VNU Journal of Science: Comp. Science & Com. Eng., Vol. 42, No. xx (2026) xx–xx

VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

Efficient English-Vietnamese Medical Machine Translation: Insights from the VLSP 2025 Shared Task

Tran Hong Viet, Tran Duy Long, Nguyen Minh Quy, Nguyen Van Vinh*

VNU University of Engineering and Technology, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam

Received 19th December 2025
Revised 24th December 2025; Accepted 26th March 2026

Abstract: In this paper, we present a comprehensive overview of the VLSP 2025 Medical Machine Translation Shared Task, which focuses on English-Vietnamese translation in the medical domain using small language models (SLMs). The shared task provides a large-scale, high quality parallel corpus and encourages participants to develop efficient, domain-adapted translation systems under resource constraints. We summarize the dataset construction, evaluation protocols, and model constraints, and review the diverse strategies adopted by participating teams-including parameter efficient fine-tuning, bidirectional training, retrieval augmented generation, and reinforcement learning with reward optimization. Our analysis highlights the strengths and limitations of SLM-based approaches for medical translation, discusses key findings from the competition, and outlines future research directions for building scalable, accurate, and practical machine translation systems in specialized domains.

1. Introduction

Machine Translation (MT) plays a vital role in overcoming language barriers and facilitating the global exchange of information [1]. Its importance is particularly evident in the medical domain, where accurate and timely access to information can directly influence clinical decision-making, medical research, and patient safety. The ability to rapidly translate medical documents, clinical reports, and scientific literature enables healthcare professionals to stay informed of the latest developments and improves the accessibility of medical knowledge across linguistic communities. However, medical translation remains one of the most challenging tasks for MT systems [2], especially for English-Vietnamese language pairs [3], due to strict accuracy requirements, complex domain-specific terminology [4], and the limited availability of high-quality parallel medical data [5].

Recent advances in large language models (LLMs) [6] have significantly improved general-

*Corresponding author.
E-mail address: vinhnv@vnu.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.6588

domain translation performance, yet their deployment in specialized and resource-constrained settings remains challenging. In response, small language models (SLMs) have emerged as a promising alternative [7], offering reduced computational costs, lower memory footprints, and faster inference while still achieving strong performance when effectively adapted to a target domain [8, 9]. These advantages make SLMs particularly attractive for medical translation scenarios, where efficiency, privacy preservation, and on-premise deployment are often required.

To foster research in this area, the Medical Translation Task organized under the Vietnam Language and Speech Processing (VLSP) evaluation campaign [10, 11] provides a standardized benchmark for English-Vietnamese medical MT. The task releases a large-scale, carefully curated medical parallel corpus collected from reliable sources such as hospitals and medical centers, resulting in approximately 500,000 high-quality sentence pairs [12]. Moreover, the task explicitly encourages the use of pretrained small language models, enabling a systematic investigation of their capabilities and limitations in domain-specific machine translation.

This article presents a comprehensive review of systems developed for the VLSP Medical Translation Task, with particular emphasis on SLM-based approaches. We analyze dataset characteristics, modeling strategies, and evaluation protocols, and discuss key findings and challenges observed across participating systems, especially those related to model size, domain adaptation [13], and translation quality. Through this review, we aim to highlight current progress and outline future research directions for efficient, scalable, and reliable English-Vietnamese medical machine translation using small language models.

2. Related Work

2.1. Medical Machine Translation

Medical machine translation (MT) is a specialized subfield of neural machine translation (NMT) that demands exceptional accuracy, terminology consistency, and semantic faithfulness due to the high stakes involved in clinical and research contexts [14, 15]. Mistranslations in medical documents can lead to significant risks, including misdiagnosis, inappropriate treatment, and compromised patient safety [16].

The challenges in medical MT are multifaceted. Domain-specific terminology and abbreviations are prevalent, often lacking direct equivalents in the target language [14]. Medical texts also feature long and complex sentence structures, further complicating translation. Moreover, the scarcity and imbalance of high-quality, domain-specific parallel data present additional hurdles compared to general-domain MT [15].

For English-Vietnamese medical MT, the situation is particularly acute. There are limited publicly available, high-quality parallel corpora, which constrains the development and evaluation of robust systems. Previous efforts, such as the MedEV dataset and initiatives under the VLSP framework, have begun to address these gaps, but resources remain limited [17, 18].

Shared-task evaluations play a crucial role in advancing the field by providing standardized datasets and evaluation protocols. They foster reproducibility and fair comparison, which are especially important in low-resource and domain-specific MT settings [19, 20]. This context motivates the necessity of the VLSP Medical MT Shared Task, as described in Section 3.

2.2. Small Language Models for Machine Translation

Transformer-based architectures have become the foundation of modern MT [21].

