

TransCore: The NewTranx Submission to the Constrained Track of the WMT26 General Machine Translation Shared Task

Hao Yu¹, Ziyan Chen², Hongyu Zhao², Yulong Zhang², Xiaogang Wang²,
Kaiyu Huang², Degen Huang^{1,3†}

¹Dalian University of Technology, ²NewTranx Technology Co., Ltd.,
³Fuyao University of Science and Technology

Abstract

This paper describes TransCore, the NewTranx submission to the WMT26 General Machine Translation Shared Task constrained track. Our system is built upon the Qwen3-8B large language model and adopts a two-stage supervised fine-tuning framework trained exclusively on publicly available data. Our approach focuses on high-quality multilingual data construction, including parallel corpus construction, knowledge-distilled corpus construction, and instruction-following data construction. For the parallel corpus, we construct a high-quality multilingual training corpus through multilingual data collection and automatic quality filtering. For knowledge distillation, we construct a KD corpus from publicly available monolingual resources using a large language model, followed by automatic quality filtering. To support the instruction-following evaluation introduced in WMT26, we further construct instruction data covering immutable entity preservation, document-level translation, and structured translation, which are incorporated into the first-stage supervised fine-tuning. Our system participates in four language directions of the WMT26 General Machine Translation Shared Task: English→Indonesian, English→Kazakh, English→Thai, and Chinese→Japanese.

1 Introduction

Recent advances in large language models (LLMs) have significantly improved machine translation performance and accelerated the adoption of instruction-tuned LLMs for multilingual machine translation (Li et al., 2024; Rei et al., 2026; Cheng et al., 2025; Luo et al., 2025; Zheng et al., 2025). Benefiting from large-scale pre-training and strong cross-lingual generalization capabilities, LLMs have achieved competitive translation quality across a wide range of language

pairs while demonstrating superior instruction-following capability in practical translation scenarios. Consequently, supervised fine-tuning of open-source LLMs has become an increasingly effective paradigm for developing high-quality multilingual machine translation systems.

Compared with previous editions, the WMT26 General Machine Translation Shared Task (Kocmi et al., 2026) introduces several important changes. The evaluation covers multiple domains, including news, social, and spoken content, while all test sets are provided at the document level. Furthermore, the task introduces instruction-following evaluation through instance-level translation instructions, requiring translation systems to satisfy diverse user-specified translation constraints. More importantly, WMT26 adopts human evaluation as the primary evaluation protocol for all submitted systems, reflecting an increasing emphasis on translation quality in realistic application scenarios rather than optimization toward automatic evaluation metrics. These changes further highlight the importance of constructing high-quality training data for LLM-based machine translation systems.

To address these challenges, we present TransCore, a translation system based on Qwen3-8B (Yang et al., 2025) using a two-stage supervised fine-tuning framework. To better support the instruction-following evaluation in WMT26, we additionally construct dedicated instruction data covering immutable entity preservation, document-level translation, and structured translation. The resulting system is submitted to four WMT26 General Machine Translation directions, English→Indonesian, English→Kazakh, English→Thai, and Chinese→Japanese. The main contributions of this work are summarized as follows:

- We present TransCore, a constrained-track multilingual machine translation system built

[†]Corresponding author.

upon Qwen3-8B for the WMT26 General Machine Translation Shared Task.

- We construct a high-quality training corpus by integrating carefully curated parallel data, knowledge-distilled (KD) data, and instruction-following data under a unified training framework.
- We design a unified instruction-following data construction framework that supports multiple translation constraints, including immutable entity preservation, document-level translation, and structured translation, through a unified instruction representation and data construction pipeline.

2 Method

2.1 Overview

Our systems, TransCore, adopts a two-stage supervised fine-tuning (SFT) pipeline for machine translation. The overall workflow consists of parallel data construction, knowledge-distilled data construction, instruction-following data construction, two-stage model training, and deterministic inference.

