Bilingual English-Vietnamese Medical Translation Model

ZERO team
Long Phung Hoang, Tuyen Nguyen The
longhoangai1102@gmail.com
nguyenthetuyen23122002@gmail.com

Abstract

This technical report describes Team ZERO's submission to the VLSP 2025 Shared Task on Medical Machine Translation. The task challenges participants to build a high-accuracy English-Vietnamese translation system under strict constraints, using pre-trained models with a maximum of 3 billion parameters. Addressing the challenges of domain-specific terminology and limited model knowledge, we propose a two-stage fine-tuning pipeline. Our approach begins by adapting the official baseline model, Qwen2.5-3B-Instruct, using Supervised Fine-Tuning (SFT) on the provided medical corpus. We then further refine the model's performance with Group Relative Policy Optimization (GRPO), a reinforcement learning technique that optimizes for translation quality using reward functions. Our constrained system demonstrates a notable improvement in translation quality over the baseline, validating the effectiveness of our training strategy.

1 Introduction

The VLSP 2025 Shared Task on Medical Machine Translation addresses a pressing real-world challenge: building accurate translation systems for specialized domains under strict resource constraints. Medical translation is particularly demanding due to the need for precise handling of complex terminology, high fidelity to avoid clinical risks, and the scarcity of high-quality bilingual medical data. The task further increases the difficulty by restricting participants to pre-trained models with no more than 3 billion parameters, such as the Qwen 2.5/3 families.

Compared to larger language models, these constrained backbones lack extensive world knowledge and deep linguistic capacity, making it difficult to capture the nuances of biomedical texts. To address these limitations, we propose a two-stage fine-tuning pipeline designed to maximize performance within the competition's constraints.

In the first stage, we apply Supervised Fine-Tuning (SFT) to adapt the general-domain knowledge of a pre-trained model to the lexicon and syntax of medical language. Since SFT primarily imitates training data, we add a second refinement stage using Group Relative Policy Optimization (GRPO) (Shao et al., 2024). This reinforcement learning–based alignment method directly optimizes model outputs against automatic metrics correlated with translation quality, without relying on costly human preference data.

This report presents our full system, including data preprocessing, training methodology, and results on the public test set. While our pipeline builds on established methods, its novelty lies in the adaptation to strict task and resource constraints. We introduce a composite BLEU/chrF++ reward tailored for English–Vietnamese medical translation and employ direction-specific adapters to reduce task interference. These design choices, though incremental, proved essential for achieving consistent improvements with a 3B parameter model. Our primary contribution is not a novel algorithm, but rather a practical and effective blueprint for adapting constrained language models to highly specialized domains under strict resource limitations. We demonstrate how a carefully engineered two-stage pipeline, leveraging both supervised learning and reinforcement learning, can achieve significant performance gains. This work provides empirical evidence that even models with a limited parameter count (under 3B) can deliver robust, domain-specific translation performance when fine-tuned with a meticulous methodology.

2 System Architecture and Methodology

Our system is developed entirely within the constrained track of the shared task, using only the provided data and permitted base models.

2.1 Base Model Selection

The selection of an optimal base model is a critical first step in our pipeline. Our choice was guided by an analysis of promising candidates from the official list provided by the VLSP 2025 organizers, with a focus on models that balance performance, size, and suitability for instruction-following tasks.

Based on these criteria, we selected `Qwen2.5-3B-Instruct` as the definitive backbone for our system. As an instruction-tuned model, it is inherently well-suited for translation, which can be framed as a prompt-following task (e.g., "Translate English to Vietnamese"). This pre-training alignment significantly simplifies the adaptation process. Furthermore, its 3-billion-parameter architecture offers substantial linguistic capacity while adhering to the competition's strict constraints. Its robust performance in general-domain benchmarks provided a strong signal of its potential as a solid foundation, which we then aimed to specialize for the medical domain through our two-stage fine-tuning pipeline. Based on these considerations, we confidently chose `Qwen2.5-3B-Instruct` for all subsequent experiments.

2.2 Data Preprocessing and Preparation

We utilized the official parallel English-Vietnamese corpora provided by the VLSP 2025 organizers. Our first critical step was a rigorous data cleaning and preparation process, applied to both the training and test sets to ensure data quality and prevent leakage.

- **Training set deduplication:** The initial training set of 500,000 sentence pairs was found to have significant redundancy. To address this and foster better model generalization, we implemented a robust deduplication pipeline based on the MinHash LSH algorithm. This approach combines two techniques: MinHash (Broder, 1997) to create compact numerical signatures for each sentence, and Locality-Sensitive Hashing (LSH) (Indyk and Motwani, 1998) to efficiently group similar sentences. This allowed us to identify near-duplicate pairs without performing an exhaustive pairwise comparison. The process resulted in a cleaned, unique training corpus of 349K sentence pairs (a 30.2% reduction).

- **Testing set deduplication:** To ensure a fair evaluation, the public test set underwent a separate curation process. First, to prevent data leakage, we removed any test samples that were found to overlap with our newly cleaned training corpus. Following this, the set was augmented with clean, non-overlapping samples from the original training data pool to restore its target size of approximately 3,000 pairs. This resulted in a final, sanitized test set for reliable performance measurement.