However, recent focus has shifted toward small language models (SLMs), which are defined by their parameter scale relative to large language models (LLMs) and are typically deployed in resource-constrained environments [22, 23].

SLMs offer several advantages in medical MT, including lower inference costs, faster decoding, and the feasibility of on-premise deployment, which is critical for privacy preservation in healthcare settings [22, 24]. These benefits make SLMs attractive for real-world medical applications, where computational resources and data privacy are significant concerns.

Nevertheless, SLMs entail trade-offs, such as reduced model capacity and the need for effective domain adaptation strategies to maintain translation quality [24]. The VLSP Medical MT Shared Task explicitly encourages the use of SLMs and evaluates systems under limited-resource inference settings, aligning with practical deployment scenarios and motivating research into efficient, high-quality MT solutions.

2.3. Parameter-Efficient Fine-Tuning Methods

Parameter-efficient fine-tuning (PEFT) methods have gained prominence for adapting large pre-trained models to specific domains without incurring prohibitive computational costs. Among these, Low-Rank Adaptation (LoRA) has emerged as a widely adopted approach [25].

$$W = W_0 + \Delta W, \quad \Delta W = AB$$

where $W_0 \in \mathbb{R}^{d\times k}$ is the pre-trained weight matrix, and $A \in \mathbb{R}^{d\times r}$, $B \in \mathbb{R}^{r\times k}$ are trainable matrices with $r \ll \min(d, k)$.

LoRA introduces low-rank updates to the attention and feed-forward layers of transformer models, enabling memory-efficient and faster training while preserving the benefits of pre-trained representations. This modularity allows for rapid experimentation and deployment across diverse tasks, making LoRA particularly suitable for the constraints of the VLSP Medical MT Shared Task.

2.4. Reinforcement Learning and Optimization for MT

Reinforcement learning (RL) has been increasingly explored in MT to directly optimize non-differentiable evaluation metrics such as BLEU [26, 27]. Unlike traditional maximum likelihood training, policy optimization approaches in MT enable models to align more closely with task-specific objectives.

Group Relative Policy Optimization (GRPO) is a recent advancement in this area, offering a high-level framework for stabilizing training and improving generation quality by leveraging flexible reward functions [28].

For GRPO, the objective generalizes standard policy optimization by allowing a flexible, possibly non-linear reward transformation $\mathcal{G}$

$$J_{\text{GRPO}}(\theta) = \mathbb{E}y \sim p\theta(\cdot|x)\,[\mathcal{G}(R(y))]$$

where :

- $\mathcal{G}$ is a general reward shaping or transformation function.
- $R(y)$ is the base reward (e.g., BLEU or other metric).

In the context of the VLSP Medical MT Shared Task, several participating teams have adopted GRPO to better align system outputs with both human and automatic evaluation metrics. However, practical considerations remain, including increased training complexity and sensitivity to reward design, which can impact the effectiveness and generalizability of RL-based approaches.

Based on the above discussion, we next describe the VLSP Medical Machine Translation Shared Task, including its dataset construction, evaluation protocol, and model constraints.

3. Task Definition and Constraint

The task focuses on developing a bilingual English-Vietnamese translation model in the medical domain that satisfies both limited-resource inference and high-quality translation.

Formally, given a source sentence

$$x = (x_1, x_2, \ldots, x_{T_x}) \in \mathcal{V}_{\text{src}}^{T_x},$$

where $\mathcal{V}_{\text{src}}$ is the source vocabulary and T_x is the sentence length, the model aims to generate a target sentence

$$y = (y_1, y_2, \ldots, y_{T_y}) \in \mathcal{V}_{\text{tgt}}^{T_y},$$

such that y is a faithful and fluent translation of x.

The translation model is parameterized by θ and defines a conditional probability distribution:

$$P_\theta(y \mid x) = \prod_{t=1}^{T_y} P_\theta(y_t \mid y_{<t}, x).$$

At inference time, the model outputs

$$\hat{y} = \arg\max_y P_\theta(y \mid x),$$

where $\hat{y}$ is the predicted translation of x.

Consistent with the discussion in Section 2, which highlights the practicality of small language models (SLMs) for domain-specific translation, this task emphasizes solutions that balance compactness and inference efficiency with translation fidelity. Participants are encouraged to pursue SLM-centered strategies, such as compact transformer variants and parameter-efficient fine-tuning, that support on-premise, low-latency deployment while preserving clinical terminology and semantic accuracy.

3.1. Building Corpus

The primary parallel training data are derived from the MedEV corpus [17], which contains approximately 360,000 English-Vietnamese sentence pairs collected from clinical institutions and professionally curated medical sources. Starting from this corpus, we performed an adaptation step using ICD-10 as a reference standard to enhance coverage of clinical terminology.

