SYSTRAN @ WMT 2025 General Translation Task

Dakun Zhang, Yara Khater, Ramzi Rahli, Anna Rebollo and **Josep Crego**
SYSTRAN by ChapsVision
5 rue Feydeau, 75002 Paris (France)
{dzhang,ykhater,rrahli,arebollo,jcrego}@chapsvision.com

Abstract

We present an English-to-Japanese translation system built upon the EuroLLM-9B (Martins et al., 2025) model. The training process involves two main stages: continue pretraining (CPT) and supervised fine-tuning (SFT). After both stages, we further tuned the model using a development set to optimize performance. For training data, we employed both basic filtering techniques and high-quality filtering strategies to ensure data cleanness. Additionally, we classify both the training data and development data into four different domains and we train and fine-tune with domain specific prompts during system training. Finally, we applied Minimum Bayes Risk (MBR) decoding and paragraph-level reranking for post-processing to enhance translation quality[1].

1 Introduction

Large language models (LLMs) are increasingly used in machine translation, taking advantage of their deep understanding of both source and target languages. Their training typically involves two main stages: continued pretraining (CPT) and supervised fine-tuning (SFT).

In the pretraining stage, the model is exposed to massive amounts of unlabeled text and learns by predicting the next token in a sequence. This allows the model to acquire a broad understanding of language structure, grammar, general world knowledge, and reasoning patterns. In the supervised fine-tuning stage, the model is trained on task-specific labeled datasets, where each input is paired with reference output. This targeted training enables the model to follow instructions more precisely and handle specialized tasks, such as question answering, summarization, classification, and translation, with greater accuracy.

For the WMT25 general translation task, we began with the pretrained LLM EuroLLM-9B (Martins et al., 2025) and performed additional training using bilingual corpora containing only English–Japanese sentence pairs. In this stage, we employed the two aforementioned training approaches: continued pretraining (CPT) and supervised fine-tuning (SFT), and trained separate systems using each method. Following the first stage, a reduced development dataset was employed to fine-tune (FT) the systems to the WMT translation tasks. The training architecture is shown in Figure 1 (left side).

Before generating translations, we segment the WMT25 test set into individual sentences using the newline character ("\n"), as our systems are trained to operate at the sentence level.

During inference, we apply Minimum Bayes Risk (MBR) decoding and reranking of diverse translation hypotheses produced by our two models. For each input sentence, we generate up to 300 translation candidates by combining outputs from both trained models with variations in decoding prompts (Table 6), greedy/nucleus decoding (5-best), and zero-shot/few-shot examples. A quality estimation step is then applied to these hypotheses, discarding the worst 50% for each input. From the remaining candidates, MBR decoding is used to select the most promising translation, following the approach of Rei et al. (2024). Finally, the translated sentences are concatenated back into paragraphs, which are reranked using CometKiwi[2] (Rei et al., 2022), with the top-ranked paragraph selected as the final system output. The inference post-process architecture is shown in Figure 1 (right side).

As a result, our submission is an ensemble of two open-weight, sentence-level, English-to-Japanese translation models with a combined total of 18B parameters. The following sections describe the

[1]We released the classification model and translation models: https://huggingface.co/Systran/collections

[2]Unbabel/wmt23-cometkiwi-da-xl. All CometKiwi scores in this work were computed using this model.

Proceedings of the Tenth Workshop on Machine Translation (WMT), pages 599–606
November 8–9, 2025.

