

HW-TSC’s Submission to the WMT 2026 General Machine Translation Track

Yu He, Xiaoqing Lan, Daimeng Wei, Jiaxin Guo, Yuanchang Luo,
Hengchao Shang, Zongyao Li, Jinlong Yang, Zhanglin Wu, Boqi Huang

Huawei Translation Service Center, Beijing, China

{heyu97, lanxiaoqing, weidaimeng, guojiaxin1, luoyuanchang1,
shanghengchao, lizongyao, yangjinlong7, wuzhanglin2, huangboqi1}@huawei.com

Abstract

Machine translation based on large language models (LLMs) has achieved remarkable progress in recent years. However, relying on a single translation model and fixed translation instructions often limits translation quality, especially for multilingual and multi-domain scenarios. To address this issue, we propose HW-TSC-Agent, a multi-stage translation agent framework that integrates translation model fine-tuning, dual-model candidate generation, and instruction-aware translation refinement into a unified pipeline. First, we perform Supervised Fine-Tuning (SFT) and Contrastive Preference Optimization (CPO) on openPangu-Embedded-7B using high-quality bilingual data to enhance its multilingual translation capabilities. The fine-tuned openPangu-Embedded-7B and HY-MT2-7B are then used to generate candidate translations. Finally, source-language analysis, candidate translations, and task-specific instructions are incorporated into a dynamic prompt, which guides DeepSeek-V4-Flash to compare, refine, and fuse the candidate translations to produce the final output. Our system is submitted to the WMT2026 General Machine Translation Shared Task, covering 13 translation directions. On the WMT2025 English-to-Chinese development test set, enriching the dynamic prompt generally improves translation quality over an instruction-only DeepSeek-V4-Flash baseline.

1 Introduction

In recent years, large language models have demonstrated strong capabilities in contextual understanding, natural language generation, and cross-lingual knowledge transfer, driving substantial advances in machine translation [Zhang et al. \(2026\)](#) [Naveed et al. \(2025\)](#) [Raza et al. \(2025\)](#). Models such as GPT-4 [Achiam et al. \(2023\)](#), Llama 3 [Grattafiori et al. \(2024\)](#), Qwen3 [Yang et al. \(2025\)](#), and DeepSeek-V3 [Liu et al. \(2024\)](#) have consistently

achieved state-of-the-art performance across a wide range of multilingual understanding and generation benchmarks, exhibiting remarkable capabilities in semantic reasoning, long-context modeling, and instruction following. These advances have significantly improved translation quality by enabling models to better capture contextual dependencies, preserve semantic consistency, and generate more fluent and natural target-language text. Consequently, large language models have become a promising foundation for modern machine translation systems, particularly in document-level translation and context-aware translation scenarios.

However, relying solely on a single translation model and fixed translation instructions remains insufficient for consistently producing high-quality translations that are accurate, fluent, and well adapted to specific tasks [Zhang et al. \(2023\)](#). Different translation models often exhibit complementary strengths [Zeng et al. \(2023\)](#) [Xu et al. \(2024\)](#), while individual samples may impose varying requirements on semantic fidelity, stylistic consistency, and output format. Therefore, we adopt a multi-source translation knowledge fusion and instruction-aware refinement strategy, integrating candidate translations from multiple models, source-language context, and task-specific instructions to further improve the fluency and accuracy of LLM-based translation.

The Conference on Machine Translation (WMT) serves as an important platform for evaluating advances in machine translation. The WMT2026 General Machine Translation Shared Task aims to assess the overall translation capabilities of systems across multilingual, multi-domain, and multimodal settings, with the test data provided at the document level. This paper presents HW-TSC-Agent, our submission to the shared task, which covers 13 translation directions. Unlike approaches that rely on a single model to generate translations directly, we adopt a multi-stage translation agent framework

