SJTU LoveFiction's System for WMT24 Discourse-Level Literary Translation

Haoxiang Sun[1]* **Tianxiang Hu**[1] **Ruize Gao**[2] **Jialong Tang**[2] **Pei Zhang**[2]
Baosong Yang[2] **Rui Wang**[1]
[1]Shanghai Jiao Tong University, Shanghai, China
[2]Alibaba Group, Hangzhou, China
{sunny_sjtu,hutianxiang,wangrui12}@sjtu.edu.cn
{xiaoyi.zp,yangbaosong.ybs}@alibaba-inc.com

Abstract

This paper describes Shanghai Jiao Tong University (SJTU LoveFiction) Discourse-Level Literary Translation systems for the WMT24 shared task. We participate in the literary translation task on Chinese → English, Chinese → German and Chinese → Russian with unconstrained tack. Our system is based on Qwen2-72B(Yang et al., 2024), Claude3.5(Anthropic, 2023) and GPT-4o(OpenAI, 2024) with novel techniques that improve literary translation performance on the target language pairs. (1) Chunk-based SFT and inference: we put several sentences together to form a chunk and try different chunksize during SFT and inference. (2) Merge multi-model translations by agents: we design a Translation Editor Agent based on GPT-4o to generate a better new translation by referencing the source text and merge 3 candidate translations generated by Qwen2-72B, Claude-3.5 and GPT-4o. (3) Terminology Intervention: to ensure terminology consistency, a Term Proofreader Agent, based on GPT-4o, is utilized to extract term pairs from source texts and translations. For each Chinese term, we decide its optimal translation and request the Term Proofreader to modify the translation generated by Translation Editor Agent. In model evaluation:(1) We employ d-BLEU for single model evaluation. (2) We design a Client Agent based on Claude-3.5 to assess the win-tie rate between two translations for cross-model evaluation.

1 Introduction

Despite great advancements in machine translation (MT) these years(Artetxe et al., 2017; Wang et al., 2022), achieving high-quality translations for literary texts remains a formidable task, primarily due to the complexities involved in maintaining coherence, consistency, and cultural context across larger text spans (Voita et al., 2019; Lopes et al., 2020).

This paper describes SJTU LoveFiction's submission to WMT24 Discourse-Level Literary Translation. We participate in all 3 language pairs (Chinese → English, Chinese → German and Chinese → Russian) with unconstrained tack.

Our system builds upon Qwen2-72B, Claude-3.5 and GPT-4o models with various practical techniques. We adopt a chunk-based strategy, grouping several sentences into a chunk during supervised fine-tuning (SFT) and inference phase.

Multi-agent structure demonstrates strong performance in discourse-level machine transaltion(Wu et al., 2024). To enhance translation quality, we develop a Translation Editor Agent based on GPT-4o. This agent references the source text and merges muliti-model translations to produce a refined output. While different models may generate varied translations for the same Chinese term, we also implement a Term Proofreader Agent powered by GPT-4o. This agent extracts term pairs from source text and corresponding translations. For each Chinese term, the optimal translation is determined manually, then the term proofreader applies these optimal terms to the merged translations.

In terms of evaluation, we use d-BLEU to assess the performance of a single model under different experimental settings. For cross-model evaluation, we design a Client Agent based on Claude-3.5. This agent references the Chinese source text to evaluate and rank the translations produced by different models by accuracy, fluency, and the preservation of stylistic elements.

This paper is structured as follows: Section 2 describes our data pre-processing strategies, followed by the details of our method in Section 3. Section 4 presents the experimental results and analysis, then we draw conclusions in Section 5.

*Work done during internship at Alibaba Group

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 967–972
November 15–16, 2024.

2 Data Processing

We perform Supervised Fine-Tuning (SFT) on the GuoFeng Webnovel Corpus (Wang et al., 2023) . Handling the noise within the dataset is crucial as it can significantly impact LLM's translation performance. We adopt a series of rigorous data filtering strategies.