- **Final data splits:** The final 349K cleaned training corpus was partitioned into three mutually exclusive splits for our pipeline: a 331K-pair set for SFT, a 15K-pair set for GRPO, and the remaining 3K pairs were used for model evaluation.

2.3 Training Methodology

Our core methodology is a sequential training process designed to first instill domain knowledge and then refine translation quality.

2.3.1 Stage 1: Supervised Fine-Tuning (SFT)

The primary goal of this stage is to adapt the base LLM to the target domain's vocabulary and syntactic structures. The process was divided into data preparation and model training configuration.

Data preparation. To construct an effective training set for the SFT stage, we performed the following steps:

- **Bidirectional training:** We utilized each of the 331K sentence pairs from the cleaned training set to create samples for both translation directions (EN→VI and VI→EN). This bidirectional approach effectively doubled the number of training instances to a total of 662K, ensuring the model received balanced exposure to both tasks.

- **Data mixing:** The combined bidirectional samples were then randomly shuffled to ensure balanced training and prevent the model from learning any spurious order-based patterns.

Training configuration. The SFT process was configured with the following parameters and strategies:

- **Method:** We employed Low-Rank Adaptation (LoRA) (Hu et al., 2022) for its parameter-efficient fine-tuning capabilities, which significantly reduces computational requirements.

- **LoRA parameters:** The LoRA rank was set to 128 and the alpha was set to 256. This configuration provides a good balance between the adapter's expressive capacity and the number of trainable parameters.
- **Training epochs:** The model was trained for a total of 6 epochs over the entire 662K-instance dataset.
- **Checkpointing strategy:** We saved a model checkpoint after each epoch and continuously monitored its performance on the development set. The checkpoint that yielded the highest BLEU score was selected as the final SFT model to serve as the backbone for the subsequent GRPO stage.

2.3.2 Stage 2: GRPO Refinement

To move beyond simple imitation of the training data, we employed GRPO to align the model's outputs more closely with desired translation properties. This was crucial for improving fluency and accuracy.

- **Motivation:** GRPO is ideal for this task as it does not require human preference labels. Instead, it improves the model by contrasting its own generated outputs with the ground-truth references from the 15K RL dataset.
- **Reward function design:** To provide a comprehensive evaluation of translation quality, we constructed a composite reward signal from two distinct and complementary automatic metrics. Each metric functions as an independent signal assessing a different aspect of the generated text:
    - **BLEU signal:** The first signal, based on BLEU (Papineni et al., 2002), serves as our primary measure for lexical accuracy and adequacy. It rewards the model for correctly translating key terms and phrases found in the reference translation.
    - **chrF++ signal:** The second signal, using chrF++ (Popović, 2015), complements BLEU by evaluating character-level fluency and morphological correctness. This is particularly crucial for Vietnamese, allowing the model to be rewarded for grammatically sound sentences even if there are minor lexical deviations.

  For the GRPO update step, these two independent signals are combined into a single scalar reward using a weighted average: 70% from the BLEU score and 30% from the chrF++ score. This weighting scheme allows us to prioritize lexical fidelity while still strongly encouraging grammatical fluency.
- **Direction-specific adapters:** We adopted a direction-specific optimization strategy, training two separate LoRA adapters—one for EN→VI and one for VI→EN. For this refinement stage, both adapters were configured with a LoRA rank of 64 and an alpha of 64, and trained in a single epoch. This specialization prevents task interference and allows each adapter to fine-tune its parameters for the specific nuances of one translation direction, leading to better overall performance compared to a single bidirectional model.

3 Experiments and Results

To validate the effectiveness of our proposed two-stage training pipeline, we conducted a comprehensive set of experiments adhering strictly to the constrained track of the VLSP 2025 shared task.

3.1 Experimental Setup

Datasets and Metrics. We evaluate our systems on the official VLSP 2025 public test set, which consists of 3,000 curated sentence pairs. This set was carefully prepared to ensure no overlap with our training or validation data, providing a fair and reliable benchmark. Following the shared task guidelines, our primary evaluation metric is the BLEU score (Papineni et al., 2002). To ensure standardized and reproducible comparisons, all BLEU scores are calculated using SacreBLEU (Post, 2018) with its default tokenization settings.

Baselines and Systems compared. To isolate and measure the contribution of our GRPO refinement stage, we compare the performance of two key systems:

- **SFT-only:** This model serves as our strong baseline. It is the result of our initial Supervised Fine-Tuning stage (Stage 1), representing a robust adaptation of the base `Qwen2.5-3B-Instruct` model to the medical domain.
- **SFT + GRPO:** This is our final proposed system and the official submission to the shared

task. It represents the SFT-only model after being further refined using our GRPO methodology (Stage 2) with the composite reward function.

Implementation details. All experiments were conducted using the Hugging Face Transformers, Unsloth, and TRL libraries. Training was performed on a single NVIDIA A100 GPU with 80GB of VRAM. Key hyperparameters for LoRA (rank, alpha) and training epochs for each stage are detailed in Section 2.

