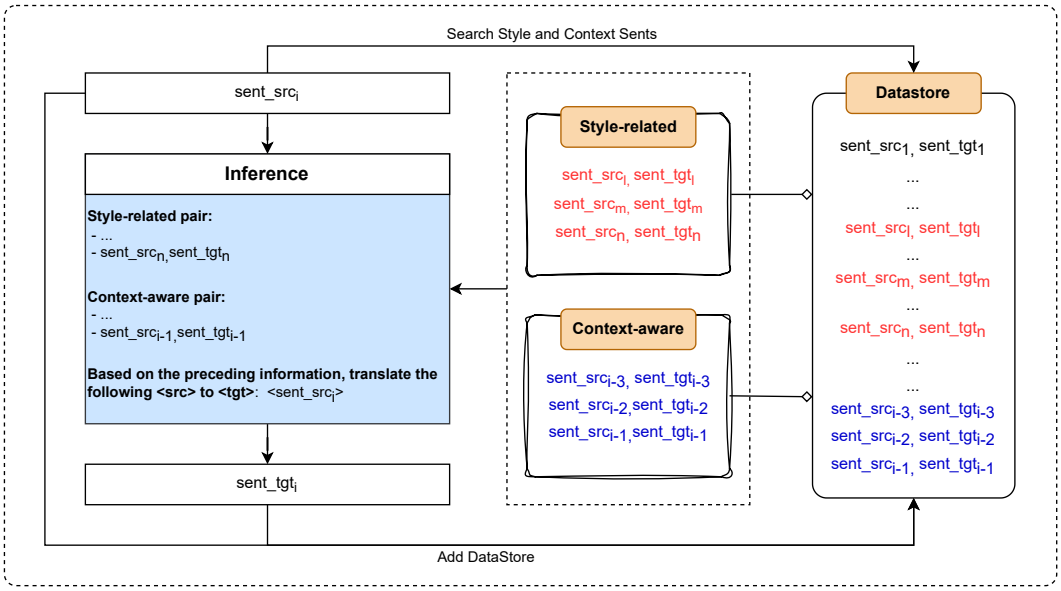

Figure 2: The overall of our incremental decoding framework.

literary translation. We treat each novel as a distinct training unit, combining sentences within each chapter into paragraphs to capture long-range dependencies and context. This approach is vital for maintaining consistency and preserving the literary quality of translations. By focusing on both Chinese and English literary data, the model gains a balanced understanding of the stylistic and structural intricacies in both languages.

Stage 2: Continual Pre-training with Aligned Chinese-English Interlinear Text Format Literary Documents In Stage 2, we enhance the model's cross-lingual translation capabilities by using aligned Chinese-English interlinear text format literary documents, building on the foundation established in Stage 1. The interlinear text format, where each source sentence is directly aligned with its translation at the word or phrase level, is essential for enabling the model to understand and map the syntactic and semantic structures between Chinese and English, which is crucial for producing high-quality translations. We implement a continual pre-training approach using LoRA (Low-Rank Adaptation of Large Language Models) (Hu et al., 2021) to efficiently adapt the model with these interlinear text documents.

Initially, the model was trained on general sentence-aligned parallel data to establish a strong cross-lingual alignment foundation. Subsequently, we performed incremental pre-training with literary-specific interlinear data. By focusing on literary documents, we ensure the model becomes finely attuned not only to general cross-lingual translation but also to the unique stylistic and structural nuances of literary texts. This approach enables the model to capture the intricate relationships between Chinese and English in a literary context, significantly improving translation quality and fidelity.

Stage 3: Supervised Fine-Tuning with Context-aware and Style-related Instructions In the final stage of our approach, we conduct supervised fine-tuning using context-aware and style-related instructions, specifically tailored to address the challenges of semantic coherence and stylistic consistency in literary translation. Unlike the traditional approach of using Source-Language Consistent Instruction, which emphasizes alignment with the source language, our method focuses on ensuring that the translated output maintains a consistent narrative flow and adheres to the stylistic nuances of the original text. This adjustment is crucial for literary translation, where preserving the author's voice and the overall tone of the work is just as important as achieving accurate translation.