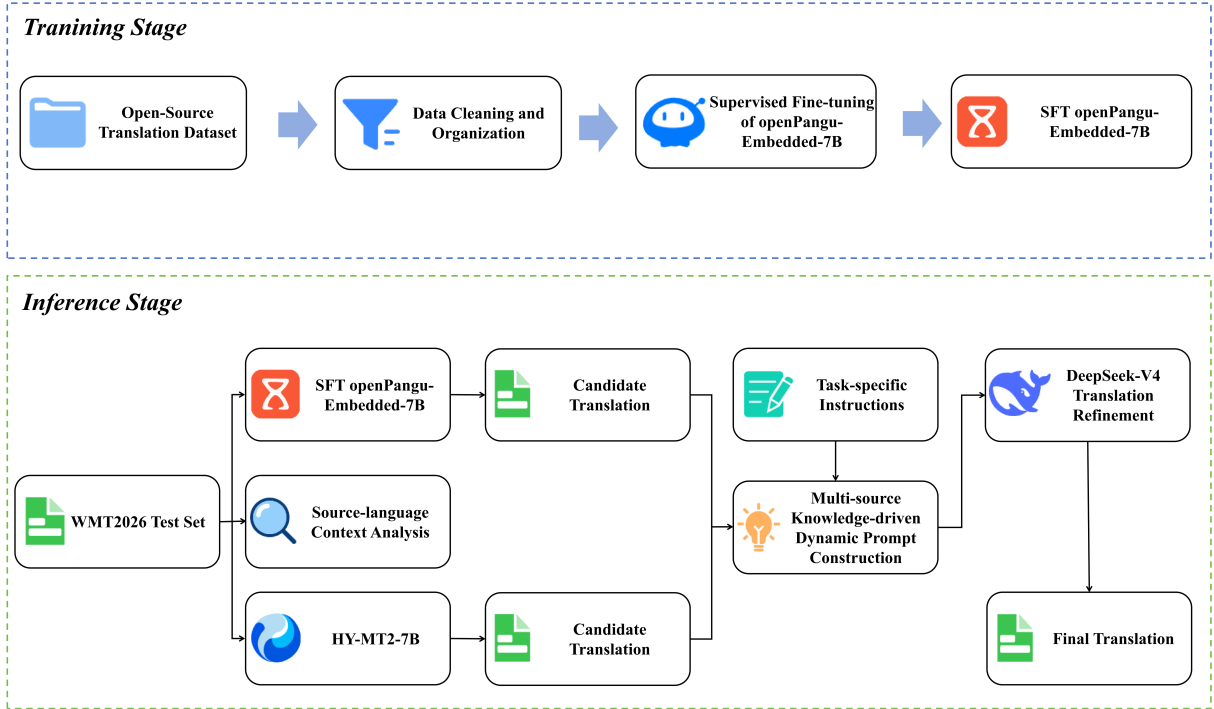


Figure 1: Overview of the HW-TSC-Agent framework, including openPangu-Embedded-7B fine-tuning (SFT and CPO), dual-model candidate generation, dynamic prompt construction, and DeepSeek-V4-Flash translation refinement.

based on large language models. Specifically, multiple models are first used to generate candidate translations, after which source-language analysis, candidate translation knowledge, and task-specific instructions are jointly incorporated into a dynamic prompt and provided to a refinement model.

We propose a multi-stage translation agent framework based on large language models to improve translation quality through three key stages: (1) fine-tuning of openPangu-Embedded-7B [Chen et al. \(2025\)](#) on high-quality bilingual data to enhance its general machine translation capabilities; (2) generating candidate translations using the fine-tuned openPangu-Embedded-7B and HY-MT2-7B [Zheng et al. \(2026\)](#) to obtain complementary translation knowledge; and (3) construction of a dynamic prompt by integrating source-language-aware context, dual-model candidate translations, and task-specific instructions, followed by the use of DeepSeek-V4-Flash to correct, fuse, and refine the candidate translations.

Our main contributions are summarized as follows:

(1) We propose a multi-stage translation agent framework based on large language models, integrating translation model fine-tuning, dual-model candidate generation, and translation refinement

into a unified translation pipeline.