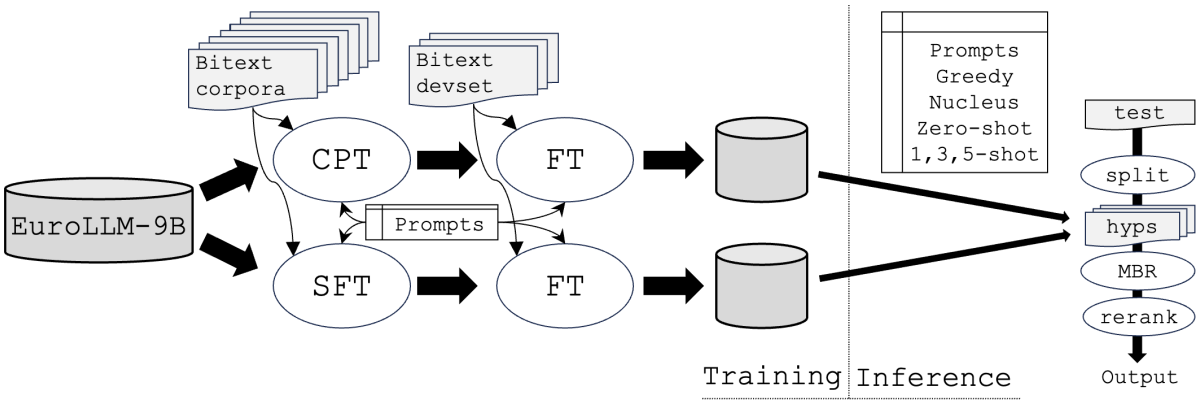

Figure 1: Frameworks for Training and Inference Post-Processing. System prompts are shown in Appendix A.

data preparation process as well as give additional details on the training procedures used to build our WMT25 English-Japanese system.

2 Data preparation

2.1 Bitext data

We use only parallel corpora to continue training LLMs for the machine translation task. We apply both basic and high-quality filtering methods to the WMT25 English-Japanese bitext data. From the original 225 million sentence pairs, we select approximately 10 million high-quality pairs for training. These filtered pairs are then used to train models using both CPT and SFT strategies. Similar filtering methods are also applied to the development datasets to further fine-tune (FT) the trained models. Table 1 summarizes the data statistics for the training and development datasets.

2.1.1 Basic Filtering Rules

Basic filtering rules are applied first to remove duplicates and noisy samples, such as misaligned sentence pairs and those with significantly different length ratios. Details of the filtering steps are as follows:

- **uniq:** Remove duplicated bitext examples.
- **1:1 example:** Discard examples where a single sentence is aligned to multiple sentences (1:N) or multiple sentences are aligned to a single sentence (M:1).
- **length ratio:** Discard examples where the length ratio between source and target exceeds 10 or is less than 0.2[3].
- **length:** Discard examples with length greater than 500 tokens.
- **character:** Retain sentences containing only Latin and Greek characters on the English side, and Latin, Greek, Hiragana, Katakana, and Han characters on the Japanese side.
- **LID:** Perform language identification using `lid.176.bin` (Joulin et al., 2016b,a).

2.1.2 High-Quality Filtering

The high-quality filtering process involves two main steps: first, generating translation hypotheses using a translation model; second, estimating the similarity between the source sentence and the target sentence, or between the source sentence and the generated hypothesis, using a quality evaluation model. For this purpose, we use EuroLLM-9B-Instruct (Martins et al., 2025) to translate English sentences into Japanese, and CometKiwi to assess the quality scores between English and Japanese segments in this year's WMT evaluation. Finally, we remove those samples which

$$CometKiwi(src, tgt) < CometKiwi(src, hyp)$$

2.2 Development data

Development data is used to fine-tune the pretrained models to adapt to WMT evaluation. We first merge all development data, both English-to-Japanese and Japanese-to-English, provided by WMT25, including WMT dev/test data in previous years, NTREX (Federmann et al., 2022) and Flores (NLLB Team et al., 2024), except wmttest2024 dataset[4]. Then we apply high-quality filtering (Section 2.1.2) after deduplication. Details are shown

[3]We use basic tokenization for length and length ratio filtering: on the English side, we tokenize by spaces, while on the Japanese side, we tokenize at the character level.