In particular, we incorporated an ICD-10 augmented version of MedEV [1] and integrated it with the original dataset. The combined data were then processed through a unified pre-processing pipeline, including length-based filtering, strict de-duplication, and text normalization. Sentence pairs containing fewer than five tokens on either the source or target side were removed.

After preprocessing, the final corpus consists of approximately 500,000 sentence pairs for training, along with a validation set of 3,000 examples and test sets of 1,000 examples per translation direction.

Table 1. Token counts for English-Vietnamese datasets. Train and Public Test sets are parallel, while Private Test is non-parallel

Dataset	En Tokens	Vi Tokens
Parallel Data		
Train	16,706,107	21,781,846
Public Test	101,803	132,320
Non-parallel Data		
Private Test	33,524	19,793

3.2. Evaluation Strategy

Automatic evaluation uses SacreBLEU following [29]. Teams may use the validation set for tuning and hyperparameter selection. Final submissions are additionally evaluated by practicing physicians with expertise in the medical domain, who assess each translation for clinical adequacy, correctness, and fluency, and assign a quality score according to the scale shown below. Each translation also

[1] https://huggingface.co/datasets/phuong123/adaption_med_ev

receives a SacreBLEU score, which quantifies a brevity-penalized geometric mean of n-gram precisions:

Table 2. Human evaluation score scale for translations

Score Range	Interpretation
90-100	Perfect
75-89	Good
56-74	Comprehensible
31-55	Partially understandable
0-30	Unable to understand

SacreBLEU is computed as a brevity-penalized geometric mean of n-gram precisions. The formula is given by:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right),$$

where p_n is the modified n-gram precision, w_n are uniform weights ($w_n = \frac{1}{N}$), and BP is the brevity penalty:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r, \\ \exp\left(1 - \frac{r}{c}\right) & \text{if } c \leq r, \end{cases}$$

with c the candidate length and r the reference length.

3.3. Model Constraint

The competition encourages teams to use small language models (SLMs) from the Qwen family [30] for the task. Teams are permitted to experiment with a variety of training and decoding strategies to improve model performance.

4. Approach Descriptions and Performance

The competition was hosted on Codabench [31], an online platform for organizing AI benchmarks and challenges. We did not limit the submission for evaluation to create more chances for participants. The leader board is public during the competition, allowing teams to refine their methods based on the results obtained on the public test set. The private test set is provided to participants seven days before the submission deadline along with final submission guideline. After teams submitted the source code and a brief system description, the organizers would verify that the submissions are reproducible. If the submissions on the platform is valid, they would be move to the final judgment phase involving doctors.

4.1. Baseline Model

To establish a common point of reference for the shared task, we evaluate **Qwen3-0.6B** [30] in a *zero-shot* setting prior to any domain-specific fine-tuning. The choice of a compact model is intentional, reflecting the task's emphasis on achieving strong medical translation performance under **parameter-efficient and resource-constrained** conditions.

In this baseline configuration, the model relies solely on its intrinsic multilingual capabilities without exposure to any task-specific training data. As such, this evaluation is not intended to represent a competitive domain-adapted system, but rather to provide a **clear lower-bound reference** that captures the model's inherent translation ability.

The resulting SacreBLEU scores on both public and private test sets are released to all participating teams through the evaluation platform, ensuring a consistent starting point for comparison across different fine-tuning and adaptation strategies.

Table 3. Translation performance (SacreBLEU scores) of the zero-shot baseline on public and private test sets

Direction	Public Test	Private Test
EN → VI	18.74	23.00
VI → EN	18.30	16.00

Building on this baseline, we next summarize the approaches of the four teams that submitted system papers. The analysis focuses on how different fine-tuning strategies, inference-time controls, and external resources enable substantial gains over the zero-shot reference while adhering to the task's efficiency-oriented design goals.

4.2. Bosch@AI

The Bosch@AI team proposed and implemented a bidirectional training strategy to fine tune small language models [32] for the bilingual machine translation task. The core idea of this approach is inspired by how humans naturally learn foreign languages: instead of being exposed to only one translation direction, learners frequently practice translating both from the source language to the target language and vice versa (e.g., from English to Vietnamese and from Vietnamese to English). This bidirectional approach reinforces bilingual associations, enhances the semantic mapping between the two languages, and improves the consistency of the model's language representations.