2.1 Chinese-English Data Filtering

1. **Remove lines without Chinese-English pairs**: Delete any line that contains only a single Chinese or English sentence.
2. **Eliminate garbled text, emojis, foreign language characters, and emoticons**: These elements can degrade model performance. We use Unicode range identification and regular expressions for precise removal.
3. **Delete lines containing only punctuation marks**: Such lines typically lack linguistic value and retaining them would introduce noise, thereby impairing model training.
4. **Standardize punctuation**: Convert all Chinese punctuation to English punctuation to enhance model consistency and coherence in translation results.

2.2 Chinese-German and Chinese-Russian Data Filtering

Chinese-German/Russian data has the following features.

1. **Chapter-Level Alignment Only**: The alignment is maintained only at chapter level. Within chapters, paragraph or sentence level alignment is not achieved.
2. **Chapter Containment Differences**: In the Chinese files, each file contains a single chapter. In contrast, the German and Russian files may contain multiple chapters per file.

The following filtering strategies are employed:

1. **Remove Unaligned Chapter Pairs**: Delete Chinese-German/Russian file pairs that are not aligned at the chapter level.
2. **Eliminate garbled text, emojis, foreign language characters, and emoticons.**
3. **Remove Chapters Exceeding 8k Tokens**: LLMs struggle with long passages, thus chapters exceeding 8k tokens are excluded.

3 Method

In this section, we describe our method and provide a comprehensive explanation of the key components.

3.1 System Overview

We depict the overview of our system in Figure 1, which can be divided into four steps:

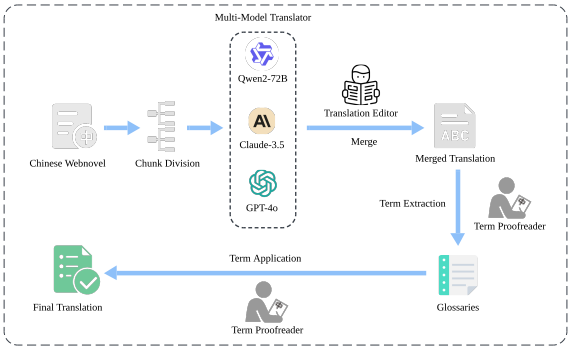

Figure 1: System Overview

1. **Chunk Division**: To maintain contextual information, we combine several sentences into a single chunk.
2. **Supervised Fine-Tuning & 1-shot Inference for Multi-Model Translator**: We SFT the Qwen2-72B model on Guofeng Webnovel Corpus. Afterwards, we use the fine-tuned Qwen2-72B, Claude-3.5, and GPT-4o to perform 1-shot inference on the test set, generating translation results.
3. **Translation Merging**: We employ a Translation Editor Agent based on GPT-4o to merge the translation outputs of the three models.
4. **Terminology Intervention**: We utilize a Term Proofreader Agent based on GPT-4o to extract term glossaries from source texts and translations. We select the optimal term pairs manually and ask the term proofreader to apply them to the merged translation as the final output.

3.2 Chunk Division

The lack of contextual information in sentence-level data poses a significant challenge for achieving high-quality translation results. Combining multiple sentences within each chapter into chunks can alleviate this problem(Zhao et al., 2023). During the SFT phase, we experiment with various

chunksizes of 5, 10, and 20 sentences to determine the optimal size for training. In the inference phase, we further extend our experiments to chunksizes of 1, 5, 10, 20, 40, and 80 sentences. This strategy aims to provide the model with more contextual information, thereby improving the translation quality.

3.3 Supervised Fine-Tuning (SFT) & 1-shot Inference

In order to find the best setting for Qwen2-72B, we SFT Qwen2-7B on the Guofeng Webnovel Corpus and conduct inference on the in-domain dev set. Given consistent distribution between the two datasets, this approach will reveal the best setting for LLM to learn the knowledge embedded in Guofeng Webnovel Corpus. d-BLEU scores under different settings are shown in Figure 2.