[4]https://data.statmt.org/wmt24/general-mt/wmt24_GeneralMT.zip

	Bitext	Dev set
wmt25 provided	225M	20,111
uniq	132M	18,019
1:1 example	57M	-
basic rules (length, LID, etc.)	55M	-
high quality	10M	5,736

Table 1: Data statistics (number of lines) for training and development dataset filtering.

in Table 1. There are in total 5,736 sentence pairs finally used for system fine-tuning (FT in Table 1).

2.3 Data classification

In WMT2024, the test dataset covers four domains: News, Social, Speech and Literary. To perform domain classification, we fine-tune Llama-3.1-8B-Instruct[5] model on the WMT24++ dataset (Deutsch et al., 2025) and use the resulting classifier to label both the training and development data. The prompt used for classification during both training and inference is shown in Table 5.

The fine-tuned model categorizes English input sentences into one of the four domains. Within the 10M cleaned parallel corpus, the distribution across these domains is 7% (News), 74% (Social), 12% (Speech), and 7% (Literary), indicating that the Social domain constitutes the majority of the training data.

This classification model is used only for training data preparation. For decoding, we generate translation hypothesis for each input with all domain related prompts (Table 6) and rely on MBR postprocessing to select the best candidate.

Table 7 shows that domain related prompts benefit the final system.

3 Model training

3.1 Continue pretraining and Supervised fine-tuning

The pretraining stage of LLM typically requires vast amounts of unlabeled monolingual data. However, since the WMT evaluation is dedicated exclusively to machine translation, we leverage parallel corpora as a stand-in for monolingual data. Accordingly, we train LLMs independently using two approaches, continued pretraining (CPT) and supervised fine-tuning (SFT), on parallel data.

[5]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

For CPT, we generate training examples of up to 2048 tokens by appending the corresponding Japanese sentence to the end of each English input, together with a domain-specific prompt (Section 2.3). To minimize dependency between bilingual sentence pairs in the synthetic example, we insert an "end-of-sentence" token (</s>) after each Japanese sentence. The input text format for CPT is:

```
sample = (domain_prompt)En\nJa\n</s>
input = [sample] +
```

where `input` is constructed by concatenating (+) one or more `sample` entries, each containing an English (`En`) and a Japanese (`Ja`) sentence. The actual set of `domain_prompts` used are shown in Table 6.

We apply 10% prompt smoothing to the domain-specific prompts, where each of the four prompt types — News, Social, Speech and Literary — is sampled with a 10% probability with generic domain-free prompt (default 1 and default 2 in Table 6). This helps mitigate the impact of annotation errors and enhances training diversity.

We apply similar domain-specific prompts for SFT training. The only difference is that training examples are not concatenated during SFT.

We use LLaMA-Factory (Zheng et al., 2024) to train CPT/SFT models from EuroLLM-9B. The training parameters are the same as described in GemmaX2-28 (Cui et al., 2025) except that we use 4 GPUs in parallel for both training. The effective batch size for both trainings is 128. The learning rate starts from 2.0e-5 and decays based on cosine_with_min_lr policy, with a miminum value of 1.0e-6 (Table 8).

We use full model tuning rather than parameter-efficient methods such as LoRA adapters for CPT/SFT training. This choice is motivated by the relatively small size of the training data, where full tuning has been shown to yield better performance than adapter-based methods in similar low-resource settings (Hu et al., 2021; Pfeiffer et al., 2021).

3.2 Fine-tuning with development data

To further adapt our models to the WMT task, we perform an additional round of fine-tuning (FT) on both the CPT and SFT models using the filtered previous year's development data except wmttest2024 (Section 2.2).