(2) We propose a dynamic prompt construction method that incorporates multi-source translation knowledge by jointly leveraging source-language analysis, external model translations, and task-specific instructions to provide multidimensional constraints for final translation generation.

(3) We design an instruction-aware translation refinement mechanism that uses DeepSeek-V4-Flash to compare and revise candidate translations generated by different models. The final translation is then produced according to the document context, linguistic style, formatting requirements, and output specifications of each sample, thereby improving the system’s adaptability to complex general machine translation scenarios.

2 Methodology

2.1 System Overview

Our translation system, HW-TSC-Agent, consists of several stages, including model training, candidate generation, knowledge integration, and translation refinement. The overall pipeline comprises five steps: (1) cleaning, filtering, and organizing high-quality bilingual translation data to construct supervised fine-tuning data; (2) fine-tuning

of openPangu-Embedded-7B to enhance its general machine translation capabilities; (3) generating candidate translations using the fine-tuned openPangu-Embedded-7B and HY-MT2-7B to obtain complementary translation knowledge; (4) constructing a dynamic prompt driven by multi-source translation knowledge through the integration of source-language-aware context, dual-model candidate translations, and task-specific instructions; and (5) using DeepSeek-V4-Flash to correct, fuse, and refine the candidate translations. The overall architecture of the system is illustrated in Fig. 1.

2.2 Data Preprocessing

This work preprocesses the multilingual datasets officially released by WMT26. For the 13 submitted language pairs, we design a multi-stage preprocessing pipeline to remove low-quality data from the raw corpus and construct high-quality parallel training data. First, we randomly sample 5 GB of raw data for each language pair and perform an initial filtering step to remove obviously low-quality sentence pairs, such as those with formatting errors or empty content. To address common issues in Chinese corpora, including mixed simplified and traditional Chinese characters and inconsistent character variants, we use OpenCC to normalize the text and eliminate inconsistencies that may interfere with model training.

Second, we perform global and hash-based deduplication to remove identical and highly similar sentence pairs, reducing repeated learning and overfitting. To further improve semantic alignment, we embed each source and target sentence with LaBSE (Language-agnostic BERT Sentence Embedding) and compute the cosine similarity of the resulting bilingual representations. Pairs whose similarity falls below 0.8 are discarded as poorly aligned. The surviving pairs are then ranked by LaBSE similarity in descending order, and the top one million pairs are retained for each language pair as the final fine-tuning set.

2.3 openPangu-Embedded-7B Fine-tuning

To fully leverage multilingual data and improve model generalization, this work adopts a two-stage alignment training strategy consisting of Supervised Fine-Tuning (SFT) and Contrastive Preference Optimization (CPO) to enhance the multilingual capabilities of openPangu-Embedded-7B. Both stages are implemented with the ModelScope SWIFT (ms-swift) training framework.

The first stage is Supervised Fine-Tuning (SFT). We combine the training data from all language pairs and perform global shuffling to eliminate language-specific data distribution bias and improve cross-lingual semantic transfer. We then fine-tune openPangu-Embedded-7B on the mixed dataset for one epoch, enabling the model to learn shared translation patterns and semantic alignment across languages.

The second stage is Contrastive Preference Optimization (CPO). To construct high-quality preference pairs, we randomly sample 10,000 held-out examples from each language pair that are not used in SFT. We then use DeepSeek-V4-Flash [Xu et al. \(2026\)](#), Qwen2-72B [Yang et al. \(2024\)](#), and Qwen3-32B [Yang et al. \(2025\)](#) to generate multiple translations for each source sentence, forming multi-translation triplets. Next, we score the candidate translations with COMET20 and select the best translation as the positive sample and the worst as the negative sample. These preference pairs are used for CPO training, which guides the model toward better translation preferences through contrastive learning. This stage corrects issues from SFT, such as translation bias, unnatural phrasing, and semantic omissions, and improves multilingual translation accuracy, fluency, and robustness.