Within the scope of this study, the team employed two parallel model adaptation strategies: (i) full parameter fine tuning and (ii) Low-Rank Adaptation [33] (LoRA), to adapt the Qwen-3 1.7B model for the English-Vietnamese and Vietnamese-English bilingual machine translation tasks. In the supervised fine-tuning setup, the training process is formalized by minimizing the negative log-likelihood loss, where the model learns to predict each target token based on the entire source context and the previously generated tokens:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\sum_{t=1}^{T_y} \log P_\theta(y_t \mid y_{<t}, x)$$

In parallel, LoRA is employed as a cost efficient adaptation mechanism, where the original model weights are frozen, and learning occurs solely through the insertion of additional low-rank matrices. Specifically, each original weight matrix W_0 is updated as the sum of itself and a low-rank corrective component:

$$W = W_0 + BA, \quad \mathrm{rank}(A), \mathrm{rank}(B) \ll d$$

This approach allows the model to acquire new knowledge with a very small number of additional trainable parameters, thereby significantly reducing computational and memory costs compared to full parameter fine-tuning.

The training data was constructed from parallel English-Vietnamese (en-vi) and Vietnamese-English (vi-en) corpora, ensuring symmetry in the bidirectional training process. In addition, the team conducted benchmarking against strong models such as GPT-4o and GPT-4.1 on the validation set, providing a clear reference point for the performance of the proposed model. Experimental results show that the best performing model was trained using full parameter fine tuning, achieving substantial improvements in both translation directions and outperforming GPT-4.1, reaching 7.13 BLEU for English-Vietnamese and 8.21 BLEU for Vietnamese-English.

From a methodological perspective, LoRA proves particularly effective for general purpose tasks, where the main objective is to adapt the model to a new task while retaining most of its pre-learned reasoning and background knowledge. By updating only a very small subset of parameters, LoRA reduces training costs, mitigates catastrophic forgetting, and is especially suitable for multi-domain settings or data that is not highly specialized.

However, in the biomedical domain characterized by complex domain specific terminology, specialized linguistic structures, and strict semantic accuracy requirements full parameter fine tuning often yields superior performance. This is because domain specific data enables the model to adjust the entire

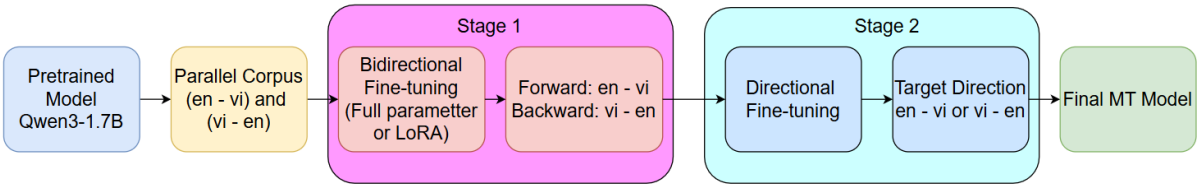

Figure 1. Bidirectional training steps..

parameter space, thereby learning semantic representations and translation patterns that are closely aligned with medical knowledge something that low-rank updates via LoRA struggle to achieve. In such contexts, accepting a trade off in generality in exchange for higher domain specific fidelity is a reasonable and effective choice.

Finally, the experimental results also indicate that combining full parameter fine tuning with effective prompt design can further enhance translation quality. This approach allows the model to fully leverage the internalized domain knowledge acquired during training, while better utilizing residual reasoning capabilities through well structured prompts. As a result, the model achieves superior performance compared to approaches relying solely on LoRA for biomedical machine translation.

4.3. BenignRhythms

BenignRhythms employs supervised fine tuning (SFT) combined with QLoRA to fine-tune the Qwen3-1.7B [30] model for both English-Vietnamese and Vietnamese-English translation directions. The team's innovation lies in integrating the fine tuned model with a Retrieval Augmented Generation (RAG) approach during inference. As a result, the model not only relies on its intrinsic knowledge but can also leverage external information particularly from a bilingual glossary to enhance the accuracy and consistency of translations, especially for domain specific terminology.

The process of retrieving glossary terms is carried out through a comprehensive pipeline. Each input sentence is first encoded using a sentence embedding model, after which the similarity between the sentence embedding and the embeddings of terms in the glossary is calculated. Terms with the highest similarity scores (top-k) that exceed a predefined threshold are selected to form a sentence-specific focused glossary. These terms are then inserted into the translation prompt, guiding the model to produce contextually accurate translations. For retrieval, MiniLM-L6-v2 [34] is used for English, while a Vietnamese SBERT [35] model handles retrieval for Vietnamese. Only documents or terms with cosine similarity above the predefined threshold are included, ensuring that the information provided to the model is of high quality and highly relevant.