Given the limited dataset size of 5,736 samples and an effective batch size of 128, this fine-tuning

Policy	Prompts	zero/few-shot	Greedy	Nucleus
Adequacy	all	all	Yes	No
Diversity	all	all	Yes	Yes

Table 2: Decoding policies (Adequacy vs. Diversity) for MBR post-processing.

wmttest2024		Adequacy policy		Diversity policy	
		BLEU	CometKiwi	BLEU	CometKiwi
Baseline	EuroLLM-9B-instruct	26.3	0.7501	22.9	0.7620
	EuroLLM-9B-instruct-FT	26.4	0.7500	22.6	0.7624
Our models	EuroLLM-9B-CPT-FT	29.5	0.7504	24.5	0.7659
	EuroLLM-9B-SFT-FT	28.4	0.7515	24.9	0.7667
	Ensemble	**29.8**	0.7586	24.6	**0.7686**

Table 3: Sentence-level evaluations on wmttest2024 English-Japanese translation. Both BLEU and CometKiwi scores are the result of MBR output. BLEU is calculated by Sacrebleu (Post, 2018) with ja-mecab tokenization. CometKiwi is the reference-free score of Unbabel/wmt23-cometkiwi-da-xl.

process runs for only 45 steps (1 epoch). In this stage, training samples are also built with domain-specific prompts as in our previous training. We apply full model training instead of using LoRa adapters for fine-tuning, with a starting learning rate 2.0e-5 and inverse_sqrt decay policy. We add NEFTune noise (Jain et al., 2023), with alpha=5, in this fine-tuning stage (Table 8).

4 Decoding and Postprocessing

We follow the idea of quality-aware decoding proposed by Rei et al. (2024). First, we apply Quality Estimation (QE) to discard the translation candidates with the lowest quality. Then, we perform Minimum Bayes Risk (MBR) decoding over the remaining hypotheses, using CometKiwi (Rei et al., 2022) as the loss function. The candidate with the lowest expected risk is selected as the final output.

To generate diverse output for each input, we apply different domain-free/domain-specific prompts, zero/few-shot examples, and greedy decoding as well as nucleus decoding. Table 2 summarizes the two decoding policies that we used in this work. Note that the policies differ only in the use of nucleus sampling to generate diverse hypotheses, which favors outputs with higher diversity (lower adequacy). Table 3[6] confirms this, as the Adequacy policy achieves correspondingly higher BLEU scores, while the Diversity policy yields higher CometKiwi scores.

[6]EuroLLM-9B-instruct-FT is the model that we directly fine-tune EuroLLM-9B-instruct with the development data described in Section 2.2.

4.1 Reference-free Quality Estimation (QE)

To filter low-quality hypotheses, we employ Comet-QE (Rei et al., 2021) in a reference-free setting. Given a source sentence src and a set of N candidate translations $\{hyp_1, hyp_2, \ldots, hyp_N\}$, we use Comet-QE score to compute quality scores $s_i = \text{Comet-QE}(src, hyp_i)$ for each hypothesis hyp_i. We then retain only the top $K = \lfloor N/2 \rfloor$ candidates with the highest scores:

$$\{hyp'_1, \ldots, hyp'_K\} = \text{Top-}K\ (\{s_1, \ldots, s_N\})$$

This step serves to remove noisy or low-quality translations that may adversely affect subsequent MBR decoding (Kondo et al., 2024).

4.2 Minimum Bayes Risk (MBR) Decoding

Following QE filtering, we apply Minimum Bayes Risk decoding (Fernandes et al., 2022) using a reference-based CometKiwi score. For each sentence, we consider the remaining K candidates $\{hyp_1, \ldots, hyp_K\}$ and compute the pairwise loss between each one of them with the others using CometKiwi scores, treating each candidate as a reference for the others: $\ell(hyp_i, hyp_j) = 1 - \text{CometKiwi}(src, hyp_i, hyp_j)$. Each hypothesis is then assigned an expected loss:

$$\mathbb{E}[\ell(hyp_i)] = \sum_{j=1}^{K} p(hyp_j) \cdot \ell(hyp_i, hyp_j)$$

where $p(hyp_j) = \frac{\exp(\log P(hyp_j))}{\sum_k \exp(\log P(hyp_k))}$ is derived from the log-probabilities assigned by the model.