In the first stage, publicly available parallel corpora are collected for all translation directions. The bilingual data are preprocessed, automatically filtered using quality estimation, and converted into a unified instruction-following format for supervised fine-tuning. In the second stage, we construct knowledge-distilled parallel data from publicly available monolingual corpora spanning multiple domains. A teacher translation model is employed to translate target-language texts into the corresponding source language, and the resulting knowledge-distilled data are further filtered before being used for continued supervised fine-tuning.

To improve instruction-following capability under the WMT26 General MT setting, we additionally construct instruction-following translation data covering three representative translation capabilities: preservation of non-translatable entities and formatting elements, document-level translation with contextual consistency, and structured translation with tag and XML structure preservation. Finally, the multilingual model is trained through two-stage full-parameter supervised fine-tuning and evaluated using deterministic decoding.

Language Pair	Original	Selected
en→id	1,402,203	450,207
en→kk	174,667	63,007
en→th	1,104,108	191,008
zh→ja	782,353	55,843
Total	3,463,331	760,065

Table 1: Statistics of the first-stage parallel corpus.

2.2 Stage I: Parallel Data Construction

The first-stage training data are constructed from publicly available parallel corpora covering the four translation directions evaluated in our submission. Multiple open-source datasets are collected for each language pair, including news, subtitles, educational resources, encyclopedic text, software localization, and benchmark datasets. Since the quality of publicly available parallel corpora varies substantially across sources, we perform a unified preprocessing and quality filtering pipeline before model training.

Data Preprocessing We first conduct basic preprocessing to ensure sentence-level alignment and remove invalid training instances. This step includes alignment checking, removal of empty sentence pairs and control characters, normalization of text formatting, and language-specific cleaning for noisy resources such as subtitle corpora. Exact deduplication is then performed independently within each dataset to remove duplicated bilingual sentence pairs and duplicated monolingual entries introduced by corpus merging while preserving the original data order.

Data Filtering After preprocessing, all sentence pairs are scored using COMETKiwi-XL quality estimation ([wmt23-cometkiwi-da-xl](#)). Following the characteristics of different datasets, we adopt either language-pair-specific score thresholds or dataset-level top-ranking selection to retain high-quality translation pairs. Software localization datasets are filtered by retaining the highest-scoring half of the samples within each dataset split, while other datasets are filtered using fixed COMETKiwi-XL thresholds selected for each translation direction. The statistics of the resulting first-stage parallel corpus after quality filtering are summarized in Table 1. Finally, the retained bilingual sentence pairs are converted into a unified instruction-following dialogue format for supervised fine-tuning. The public datasets used for each translation direction are

Category	Dataset	en→id	en→kk	en→th	zh→ja
TED	Neulab TED Talks	✓	✓	✓	✓
	OPUS TED2020	v1	v1	v1	v1
News	News Commentary	v18	v18	N/A	v18
	OPUS GlobalVoices	v2018q4	N/A	N/A	N/A
Subtitle	OPUS OpenSubtitles	v2024	v2024	v2024	v2016
Software	OPUS GNOME	v1	v1	v1	N/A
	OPUS KDE4	v2	v2	v2	N/A
	OPUS Ubuntu	v14.10	v14.10	v14.10	v14.10
	OPUS PHP	N/A	N/A	N/A	v1
Education	OPUS QED	v2.0a	v2.0a	v2.0a	v2.0a
Encyclopedia	OPUS Wikimedia	v20230407	v20230407	v20230407	v20230407
Benchmark	FLORES200 (Costa-Jussà et al., 2022)	dev	dev	dev	dev
	NTREX-128 (Aepli et al., 2023)	✓	✓	✓	✓
	WMT Development Sets	N/A	✓	N/A	✓

Note: Unless otherwise specified, all parallel corpora are downloaded using the official [WMT26 constrained-track mtdata](#) recipe. Benchmark datasets are obtained from their respective official sources.

Table 2: Public parallel datasets used for first-stage training.