2.4 Multi-source Knowledge-aware Dynamic Prompt Construction

To improve the model’s understanding of complex translation scenarios and the quality of translation generation, we design a source-knowledge-aware dynamic prompt construction method. For each input sample, the method integrates nine types of information: a source-text summary, translation scenario identification, translation considerations, source-format preservation requirements, domain identification, style alignment, terminology extraction, external-model candidate translations, and task-specific translation instructions (Table 1). Unlike systems that rely on a fixed language-pair instruction alone, the prompt is assembled from the source content, domain characteristics, text format, and task requirements of the current sample, so that DeepSeek-V4-Flash can jointly exploit global document context, local translation constraints, complementary candidate translations, and official task instructions when producing the final output.

Concretely, the prompt is organized in three blocks. The first block is source-language-aware context obtained by prompting a large language

Translation Information	Main Function
Source-Text Summary	Summarize the main content of long texts to ensure document-level consistency.
Translation Scenario Identification	Identify the usage scenario and select an appropriate translation strategy.
Translation Considerations	Identify key translation points, ambiguities, and error-prone expressions.
Source-Format Preservation	Identify the source layout and preserve the corresponding format in the translation.
Domain Identification	Identify the domain and apply domain-specific terminology and translation conventions.
Style Alignment	Match the target style, such as formal, conversational, promotional, literary, or academic.
Terminology Extraction	Extract terms, named entities, and domain-specific abbreviations to ensure translation consistency.
External Model Knowledge	Incorporate translations from external models as complementary references.
User Translation Requirements	Inject task-specific translation rules and output requirements.

Table 1: Nine types of translation information used for source-knowledge-aware dynamic prompt construction.

model to analyze the source document along the first seven dimensions in Table 1. The summary captures the main topics and discourse relations of long documents and supports document-level consistency. Scenario identification determines the practical setting of the text and guides strategy selection. Translation considerations highlight polysemous words, coreference, puns, and other spans that are prone to mistranslation or omission. Format preservation recovers titles, lists, tables, and related layout cues so that the target side can mirror the source structure. Domain identification and style alignment supply domain conventions and the intended register (e.g., formal, conversational, promotional, literary, or academic), optionally with short exemplars. Terminology extraction lists technical terms, named entities, and domain abbreviations to encourage consistent rendering throughout the document.

The second block injects external model knowledge: the candidate translations from the fine-tuned openPangu-Embedded-7B and HY-MT2-7B. Because the two models differ in lexical choice, terminology handling, completeness, and fluency, agreement between them provides a high-confidence reference, whereas disagreement encourages the refiner to re-examine the corresponding source span.

The third block injects user translation require-

ments from the WMT2026 test set. We focus on spoken dialogue and social media. Spoken sources may contain ASR errors, so the prompt asks the model to recover intended meaning while preserving conversational tone. Social-media sources derived from screenshots may contain spelling noise, so the prompt asks the model to correct surface errors while retaining emotional cues. Together, the three blocks turn the nine information types into a sample-specific instruction rather than a static template shared by an entire language pair.

2.5 Translation Refinement

DeepSeek-V4-Flash receives the assembled dynamic prompt and produces the final translation by comparing, revising, and fusing the two candidates under the source-analysis and task constraints. The refiner is instructed to prefer information that is supported by the source, to preserve required formatting, and to follow scenario-specific rules when applicable. The DeepSeek-V4-Flash output is used as the submitted translation.

3 Experiments

3.1 Experimental Setup

Submission coverage. Our WMT2026 submission covers 13 translation directions: en→{ar, cs, id,

ja, kk, ko, th, uk, zh, zh_TW}, cs→{uk, vi}, and zh→ja.

Development evaluation. We report automatic scores on the WMT2025 General Machine Translation English-to-Chinese test set as a development benchmark for comparing prompt variants under a fixed DeepSeek-V4-Flash refinement backbone. Translation quality is evaluated with COMETKiwi-XL Rei et al. (2023) and XCOMET-XL Guerreiro et al. (2024).

Inference setup. All candidate-generation and refinement models are served through self-hosted deployments.