wmttest2025	CometKiwi	
	Adequacy policy	Diversity policy
EuroLLM-9B-CPT-FT	0.6769	0.6801
EuroLLM-9B-SFT-FT	0.6826	0.6842
Ensemble	0.6843	0.6859
Reranking (main submission)	**0.7033**	

Table 4: Paragraph-level evaluation (Kocmi et al., 2025) for wmttest2025. CometKiwi is the reference-free score of Unbabel/wmt23-cometkiwi-da-xl.

We select the hypothesis y^* that minimizes the expected loss:

$$y^* = \arg\min_{hyp_i} \mathbb{E}[\ell(hyp_i)]$$

and thus, selecting the hypothesis that is most representative of the overall candidate distribution.

4.3 Paragraph-level re-ranking (WMT25)

For the WMT25 test set, we first split each document into sentences using the newline character "\n". Then we apply MBR to select the best candidate for each sentence individually. The selected sentences are reassembled into their original paragraph structure. Finally, we compute paragraph-level CometKiwi scores and select the combination of sentences that yields the highest overall score as the final output.

Let $D = \{d_1, d_2, \ldots, d_N\}$ be the set of documents in the WMT25 test set. Each document d is split into a list of sentences:

$$S_d = [s_1, s_2, \ldots, s_{n_d}]$$

For each sentence s_i, let $y_i = \{y_{i_1}, y_{i_2}, \ldots, y_{i_k}\}$ be a set of corresponding candidate translations. Apply MBR decoding to select the best candidate $\hat{y}_i$, and reconstruct the sentence list with best translation candidates:

$$\hat{S}_d = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{n_d}]$$

These selected sentences are then reassembled into a paragraph structure:

$$\hat{P}_d = \text{Assemble}(\hat{S}_d)$$

Finally, among all possible sentence combinations $P \in \mathcal{P}_d$, select the one with the highest paragraph-level CometKiwi score:

$$\hat{P}_d^* = \arg\max_{P \in \mathcal{P}_d} \text{CometKiwi}(P)$$

As shown in Table 4, incorporating paragraph-level reranking further improves the quality of the final submission.

5 Conclusions

In this paper, we present our English-to-Japanese translation system for the WMT25 General Translation Task. The final output is generated by ensembling two models trained with different strategies: continued pretraining (CPT) and supervised fine-tuning (SFT). Both models are trained on a cleaned parallel corpora of 10 million sentence pairs and further fine-tuned (FT) on a development set consisting of 5,736 sentences. MBR and re-ranking inference post-processing are also successfully performed to obtain the final quality boost. Additionally, we release one classification model[7] and two translation models[8] in our HuggingFace repository.

Acknowledgments

This research was funded by the French Agence Nationale de la Recherche (ANR) under the project TraLaLaM ("ANR-23-IAS1-0006"). This work was performed using HPC resources from GENCI–IDRIS (Grant 2024-A0161015117).

References

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the Language Coverage of WMT24 to 55 Languages Dialects.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First*

[7] https://huggingface.co/Systran/Llama-3.1-8B-Instruct-ft-wmt25-classifier

[8] https://huggingface.co/Systran/EuroLLM-9B-cpt-ft-wmt25-en-ja, https://huggingface.co/Systran/EuroLLM-9B-sft-ft-wmt25-en-ja

Workshop on Scaling Up Multilingual Evaluation, pages 21–24, Online. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. 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.

Neel Jain, Ping yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R. Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Neftune: Noisy embeddings improve instruction finetuning.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025. Preliminary ranking of wmt25 general machine translation systems.

Minato Kondo, Ryo Fukuda, Xiaotian Wang, Katsuki Chousa, Masato Nishimura, Kosei Buma, Takatomo Kano, and Takehito Utsuro. 2024. NTTSU at WMT2024 general translation task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 270–279, Miami, Florida, USA. Association for Computational Linguistics.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA. 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.