Category	Dataset	en→id	en→kk	en→th	zh→ja
News	Indonesian News Datasets	✓	–	–	–
	Kazakh News Corpus	–	✓	–	–
	News Commentary v18.1	✓	✓	–	✓
	ThaiSum	–	–	✓	–
Social Media	IndoDiscourse	✓	–	–	–
	KazSANRA	–	✓	–	–
	Wisesight Sentiment	–	–	✓	–
Spoken Language	Indonesian Voice Transcription	✓	–	–	–
	Kazakh Speech Corpus 2 (KSC2)	–	✓	–	–
	LOTUSDIS	–	–	✓	–
Image Caption	DEJIMA Caption	–	–	–	✓

Table 3: Public monolingual datasets used for KD data construction.

summarized in Table 2. In addition to the filtered parallel corpora, the first-stage training data also include a collection of general translation instruction-following examples to improve the model’s ability to follow translation requests with diverse instructions. The final first-stage corpus contains approximately 760k high-quality parallel sentence pairs together with 10.7k general translation instruction-following samples.

2.3 Stage II: Knowledge Distillation Data Construction

To further improve translation quality, we construct a knowledge-distilled parallel corpus from publicly available target-language monolingual datasets. The collected corpora cover diverse domains, including news, news commentary, online discussions, social media, spoken language, meeting conversations, and image captions. The mono-

lingual datasets used for each translation direction are summarized in Table 3.

Field Selection and Data Preparation Public monolingual datasets typically contain multiple fields, including metadata, labels, multimedia content, and natural language text. We retain only text fields suitable for translation, such as article content, summaries, comments, transcripts, and image captions, while discarding non-textual information. The extracted texts are subsequently processed through text normalization, sentence segmentation for long documents or paragraphs, language filtering, duplicate removal, and domain-aware sampling to construct candidate monolingual corpora for knowledge distillation. The preprocessing strategy is adapted to the characteristics of each dataset while preserving high-quality natural language content.

Instruction Category	Translation Constraint	Data Source
Immutable Entity Preservation	Preserve predefined entities verbatim	Do Not Change Me
Document-level Translation	Utilize document-level context	BWB , mZPRT , PDC
Terminology-constrained Translation	Soft / Hard terminology constraints	WMT21 Terminology , WMT23 Terminology
Structured Translation	Preserve XML / HTML structure	Salesforce , Amazon MT
Formula Preservation	Preserve mathematical and chemical formulas	latex-formulas , VNHSGE

Table 4: Instruction-following data constructed in this work.

Language Pair	Candidates	Final
en→id	20,000	19,955
en→kk	20,000	19,989
en→th	20,000	19,963
zh→ja	19,978	19,962
Total	79,978	79,869

Table 5: Statistics of the KD corpus.

Configuration	Value
Training Strategy	Full SFT
Optimizer	AdamW
Learning Rate	1×10^{-5}
Learning Rate Scheduler	Cosine
Warmup Ratio	0.03
Weight Decay	0.01
Batch Size	128
Epochs	2
Max Sequence Length	2048
Precision	BF16
Optimization	DeepSpeed ZeRO-2
Hardware	$8 \times$ NVIDIA A100 80GB

Table 6: Training configuration.

Knowledge Distillation For each translation direction, a teacher translation model is employed to translate the prepared target-language texts into the corresponding source language. The generated source-language translations are then paired with the original target-language texts to construct the KD corpus. Finally, all translation pairs are converted into the unified instruction-following dialogue format adopted in Stage I for subsequent supervised fine-tuning.

Data Filtering To improve the quality of the KD corpus, we further filter the generated translation pairs. For the English-target translation directions, copied translations and highly similar source-target pairs are removed. For the Chinese-Japanese translation direction, abnormal generation patterns, such as refusal responses and incomplete outputs, are discarded. We additionally apply Qwen3 tokenizer-based length filtering by removing sentence pairs whose source or target side exceeds 512 tokens.

The resulting KD corpus is used for the second-stage supervised fine-tuning, and its statistics are summarized in Table 5.

2.4 Instruction-following Data Construction

To improve the instruction-following capability of the translation model, we construct a unified instruction-following translation corpus covering multiple constrained translation scenarios. During data development, five instruction categories are designed, including immutable entity preservation, document-level translation, terminology-constrained translation, structured translation, and formula preservation. Following the requirements of the WMT26 General MT task, this paper focuses on immutable entity preservation, document-level translation, and structured translation, while the remaining instruction categories are omitted from the discussion. An overview of the constructed instruction-following data is presented in Table 4.