3.2 Main Results

Table 2 presents the evaluation results of the DeepSeek-V4-Flash baseline and four Agent variants on the WMT2025 English-to-Chinese test set. The baseline uses only the official translation instructions as the prompt, achieving COMETKiwi-XL and XCOMET-XL scores of 0.5581 and 0.3674, respectively. After introducing seven types of source-language-aware information, Agent-v1 improves the COMETKiwi-XL score to 0.5669. Agent-v2 and Agent-v3 further incorporate ASR error correction requirements for spoken transcripts and spelling correction requirements for social media texts, respectively, leading to additional performance gains. Among them, Agent-v3 achieves the best COMETKiwi-XL score of 0.5706, while its XCOMET-XL score increases to 0.3699.

Model	COMETKiwi-XL	XCOMET-XL
DeepSeek-V4-Flash	0.5581	0.3674
Agent-v1	0.5669	0.3624
Agent-v2	0.5700	0.3664
Agent-v3	0.5706	0.3699
Agent-v4	0.5670	0.3776

Table 2: Development results on the WMT2025 English-to-Chinese test set. The best results are highlighted in bold.

Agent-v4 further incorporates explicit contextual understanding cues, achieving an XCOMET-XL score of 0.3776, which is 0.0102 higher than the baseline. Its COMETKiwi-XL score is 0.5670, which is lower than that of Agent-v3 but still exceeds the baseline. Prompt enrichment generally improves at least one of the two metrics, although the variants exhibit different trade-offs between COMETKiwi-XL and XCOMET-XL. The two metrics respond differently to the added prompt components, suggesting that prompt enrichment does

not uniformly improve all dimensions of translation quality.

4 Conclusion

This paper presents HW-TSC-Agent, our system developed for the WMT2026 General Machine Translation Shared Task across 13 directions. We propose a multi-stage translation agent framework that combines supervised fine-tuning and preference optimization of openPangu-Embedded-7B, dual-model candidate generation, dynamic prompt construction with multi-source translation knowledge, and DeepSeek-V4-Flash refinement. On the WMT2025 English-to-Chinese development set, enriching the dynamic prompt improves COMETKiwi-XL and XCOMET-XL relative to an instruction-only DeepSeek-V4-Flash baseline for the strongest prompt variants.

Acknowledgments

We thank the WMT2026 organizers for the task design and data, and colleagues at HW-TSC for discussion and engineering support.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Hanting Chen, Yasheng Wang, Kai Han, Dong Li, Lin Li, Zhenni Bi, Jinpeng Li, Haoyu Wang, Fei Mi, Mingjian Zhu, et al. 2025. Pangu embedded: An efficient dual-system llm reasoner with metacognition. *arXiv preprint arXiv:2505.22375*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

- Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.
- Mubashar Raza, Zarmina Jahangir, Muhammad Bilal Riaz, Muhammad Jasim Saeed, and Muhammad Awais Sattar. 2025. Industrial applications of large language models. *Scientific Reports*, 15(1):13755.
- Ricardo Rei, Nuno M Guerreiro, José Pombal, Daan Van Stigt, Marcos Treviso, Luisa Coheur, José GC de Souza, and André FT Martins. 2023. Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848.
- Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chencheng Ling, et al. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*.
- Yangyifan Xu, Jinliang Lu, and Jiajun Zhang. 2024. Bridging the gap between different vocabularies for llm ensemble. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7140–7152.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023. Improving machine translation with large language models: A preliminary study with cooperative decoding. *arXiv preprint arXiv:2311.02851*.
- Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International conference on machine learning*, pages 41092–41110. PMLR.
- Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, et al. 2026. Instruction tuning for large language models: A survey. *ACM Computing Surveys*, 58(7):1–36.
- Mao Zheng, Zheng Li, Tao Chen, Bo Lv, Mingrui Sun, Mingyang Song, Jinlong Song, Hong Huang, Decheng Wu, Hai Wang, et al. 2026. Hy-mt2: A family of fast, efficient and powerful multilingual translation models in the wild. *arXiv preprint arXiv:2605.22064*.