Algorithm 1 Glossary Term Retrieval

procedure RETRIEVEGLOSSARY(x, D, k, τ)
  $u \leftarrow f(x)$ ▷ f: SBERT encoder
  for $i \leftarrow 1$ **to** $|D|$ **do**
    $s_i \leftarrow \cos(u, e(t_i))$ ▷ $e(t_i)$: embedding of glossary term t_i
  end for
  $T \leftarrow \text{TopK}(\{s_i\}_{i=1}^{|D|}, k)$ ▷ indices of top-k scores
  $T(x) \leftarrow \{t_i \mid i \in T \wedge s_i \geq \tau\}$
  $G(x) \leftarrow \{(t_i, D(t_i)) \mid t_i \in T(x)\}$
  return $G(x)$
end procedure

This approach offers several notable advantages. The innovative application of glossary driven prompting enhances translation

accuracy, particularly for domain specific terminology. The comprehensive data processing pipeline, ranging from sentence encoding to relevant term filtering, ensures stable system performance and scalability. Additionally, combining SFT with LoRA enables effective model fine tuning without updating all parameters, significantly reducing computational resource requirements.

However, the method also has certain limitations. The model primarily relies on SFT with LoRA applied to a moderately sized model, which may restrict its capacity to capture complex linguistic features. SBERT-based retrieval does not guarantee consistently high-quality results, and some critical terms may be overlooked. Furthermore, the absence of data augmentation techniques limits the model's generalization ability. Overall, the approach achieves a reasonable balance between translation quality and computational efficiency, yet it remains dependent on the quality of the glossary data and the SBERT model used for retrieval.

4.4. Team ZERO

Team ZERO developed a comprehensive system centered on a novel two stage fine tuning pipeline, aiming to maximize the performance of a constrained size language model while adapting it to a specialized bilingual translation task. Their approach demonstrates that with careful data curation, targeted supervised fine tuning, and reinforcement learning, it is possible to achieve significant improvements even on moderately sized models. This provides a practical blueprint for low resource adaptation and domain specific translation tasks, highlighting the effectiveness of staged training in extracting maximal performance from pre-trained language models.

Data Curation and Partitioning: The preprocessing phase was a critical component of the pipeline. The team applied a deduplication process based on **MinHash** [36] and **Locality-Sensitive Hashing (LSH)** [37] to identify and remove near-duplicate sentence pairs from the original 500,000-pair corpus. MinHash allows efficient estimation of Jaccard similarity between sets, which, when combined with LSH, enables the detection of highly similar or duplicate pairs without performing costly exhaustive comparisons. This step was particularly important in avoiding overfitting, improving data diversity, and reducing noise introduced by repeated examples. After this filtering, the corpus was reduced by 30.8%, yielding 346,000 high-quality, unique sentence pairs.

The curated data was then strategically partitioned:

- **331,000 pairs** for the initial Supervised Fine-Tuning (SFT) stage
- **15,000 pairs** for the subsequent reinforcement learning stage
- **3,000 pairs** for model evaluation

This partitioning strategy reflects a deliberate trade-off: the large SFT subset ensures the model learns robust general translation mappings, while the smaller GRPO subset enables targeted reward optimization for higher translation quality. The evaluation subset ensures the reported metrics reflect genuine generalization rather than overfitting to training data.

Two-Stage Training Methodology: At the heart of Team ZERO's pipeline is a sequential training methodology built upon the **Qwen2.5-3B-Instruct** base model.

The first stage follows a supervised fine-tuning approach on bidirectional translation pairs (English ↔ Vietnamese), analogous to strategies previously explored by Bosch@AI. This phase allows the model to establish strong foundational mappings between source and target languages, capturing lexical correspondences, syntactic structures, and domain-specific terminology. Supervised learning provides a stable initial point for subsequent reward-based optimization

and ensures that the model generates coherent translations with minimal divergence from reference sentences.

The second stage introduces **GRPO** [38] (Group Relative Policy Optimization), a reinforcement learning approach designed to fine tune model outputs by directly optimizing task specific rewards. Here, the policy parameters θ are updated to maximize expected rewards based on a weighted combination of BLEU and ChrF++ [39] scores:

$$\text{ChrF++} = (1 + \beta^2) \cdot \frac{P \cdot R}{\beta^2 P + R}$$

The model's outputs are contrasted against ground-truth references and refined through a composite reward function balancing lexical accuracy and morphological fluency, specifically:

- **BLEU Score** (70% weight): Rewarded the model for correctly translating key terms and phrases found in the reference.
- **chrF++ Score** (30% weight): Evaluated character-level fluency and grammatical correctness.

By applying GRPO, the system can adjust translations toward higher scoring outputs even when such improvements are subtle or context specific, effectively bridging the gap between purely supervised learning and human-like translation quality refinement.

Evaluation and Observations: The two stage SFT + GRPO pipeline achieved measurable improvements over the SFT only baseline. BLEU gains were +1.59 for English-Vietnamese and +1.48 for Vietnamese-English translations. Qualitative observations indicated that outputs were not only more fluent but also more accurate in terminology usage, reflecting the influence of reward driven fine tuning. The method demonstrates that reinforcement learning can effectively complement supervised fine tuning, especially in cases where lexical accuracy and fluency are not perfectly aligned.