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

A Prompts

classifier	{"role": "system", "content": "You are a language expert."} {"role": "user", "content": "Classify the following English sentence into one of the following categories based on its content and style:\nNews: Factual reporting or informative text typically found in journalism.\nSocial: Informal, conversational, or casual text often used on social media or in personal messages.\nSpeech: Spoken or scripted verbal communication, such as political speeches, interviews, or lectures.\nLiterary: Creative or artistic writing, including fiction, poetry, or other literary works.\n\nSentence: {line}\n\nYour task: Identify the most appropriate category from the four above. Just respond with the category name: News, Social, Speech, or Literary."}

Table 5: Prompt used to fine-tune classification model and inference

default.1	Please translate the following {source_language} text into {target_language}.\nInput: {line}\nOutput:
default.2	You're a professional {target_language} translator. Please translate the following {source_language} sentence into {target_language}. You must only answer with the translation.\n{source_language} sentence: {line}\n{target_language} sentence:
domain.news	You're a professional {target_language} translator. You are translating a news article. The translation should be clear, objective, and concise, preserving factual accuracy and the original journalistic tone. Do not add opinions or interpretations. Please translate the following {source_language} sentence into {target_language}. You must only answer with the translation.\n{source_language} sentence: {line}\n{target_language} sentence:
domain.literary	You're a professional {target_language} translator. You are translating a literary text. Pay attention to stylistic features such as imagery, rhythm, and narrative voice. The translation should be faithful to the original tone and emotional depth, while adapting gracefully into the target language. Please translate the following {source_language} sentence into {target_language}. You must only answer with the translation.\n{source_language} sentence: {line}\n{target_language} sentence:
domain.speech	You're a professional {target_language} translator. You are translating a speech transcript. Maintain a fluent, persuasive tone suitable for public speaking. The output should be easy to read aloud and emotionally engaging, while staying faithful to the speaker's intent. Please translate the following {source_language} sentence into {target_language}. You must only answer with the translation.\n{source_language} sentence: {line}\n{target_language} sentence:
domain.social	You're a professional {target_language} translator. You are translating a social media post or informal message. The tone should be natural, conversational, and culturally relevant. Preserve emojis, slang, and informal expressions when appropriate. Please translate the following {source_language} sentence into {target_language}. You must only answer with the translation.\n{source_language} sentence: {line}\n{target_language} sentence:

Table 6: Prompt used to train translation model and inference

B Results on domain related prompts

EuroLLM-9B-CPT	+FT	wmttest2024	
		BLEU	CometKiwi
✗		20.7	0.7120
✓		22.9	0.7140
✗	✗	27.7	0.7373
✗	✓	28.1	0.7369
✓	✗	28.7	0.7377
✓	✓	28.6	0.7376

Table 7: Evaluation of the EuroLLM-9B-CPT-FT model trained with domain-related prompts (✓) and without domain-related prompts (✗). Scores are computed on wmttest2024 using the default prompt during the decoding phase.

C Training parameters

	CPT	SFT/FT
per_device_train_batch_size	4	4
Number of GPUs	4	4
gradient_accumulation_steps	8	8
Data cutoff length	2048	2048
Number of epochs	1	1
Max Learning Rate	2.0e-5	2.0e-5
Min Learning Rate	1.0e-6	/
lr_scheduler_type	cosine_with_min_lr	inverse_sqrt
Finetuning Type	full	full
bf16	true	true
Template	empty	empty
warmup_ratio	0.01	0.01
weight_decay	0.01	0.01
Optimizer	AdamW	AdamW
Deepspeed	ZeRO2	ZeRO2
neftune_noise_alpha	/	5

Table 8: The training parameters for CPT, SFT and FT are configured according to Cui et al. (2025).