All instruction-following samples are represented using a unified instruction-input-output format. Each instruction consists of a task-specific constraint and a general translation instruction that specifies the translation direction. For both Chinese and English instructions, five semantically equivalent translation templates are designed and uniformly sampled using fixed random seeds to improve instruction diversity while maintaining reproducibility.

The instruction-following corpus is constructed from both publicly available bilingual datasets and automatically generated bilingual resources. For document-level translation, terminology-constrained translation, and structured translation, the original bilingual alignment is preserved throughout the construction process. For document-level translation, aligned documents are first segmented into pseudo-documents with a maximum length of 512 tokens while ensuring that each pseudo-document does not cross original document boundaries. To improve the model’s robust-

COMETKiwi-XL									
Model	En→Id	En→Kk	En→Th	Zh→Ja	Avg	En→Id	En→Th	Zh→Ja	Avg
			Flores200				WMT24++		
Base	79.97	55.91	76.27	76.56	72.18	69.82	69.64	69.91	69.79
+ Stage I	80.14	78.06	76.09	75.36	77.41	69.32	66.55	67.89	67.92
+ Stage I + Stage II (Ours)	80.54	78.34	77.34	74.31	77.63	71.31	69.42	66.09	68.94

sp-BLEU									
Model	En→Id	En→Kk	En→Th	Zh→Ja	Avg	En→Id	En→Th	Zh→Ja	Avg
			Flores200				WMT24++		
Base	36.81	6.48	35.98	18.59	24.47	32.30	25.93	14.19	24.14
+ Stage I	50.99	39.02	40.68	29.26	39.99	32.42	19.70	17.67	23.26
+ Stage I + Stage II (Ours)	51.71	50.35	47.12	43.27	48.11	36.50	27.72	22.20	28.81

Note: COMETKiwi-XL scores are computed using the Unbabel/wmt23-cometkiwi-da-xl quality estimation model (Rei et al., 2022). BLEU scores are reported using sp-BLEU (SentencePiece BLEU) (Goyal et al., 2022).

Table 7: Translation performance on Flores200 and WMT24++.

ness to different document organizations, each pseudo-document is randomly formatted into one of three variants, namely sentence-separated, space-separated, and newline-separated forms. For immutable entity preservation, English monolingual sentences collected from the Do Not Change Me dataset are translated into Chinese by a teacher model with customized system prompts that explicitly require predefined entities, including alphanumeric strings, URLs, email addresses, emojis, IP addresses, ISBNs, phone numbers, social media handles, and date expressions, to be preserved verbatim. The generated bilingual sentence pairs are subsequently expanded into both translation directions. Finally, all bilingual sentence pairs are converted into a unified instruction-following dialogue format using the corresponding task-specific instruction templates. This unified construction framework enables different translation constraints to share the same data representation while maintaining their task-specific instruction semantics, facilitating joint optimization within a single multilingual translation model.

2.5 Training and Inference

Our translation system is built upon Qwen3-8B (Yang et al., 2025) and optimized through a two-stage full-parameter supervised fine-tuning (SFT) strategy. The first stage uses the parallel corpus and instruction-following data, while the second stage further incorporates the knowledge-distilled corpus to improve translation quality while preserving instruction-following capability learned in the first stage across different translation directions. This two-stage training strategy enables the model

to first acquire high-quality bilingual translation knowledge and subsequently benefit from large-scale synthetic parallel data.

Training is conducted using the LlamaFactory framework (Zheng et al., 2024) with full-parameter supervised fine-tuning. We employ the AdamW optimizer with a learning rate of 1×10^{-5} and train the model for 2 epochs in each stage. All experiments are performed on eight NVIDIA A100 80GB GPUs using DeepSpeed ZeRO-2 optimization (Rajbhandari et al., 2020). The maximum sequence length is set to 2048 tokens. Additional training hyperparameters are summarized in Table 6.