3.2 Main Results

The performance of both systems on the public test set is presented in Table 1. The results clearly demonstrate a significant and consistent improvement achieved by our GRPO refinement stage across both translation directions.

Training Method	EN→VI BLEU	VI→EN BLEU
SFT only	49.54	38.00
SFT + GRPO	**51.13**	**39.48**
Improvement	+1.59	+1.48

Table 1: Performance comparison on the VLSP 2025 public test set. Our final system (SFT+GRPO) shows improvements over the SFT baseline.

3.3 Analysis and Discussion

Our final SFT + GRPO system achieved a significant improvement of 1.59 BLEU points (a 3.2% relative gain) for English-to-Vietnamese translation and 1.48 BLEU points (a 3.9% relative gain) for Vietnamese-to-English. This consistent improvement across both directions validates our core hypothesis: for specialized domains, a second stage of RL-based alignment can effectively refine a model's performance beyond what SFT alone can achieve, even with limited-parameter models.

The gains can be attributed to GRPO's ability to directly optimize for metrics correlated with translation quality. Our reward function, with a 70/30 weighting for BLEU and chrF++, was determined empirically on a validation set to strike a balance between lexical fidelity and grammatical fluency. This composite reward guided the model to produce translations that are not only statistically likely but also qualitatively superior.

Evaluation Strategy and Qualitative Findings. Due to the limited timeframe of the shared task, our evaluation emphasized BLEU, the official metric that directly determined leaderboard ranking. To complement BLEU's focus on lexical fidelity, we incorporated chrF++, a metric more sensitive to fluency and morphological correctness, which is particularly crucial for a morphologically rich language like Vietnamese.

While a systematic, large-scale human evaluation was not feasible, we conducted a manual review of several hundred sentences from the test set. This qualitative analysis confirmed that the GRPO-refined outputs were consistently more fluent and terminologically accurate than the SFT-only baseline. Key improvements included the correct usage of specific medical terminology and the generation of more natural-sounding syntax, addressing common errors from the SFT model. These observations, though not a substitute for a formal human evaluation, suggest that the quantitative gains reflect meaningful improvements in translation quality. We acknowledge that a more systematic error analysis and a structured human evaluation are key priorities for future work.

4 Deployment and Efficiency

As per the task requirements, our final system is packaged in a self-contained Docker image. This image includes the base `Qwen2.5-3B-Instruct` model and our two trained, direction-specific LoRA adapters. For inference, the system is powered by the vLLM framework, which is configured with a maximum context length of 4096 tokens to handle longer inputs. This setup ensures very fast inference speeds and high throughput. A provided Bash script handles the interaction, allowing for offline translation of an input text file.

This adapter-based serving mechanism is not only a submission requirement but also a highly efficient deployment strategy. By leveraging vLLM's PagedAttention mechanism to optimize GPU memory and batching, we achieve significant gains in translation speed. Loading the large base model once and dynamically applying the lightweight adapters (each only a few megabytes in size) as needed further reduces VRAM consumption and enables near-instantaneous switching between translation directions, a process much faster and more memory-efficient than loading two full models

5 Conclusion and Future Work

In this technical report, we presented Team ZERO's system for the VLSP 2025 Medical MT Shared Task. By implementing a careful data preprocessing strategy and a two-stage training pipeline combining SFT and GRPO, we successfully adapted a constrained 3B-parameter model to this challenging, specialized domain. Our results demonstrate the effectiveness of this approach, yielding significant improvements in translation quality while adhering to all task constraints.

Despite these gains, we acknowledge several limitations and identify promising directions for future work. First, our choice of GRPO was motivated by its practicality in this constrained setting, as it optimizes directly on automatic reward functions without requiring costly human preference data. However, we did not conduct ablation studies to disentangle the contributions of our reward shaping from the GRPO algorithm itself. Future work should include controlled comparisons against other alignment methods like PPO or DPO to precisely quantify the impact of each component.

Second, our evaluation was primarily driven by the task's official metrics. To move beyond this, we plan to broaden our evaluation framework. This includes conducting a systematic human evaluation with multiple annotators and incorporating stronger learned metrics (e.g., COMET, BLEURT) for a more comprehensive assessment of clinical accuracy and safety.

Third, our experiments were restricted to the official VLSP dataset. To assess and improve real-world robustness, we plan to evaluate the system on external biomedical and clinical corpora and explore data augmentation with large-scale monolingual resources.

Looking ahead, we aim to enhance the system's core capabilities. A key direction is to explore more advanced reward functions that incorporate semantic similarity metrics or pretrained scoring models to better preserve nuanced medical meaning. Furthermore, while our current system adheres to the constrained track, a natural next step is to investigate the impact of external parallel resources, which could substantially boost performance and bring the system closer to human-level translation quality.

References

Andrei Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of Sequences 1997 (Cat. No.97TB100171)*, pages 21–29. IEEE.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Piotr Indyk and Rajeev Motwani. 1998. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the 30th Annual ACM Symposium on Theory of Computing (STOC '98)*, pages 604–613. ACM.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. ACL.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*, pages 392–395. ACL.

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. Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.