The fine-tuning process leverages the LoRA to refine specific parameters of the model efficiently. By applying LoRA, we can update the model with low-rank adaptations, which helps in preventing overfitting while ensuring that the model adapts ef-

fectively to the task-specific requirements. This targeted fine-tuning allows the model to better capture the long-range dependencies and stylistic elements that are essential for producing translations that are not only accurate but also faithful to the literary qualities of the source text.

3.2 Incremental Decoding framework

In traditional machine translation, sentences are often translated independently of one another, leading to issues with semantic coherence and stylistic consistency when viewed from a broader, document-level perspective. To address these challenges, we propose an Incremental Decoding framework that considers the translation of each sentence as part of a continuous process, taking into account the translations of previous sentences. This method ensures that the translated text maintains a cohesive flow and consistent style throughout the entire document.

The Incremental Decoding framework incorporates two key components: Context-aware information and Style-related information. Context-aware information involves using the translations of the previous n sentences as historical context when translating the current sentence. This helps maintain continuity in the narrative and ensures that the translation aligns with the broader context established in earlier sentences.

Style-related information further refines this process by incorporating translations of sentences that are similar to the current sentence in terms of content and style. These sentences are selected based on sentence and keyword similarity, ensuring that the translation reflects the stylistic nuances present in the original text. By integrating both context-aware and style-related information, the Incremental Decoding framework produces translations that are not only accurate but also harmonious in tone and structure, closely mirroring the original literary work.

4 Experiments

4.1 Datasets and Evaluation Metrics

We utilized data from the general MT shared task and the GuoFeng Webnovel Corpus. The GuoFeng Webnovel Corpus was employed in Stages 1, 2, and 3, while the general MT data was used exclusively in Stage 2. Detailed statistics of the data are presented in Table 2.

For the evaluation metrics, we utilized SacreBLEU (Papineni et al., 2002) to assess system performance. Given that the test set was segmented into sentence-level units, we conducted evaluations using both s-BLEU (sentence-level BLEU) and d-BLEU (document-level BLEU) scores to provide a comprehensive analysis of the translation quality.

4.2 Experiment Settings

In our experiments, we used Chinese-LLaMA2 (Cui et al., 2023) as the foundation model. Chinese-LLaMA2 is an enhanced and optimized version of Llama-2, specifically designed for Chinese language understanding and instruction comprehension. This model includes a larger Chinese vocabulary and benefits from incremental pretraining on a large-scale Chinese dataset, which significantly improves its semantic understanding capabilities.

For both the Continual Pre-training and Supervised Fine-Tuning stages, we adhered to the hyperparameters utilized in the Chinese-LLaMA2 project. During Stage 2, the model was trained for 1 epoch, while in Stage 3, the training was extended to 3 epochs to ensure more refined adjustments.

All experiments were conducted using 8 Nvidia GPUs, each with 64GB of memory, and employed DeepSpeed (Rasley et al., 2020) ZeRO 2 for model parallelization, which allowed for efficient handling of the large-scale model and dataset.

4.3 Compared Baselines

- **General Sent-Trans**: In this baseline, we directly create sentence-level translation instruction data and use it to perform Supervised Fine-Tuning on the Chinese-LLaMA2 model. This approach focuses on training the model with general sentence-level translation tasks without any specialized pre-training.

- **Literary Sent-Trans**: This baseline builds on the previous stages, as outlined in Stage 1 and Stage 2. We first subject the Chinese-LLaMA2 model to Continual Pre-training using monolingual and bilingual literary data. Following this pre-training, the model undergoes Supervised Fine-Tuning using the same sentence-level translation instruction data as in the General Sent-Trans baseline. This approach is designed to adapt the model to the literary domain before fine-tuning it with general sentence-level instructions.

	Valid 1		Valid 2		Test 1		Test 2	
	s-BLEU	d-BLEU	s-BLEU	d-BLEU	s-BLEU	d-BLEU	s-BLEU	d-BLEU
General Sent-Trans	16.81	24.1	10.74	17.39	17.97	25.87	13.32	20.37
Literary Sent-Trans	23.35	30.51	14.64	21.81	20.91	28.51	18.02	25.38
Literary Doc-Trans	23.78	31.85	14.94	22.12	20.97	29.43	18.28	25.62

Table 1: **The overall results.**

Data Source	Data Size
General MT	25M
GuoFeng Webnovel Corpus	1.9M

Table 2: Data Statistics.