Strengths and Contributions: Team ZERO leveraged the pre-trained Qwen2.5-3B-Instruct model, enabling high quality fine tuning with moderate computational resources. This demonstrates that small to mid sized models can achieve competitive translation performance when properly optimized. The proposed two stage pipeline effectively separates general language learning through supervised fine tuning (SFT) from reward driven optimization via GRPO, allowing precise control over model behavior and enabling targeted improvements in translation quality. Careful data curation and deduplication enhance the diversity and quality of training samples, which reduces overfitting and promotes better generalization. Furthermore, the combination of BLEU and ChrF++ in the reward system balances lexical fidelity with fluency, providing a principled framework for reinforcement learning in translation.

Limitations and Areas for Future Work: Despite these strengths, some limitations remain. The isolated effect of the deduplication pipeline was not explicitly evaluated, leaving open questions about its exact contribution to performance. The 70-30 weighting between BLEU and ChrF++ was chosen heuristically; exploring alternative weightings or integrating semantic oriented reward metrics, such as BERTScore or COMET, may further improve translation quality. Comparisons were limited to SFT only baselines, and additional ablation studie such as GRPO only experiments, varied reward functions, or different training schedules could provide deeper insights into the effectiveness of each component. Moreover, reliance on BLEU and ChrF++ may insufficiently penalize semantic errors, suggesting the need for semantic aware evaluation metrics. Finally, the pipeline's efficiency and scalability for larger datasets or larger models remain areas for future exploration.

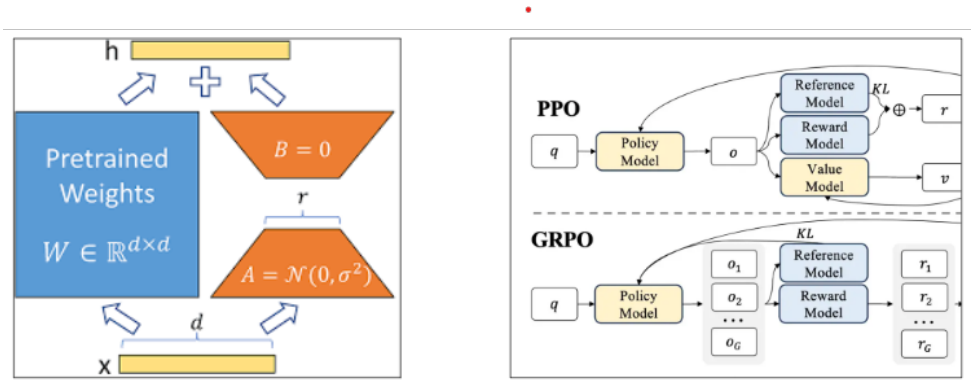

Figure 2. Using two training steps.

4.5. *JustGraduate*

The JustGraduate Team implemented a multi-stage training strategy, starting with data preprocessing, followed by SFT and finally Group Relative Policy Optimization (GRPO), similar to the approach adopted by the ZERO team. A key distinction in their method is the use of a *length penalty* combined with the BLEU score as a reward signal in GRPO. This design helps prevent outputs that are excessively long or short while guiding the model toward greater fluency, adherence to sentence structure, and improved stability in medical translation tasks.

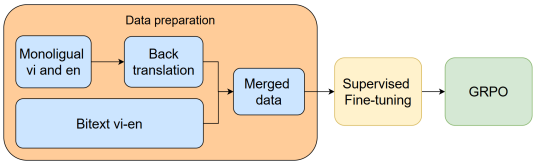

Figure 3. Back-translation for training.

During the data preparation phase for GRPO, the JustGraduate Team filtered sentence pairs based on semantic similarity. They reused the SFT dataset and applied a semantic embedding model to select only the most semantically relevant pairs. This ensured that the GRPO stage focused on high quality, domain specific examples, resulting in a carefully curated and balanced subset of approximately 50,000 sentence pairs, evenly divided between 25,000 Vi-En and 25,000 En-Vi pairs. Additionally, the team performed *semantic and BLEU-based deduplication* to remove duplicate or nearly identical sentence pairs, further enhancing the quality of the training data.

The final system combining SFT and GRPO demonstrated notable effectiveness. The model achieved BLEU scores of 43.38 for English to Vietnamese and 31.49 for Vietnamese to English, representing substantial improvements over the SFT only baseline, with gains of +2.46 (EN→VI) and +4.17 (VI→EN). Moreover, the integration of multiple techniques, including SFT, back-translation, and GRPO, helped optimize both BLEU scores and sentence structure, yielding stronger performance in specialized translation tasks.

However, the study has several limitations. The authors did not provide a clear justification for using QLoRA instead of traditional LoRA. Given that the experiments were conducted on a single 48GB GPU, which is generally sufficient to fine tune a 3B-parameter model using standard LoRA, the decision to employ 4-bit quantization via QLoRA warrants further explanation.