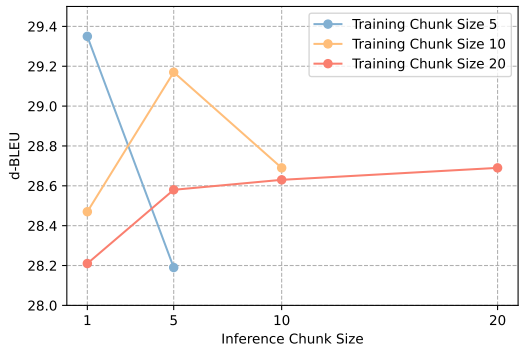

Figure 2: d-BLEU for Qwen2-7B on In-domain Dev Set

Although training with 5-sentence chunks and inferring with 1-sentence chunks yields the highest d-BLEU score of 29.35, we prefer 10-sentence chunks for training and 5-sentence chunks for inference. This configuration, with a d-BLEU score of 29.17, maintains nearly equivalent performance while preserving contextual information during inference.

As we aim to capture more context, the model must handle longer inputs. However, LLM's ability to handle long inputs is inherently limited. It's essential to acknowledge that we need to strike a balance, i.e. **maintaining sufficient contextual information without exceeding the model's capacity for processing long inputs.**

In our inference experiments with Claude-3.5 and GPT-4o, we employed 1-shot inference, a form of few-shot learning. Few-shot learning aims to enable models to generalize from a limited number of examples(Brown et al., 2020).We determine the best inference chunksize according to the following results.

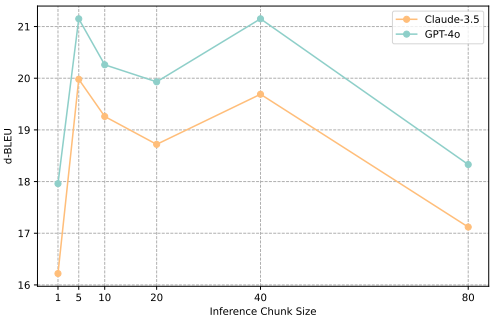

Figure 3: d-BLEU for Claude-3.5 & GPT-4o on OOD Dev Set

Both models perform best with 5-sentence chunk size, achieving d-BLEU scores of 19.98 and 21.15, respectively. We choose 5-sentence chunksize as the inference setting for Claude-3.5 and GPT-4o.

3.4 Translation Merging

After obtaining multi-model translations, we randomly select a chapter for manual verification and observe that different models exhibit distinct strengths in their translations for the same chunk. **To leverage the advantages of all three translations, we employ a Translation Editor agent based on GPT-4o, which is prompted to merge the three candidate translations into an improved version.** Workflow of the Translation Editor is as follows.

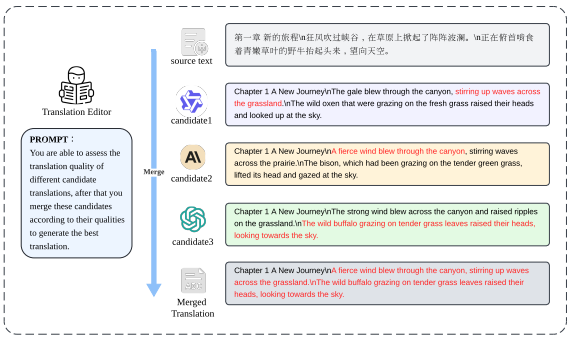

Figure 4: Workflow of The Translation Editor

1. **Quality Assessment.** Assess the quality of different translation referencing the source text. After this step, the agent knows the relatively better part in each translation.

2. **Translation Merging.** Put these parts together to form the merged translation.

This process allows the Translation Editor agent to integrate the best elements (highlighted in red in Figure 4) of the three candidate translations, generating a superior translation.