- **Literary Sent-Trans**: **This represents our final proposed approach**. After the Continual Pre-training conducted in Stage 1 and Stage 2, we further train the model using the Supervised Fine-Tuning method from Stage 3, which incorporates Context-aware and Style-related Instructions. This method aims to enhance the model's ability to maintain semantic coherence and stylistic consistency across sentences in literary document translation.

4.4 Results

The comparison between Literary Sent-Trans and General Sent-Trans reveals significant improvements in both s-BLEU and d-BLEU scores across various test sets, indicating that Stage 1 and Stage 2 effectively incorporated literary knowledge into the model. Furthermore, when comparing Literary Doc-Trans with Literary Sent-Trans, we observe additional gains in both s-BLEU and d-BLEU metrics, demonstrating the effectiveness of Stage 3's Context-aware and Style-related Instructions. These results collectively highlight the incremental benefits of each stage in enhancing the model's performance in literary translation. The detailed results are presented in Table 1.

5 Conclusion

In this work, we addressed the complex task of literary translation within the WMT24 Discourse-Level Literary Translation Task, focusing on the Chinese-English language pair. By leveraging the Chinese-Llama2 model, enhanced through Continual Pre-training and Supervised Fine-Tuning, we successfully adapted the model to capture the unique nuances of literary texts. Our Incremental Decoding framework further ensured that each sentence was translated with awareness of its broader context, resulting in more coherent and stylistically consistent translations. The improvements observed in both sentence-level and document-level BLEU scores validate the effectiveness of our approach. These results highlight the potential of combining advanced language models with specialized training strategies to tackle the intricacies of literary translation, paving the way for further research in this challenging domain.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2023. Improving translation faithfulness of large language models via augmenting instructions.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov,

and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

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*.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6111–6120. Association for Computational Linguistics.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Jiaxin Guo, Minghan Wang, Daimeng Wei, Hengchao Shang, Yuxia Wang, Zongyao Li, Zhengzhe Yu, Zhanglin Wu, Yimeng Chen, Chang Su, Min Zhang, Lizhi Lei, Shimin Tao, and Hao Yang. 2021. Self-distillation mixup training for non-autoregressive neural machine translation. *CoRR*, abs/2112.11640.

Jiaxin Guo, Hao Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, and Xiaoyu Chen. 2024. A novel paradigm boosting translation capabilities of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 639–649. Association for Computational Linguistics.

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.

Wenxiang Jiao, Jen tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Parrot: Translating during chat using large language models tuned with human translation and feedback.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3505–3506. ACM.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2020. Capturing longer context for document-level neural machine translation: A multi-resolutional approach. *Cornell University - arXiv,Cornell University - arXiv*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Minghan Wang, Jiaxin Guo, Yuxia Wang, Yimeng Chen, Chang Su, Daimeng Wei, Min Zhang, Shimin Tao, and Hao Yang. 2021. HI-CMLM: improve CMLM with hybrid decoder input. In *Proceedings of the 14th International Conference on Natural Language Generation, INLG 2021, Aberdeen, Scotland, UK, 20-24 September, 2021*, pages 167–171. Association for Computational Linguistics.

Daimeng Wei, Zhanglin Wu, Hengchao Shang, Zongyao Li, Minghan Wang, Jiaxin Guo, Xiaoyu Chen, Zhengzhe Yu, and Hao Yang. 2023. Text style transfer back-translation.

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*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages.

Zhengzhe Yu, Jiaxin Guo, Minghan Wang, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu, Yuxia Wang, Yimeng Chen, Chang Su, Min Zhang, Lizhi Lei, Shimin Tao, and Hao Yang. 2021. Joint-training on symbiosis networks for deep nueral machine translation models. *CoRR*, abs/2112.11642.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023. Tim: Teaching large language models to translate with comparison.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models.