Another limitation related to GRPO is the model's heavy reliance on BLEU-based reward signals. While BLEU is a widely used metric in machine translation evaluation, it does not always

accurately reflect the naturalness and semantic correctness of medical translations. Additionally, the use of back-translation carries the risk of the model reinforcing its own translation errors if appropriate control mechanisms are not applied.

5. Result and Discussion

This section summarizes the competition outcomes under multiple evaluation protocols and discusses common patterns observed across submissions. Overall, participating teams achieved substantial gains over the baseline (Table 3), confirming that domain adaptation through data curation, supervised fine-tuning, and inference-time controls is critical for medical translation.

Table 4. SacreBLEU Leaderboard

Team	AVG	EN-VI	VI-EN
ZERO	42.6	59.8	25.4
Bosch@AI	37.8	50.6	25.1
JustGraduate	31.9	43.3	20.6
BenignRhythms	27.1	39.9	14.3

Table 5. Human preference Leaderboard

Team	AVG	EN-VI	VI-EN
Bosch@AI	85.7	88.2	83.8
ZERO	84.0	87.7	80.4
BenignRhythms	81.0	84.1	78.0
JustGraduate	77.1	74.4	79.9

Table 4 reports SacreBLEU on the shared-task benchmark. While the absolute SacreBLEU values remain moderate in the medical setting, all submitted systems outperform the zero-shot baseline by a wide margin. A consistent directionality gap is also observed across metrics: English→Vietnamese tends to score higher than Vietnamese→English. This pattern is compatible with the asymmetry of domain terminology, stylistic conventions, and available training signals.

Human evaluation (Table 5) serves as the primary indicator of practical translation quality in this task, as it better captures adequacy, clinical correctness, and fluency than n-gram overlap alone. Under human preference, **Bosch@AI** achieves the highest overall score, while **ZERO** is very close; the difference between these two systems is small, suggesting comparable quality at the top of the leaderboard.

Table 6. GPT-4o Leaderboard

Team	AVG	EN-VI	VI-EN
Bosch@AI	83.54	88.41	78.66
ZERO	80.02	88.87	71.17
BenignRhythms	74.40	84.19	64.6
JustGraduate	70.22	68.91	71.52

To provide an additional reference signal, we also report scores from GPT-4o (Table 6). Model-based evaluation can be useful for large-scale comparisons and may correlate with human judgments in many cases; however, it remains a supplementary measure rather than a replacement for expert human assessment, particularly in safety-critical medical contexts.

Discussion. Two broader observations emerge when comparing automatic and human-centered evaluations. First, the gap between SacreBLEU and human preference indicates that surface-form overlap may underestimate translation quality when multiple valid renderings exist (e.g., paraphrasing, reordering, and alternative acceptable terminology). This effect is amplified in medical translation, where correct meaning and terminology are essential but can be expressed with legitimate lexical variation.

Second, most successful systems rely on supervised fine-tuning (SFT) paired with carefully designed prompts, highlighting the importance of instruction formatting and consistent input-output structure. Differences in computational resources and training budgets make it difficult to attribute performance to a

single best practice; nonetheless, the results suggest that full-parameter fine-tuning can be advantageous for biomedical domain adaptation, as seen in the strong human preference scores of **Bosch@AI**.

We also observe that inference-time augmentation can improve perceived translation quality even when SacreBLEU gains are limited. For example, **BenignRhythms** integrates a glossary-driven retrieval mechanism, which can enhance terminology consistency and reduce critical term errors. Finally, reinforcement learning methods such as GRPO (used by **ZERO** and **JustGraduate**) show potential to refine outputs, but their effectiveness depends strongly on the reward design. Optimizing primarily for *n*-gram-based rewards (e.g., BLEU) may not fully capture medical adequacy and can diverge from human preference; incorporating semantic or preference-aligned signals remains an important direction for future work.

6. Conclusion and Future Work

The VLSP Medical Machine Translation Shared Task demonstrates the practicality of small language models (SLMs) for specialized English-Vietnamese translation. Across submissions, effective domain adaptation-including supervised fine-tuning, prompt engineering, and (in some cases) reinforcement learning-yields clear gains over the baseline and produces translations that are broadly usable under human evaluation.

Nevertheless, a persistent performance gap remains between the two translation directions, indicating that medical MT for English-Vietnamese is not yet a solved problem. Future work should focus on improving terminology fidelity and semantic adequacy, broadening coverage to additional medical subdomains and text types, and incorporating context-aware modeling beyond sentence-level translation.

7. Limitations

While the corpus provides a valuable resource for both training and evaluation of machine translation models, several limitations should be acknowledged.