3.5 Terminology Intervention

While the Translation Editor agent generates improved results by blending three candidate translations, **different models may produce different translations for the same Chinese term, leading to consistency issues.** To address this, we develop a **Term Proofreader Agent**. Workflow of the agent is as follows.

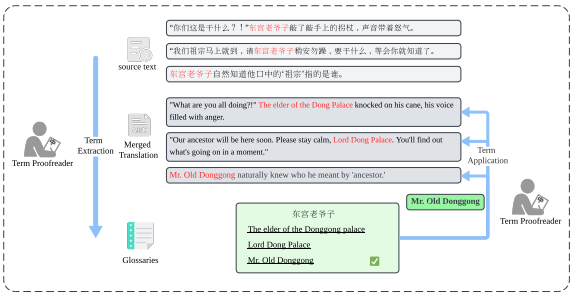

Figure 5: Workflow of The Term Proofreader

1. **Term Extraction.** The terminology proofreader agent begins by extracting term pairs from the Translation Editor's output, referencing its Chinese source. Glossaries are obtained after this step.

2. **Manual Determination.** For each Chinese term in the glossaries, we manually determine the optimal translation. This step involves reviewing the context and ensuring that the chosen translation accurately reflects the meaning and nuance of the original term.

3. **Term Application.** Once the optimal translations are determined, the terminology proofreader agent applies these optimal translations to Translation Editor's output.

3.6 Evaluation

3.6.1 Single Model Evaluation

We calculate d-BLEU scores between our translations and reference texts to evaluate single model performance and determine the optimal experimental settings (i.e. training & inference chunksize). d-BLEU measures N-gram matching, reflecting the similarity between two distributions.The distribution of the in-domain dev set and the train set are consistent. Thus d-BLEU can assess the model's learning of train set during SFT stage, enabling us to select the optimal SFT setting by d-BLEU. On the other hand, distribution of the ood dev set is inconsistent with the train set. d-BLEU can assess the model's fitting to the ood dev set distribution. Thus we can select the optimal inference setting by d-BLEU.

3.6.2 Cross-model Evaluation

For cross-model evaluation, we find that human-preferred translation can have low d-BLEU score. This discrepancy arises because d-BLEU relies solely on N-gram matching and is unable to capture deeper semantic information. For human-preferred translation, there can be significant lexical differences from the reference translations, even though the semantic content is accurately conveyed. d-BLEU is ineffective in evaluating such cases.

Previous works reveal that LLM-Evaluators can achieve high consistency with human expert on system-level evaluation(Kocmi and Federmann, 2023; Moosa et al., 2024). We build a Client Agent based on Claude-3.5, which considers accuracy, fluency, and the preservation of stylistic elements.

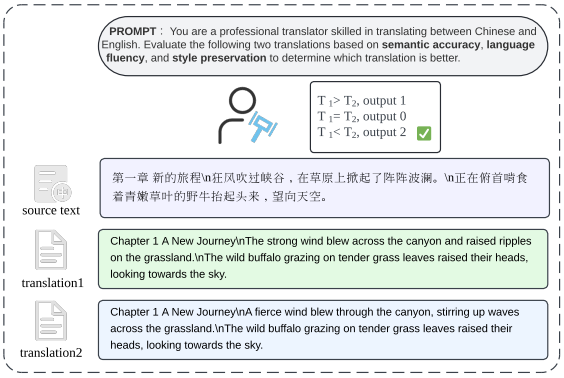

Figure 6: Workflow of The Client

3.6.3 Human Evaluation

We employ 3 language experts to do fine-grained evaluation. They are requested to perform Linguistic Quality Rating (LQR) by the following standard in Table 1.

4 Results

We present the effect of our method in this section.

Score	Quality Description
1	Incomprehensible or incorrect.
2	Severe errors, hard to understand.
3	Some errors, but understandable.
4	Mostly correct, minor errors.
5	Completely correct and fluent.