During inference, we adopt deterministic decoding with a temperature of 0.0 for instruction-following data generation. For translation evaluation, decoding is performed with a maximum generation length of 2048 tokens. Unless otherwise specified, all experiments use the same inference configuration across different translation directions.

3 Main Results

Table 7 presents the translation performance of the base model, the Stage I model, and the final system after Stage II training on Flores200 devtest and WMT24++. The evaluation is conducted using the Unbabel/wmt23-cometkiwi-da-xl quality estimation model and sp-BLEU.

Compared with the base model, Stage I substantially improves translation quality across all four language directions on Flores200. The average COMETKiwi-XL score increases from 72.18 to 77.41, while the average sp-BLEU score improves from 24.47 to 39.99.

Further fine-tuning on the knowledge-distilled corpus improves the average performance of the Stage I model. After Stage II training, the average COMETKiwi-XL score on Flores200 further increases from 77.41 to 77.63, while the average sp-BLEU score reaches 48.11. On WMT24++, Stage II improves over Stage I on average for both metrics and achieves the best overall sp-BLEU score, although its average COMETKiwi-XL score remains below that of the base model. These results indicate that the synthetic parallel data constructed from high-quality monolingual corpora effectively complements the bilingual parallel corpus and further enhances multilingual translation performance.

4 Conclusion

In this paper, we present TransCore, our submission to the constrained track of the WMT26 General Machine Translation Shared Task. Our system is built upon Qwen3-8B and adopts a two-stage supervised fine-tuning framework. We construct a high-quality training corpus consisting of carefully curated parallel data, knowledge-distilled data, and instruction-following data. To support the instruction-following evaluation introduced in WMT26, we further construct a unified instruction-following data construction framework covering immutable entity preservation, document-level translation, and structured translation. The proposed data construction pipeline provides a practical data construction strategy for multilingual LLM-based machine translation under the WMT26 instruction-following setting. Future work will investigate more effective multilingual data construction, synthetic data generation, and instruction-following optimization for LLM-based machine translation.

References

- Noëmi Aepli, Chantal Amrhein, Florian Schottmann, and Rico Sennrich. 2023. [A benchmark for evaluating machine translation metrics on dialects without standard orthography](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 1045–1065, Singapore. Association for Computational Linguistics.
- Shanbo Cheng, Yu Bao, Qian Cao, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, Wenhao Zhu, Jingwen Chen, Zhichao Huang, et al. 2025. Seed-x: Building strong multilingual translation llm with 7b parameters. *arXiv preprint arXiv:2507.13618*.
- Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. [The Flores-101 evaluation benchmark for low-resource and multilingual machine translation](#). *Transactions of the Association for Computational Linguistics*, 10:522–538.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Sergey Dukanov, Anton Dvorkovich, Mark Fishel, Markus Freitag, Samuel Frontull, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Stan Kharevich, Philipp Koehn, Zheng Li, Jean Maillard, Christof Monz, Alexander Murauski, and 12 others. 2026. Findings of the WMT26 general machine translation shared task: Contrastive dynamic human evaluation at scale. In *Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics.
- Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024. [Eliciting the translation ability of large language models via multilingual finetuning with translation instructions](#). *Transactions of the Association for Computational Linguistics*, 12:576–592.
- Yingfeng Luo, Ziqiang Xu, Yuxuan Ouyang, Murun Yang, Dingyang Lin, Kaiyan Chang, Tong Zheng, Bei Li, Peinan Feng, Quan Du, et al. 2025. Beyond english: Toward inclusive and scalable multilingual machine translation with llms. *arXiv e-prints*, pages arXiv–2511.
- Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC ’20. IEEE Press.
- Ricardo Rei, Nuno M Guerreiro, José Pombal, João Alves, Amin Farajian, Pedro Teixeira, and Andre Martins. 2026. [TOWER+: Bridging generality and translation specialization in multilingual LLMs](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29614–29635, San Diego, California, United States. Association for Computational Linguistics.
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. [CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task](#). In

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025. Hunyuan-mt technical report. *arXiv preprint arXiv:2509.05209*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyang 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.