Table 1: LQR Scoring Standards

4.1 Supervised Fine-Tuning (SFT) & Inference

We train Qwen2-72B on GuoFeng Webnovel Corpus with 10-sentence chunksize. The following table shows the d-BLEU scores for various inference chunksizes on both in-domain and out-of-domain dev sets.

Inference Chunksize	In-domain	OOD
1	27.32	22.51
5	27.05	24.64
10	26.05	**24.74**
20	27.74	24.65
40	**28.11**	24.25
80	20.49	24.04

Table 2: d-BLEU of Qwen2-72B on In-domain & OOD Dev Set

Qwen2-72B achieves best performance under 40-sentence inference chunksize on in-domain dev set while the best performance on OOD dev set is achieved with the 10-sentence chunksize. This indicates that although Qwen2-72B has a stronger capability for handling long texts, out-of-domain data distribution still poses difficulties for translation.

Results for Claude-3.5 and GPT-4o on OOD dev set is in Figure 3.

4.2 Translation Merging

We randomly selected 200 chunks from the final test set to evaluate the performance of individual models and our translation merging strategy.

GPT-4o ranks 1st place in single model performance while our translation merging strategy surpasses every single model, indicating that better translation is generated by the Translation Editor Agent.

Model	LQR3	LQR4	LQR5
Translation Merging	**65%**	**44%**	**24%**
GPT-4o	60%	30%	9%
Claude-3.5	54%	33%	12%
Qwen2-72b	42%	24%	3%

Table 3: LQR Scores for Different Models

We also employed the Client Agent to compare GPT-4o's results and the merged translations. Table 4 presents the win-tie rate relative to GPT-4o.

Metric	Rate
Win	41%
Tie	21%
Lose	38%
Net Win Rate	**3%**

Table 4: Win-tie Rate Compared to GPT-4o

The LLM evaluator also acknowledges that our translation merging strategy brings a slight improvement.

4.3 Terminology Intervention

We employ the Term Proofreader Agent to extract term pairs from the entire test set. The following table presents the results before and after the terminology intervention.

	Before	After
Chinese Terms	806	806
English Translations	3012	902
Average Correspondence	3.73	1.12

Table 5: Term Correspondence Before and After Intervention

Before the intervention, 806 unique Chinese terms correspond to 3012 English translations, with an average of 3.73 English translations per Chinese term, indicating high variability and inconsistency.

After the intervention, the number of English translations is reduced to 902. This significant reduction demonstrates that the Term Proofreader Agent effectively standardized the terminology, ensuring consistent translations for each Chinese term.

5 Conclusion

Through chunk splitting, multi-model translation merging, and terminology intervention, our system demonstrates strong performance in the WMT24 Discourse-Level Literary Translation task. The translation merging strategy surpasses all individual models in LQR scores. Terminology intervention significantly improves terminology consistency, reducing the average correspondence from 3.73 translations to 1.12. Future work will focus on further optimizing these techniques and exploring new strategies to enhance translation quality, especially in handling long texts and preserving literary styles.

References

Anthropic. 2023. Claude 3.5.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Ibraheem Muhammad Moosa, Rui Zhang, and Wenpeng Yin. 2024. Mt-ranker: Reference-free machine translation evaluation by inter-system ranking. In *The Twelfth International Conference on Learning Representations*.

OpenAI. 2024. Gpt-4o.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2022. Progress in machine translation. *Engineering*, 18:143–153.

Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, et al. 2023. Findings of the wmt 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of llms. In *Proceedings of the Eighth Conference on Machine Translation*, pages 55–67.

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

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Anqi Zhao, Kaiyu Huang, Hao Yu, and Degen Huang. 2023. Dutnlp system for the wmt2023 discourse-level literary translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 296–301, Singapore. Association for Computational Linguistics.