First, despite the overall translation quality, the dataset is largely sentence-aligned and does not provide document-level context. In real clinical settings, multiple translations can be acceptable depending on audience, register, and surrounding discourse, and such contextual variation is not represented.

Second, the corpus focuses on written medical documents. While this scope supports strong domain specificity, it limits diversity in genre and interaction patterns. In particular, conversational and semi-structured sources-such as clinical dialogues, patient-doctor communication, and electronic health records-are not covered, which may reduce the robustness of models trained exclusively on this dataset.

References

[1] W. J. Hutchins. Machine translation: A brief history. In Concise History of the Language Sciences. Pergamon, 1995. doi: https://doi.org/10.1016/B978-0-08-042580-1.50066-0.

[2] P. Koehn and R. Knowles. Six challenges for neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation, pages 28–39. Association for Computational Linguistics, 2017. doi: https://doi.org/10.18653/v1/W17-3204.

[3] Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 845–850. Association for Computational Linguistics, 2017. doi: https://doi.org/10.3115/v1/P15-2139.

[4] Mariana Neves, Antonio Jimeno Yepes, and Aurélie Névéol. Domain adaptation in biomedical text mining: A survey. *Briefings in Bioinformatics*, 18(1):1–20, 2016. doi: https://doi.org/10.1016/j.jbi.2023.104418.

[5] Yujie Zhang, Qun Liu, and Yan Song. Improving neural machine translation in the medical domain via domain adaptation. *Journal of Biomedical Informatics*, 105:103421, 2020.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, and Dario Amodei. Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877–1901, 2020. URL https://mlanthology.org/neurips/2020/brown2020neurips-language/.

[7] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, and Guillaume Lample. LLaMA: Open and efficient foundation language models, 2023. arXiv:2302.13971.

[8] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. In Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing, 2020. doi: https://doi.org/10.48550/arXiv.1910.01108.

[9] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4163–4174. Association for Computational Linguistics, 2020. doi: https://doi.org/10.18653/v1/2020.findings-emnlp.372.

[10] Van Tai Nguyen, Huu Manh Nguyen, Thi Thu Pham, and Thanh Tung Vu. Goals, challenges and findings of the vlsp 2020 english-vietnamese news translation shared task. In Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing, 2020. URL https://aclanthology.org/2020.vlsp-1.18.pdf.

[11] Tran Hong Viet, Nguyen Minh Quy, and Nguyen Van Vinh. Vibidirectionmt - eval: Machine translation for vietnamese-chinese and vietnamese-lao language pair. *Journal of Computer Science and Cybernetics*, 41(3), 2025. doi: 10.15625/1813-9663/21055. URL https://vjs.ac.vn/jcc/article/view/21055.

[12] VLSP Organization. VLSP medical translation shared task, 2023. Dataset and task description.

[13] Chenhui Chu, Raj Dabre, and Sadao Kurohashi. An empirical comparison of domain adaptation methods for neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 385–391. Association for Computational Linguistics, 2017. doi: https://doi.org/10.18653/v1/P17-2061.

[14] Mariana Neves and Ulf Leser. A survey on biomedical natural language processing. *Briefings in Bioinformatics*, 17(1):31–43, 2016. URL https://doi.org/10.1093/bib/bbv059.

[15] Xiaojun Chen, Yining Zhang, Yifan Wang, Yuxuan Wang, Yining Wang, and Yifan Wang. Clinical domain machine translation: A survey. *Journal of Biomedical Informatics*, 108:103482, 2020. URL https://doi.org/10.1016/j.jbi.2020.103482.

[16] Yuxin Wu, Yining Zhang, Yifan Wang, and Yuxuan Wang. Machine translation for medical text: A systematic review. *JMIR Medical Informatics*, 7(4): e13016, 2019. URL https://medinform.jmir.org/2019/4/e13016/.

[17] Nhu Vo, Dat Quoc Nguyen, Dung D. Le, Massimo Piccardi, and Wray L. Buntine. Improving vietnamese-english medical machine translation. *ArXiv*, abs/2403.19161, 2024. URL https://api.semanticscholar.org/CorpusID:268732599.

[18] VLSP Organizing Committee. Vlsp 2023 shared task, 2023. URL https://vlsp.org.vn/.

[19] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, et al. Findings of the 2016 conference on machine translation (wmt16). In Proceedings of the First Conference on Machine Translation (WMT), pages 131–198, 2016. URL https://aclanthology.org/W16-2301.

[20] Yang Liu, Yifan Wang, and Yining Zhang. Shared tasks in machine translation: A review. *ACM Computing Surveys*, 54(8):1–36, 2021. URL https://doi.org/10.1145/3477600.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems (NeurIPS), pages 5998–6008, 2017. URL https://arxiv.org/abs/1706.03762.

[22] Shizhuo Zhang et al. Llm in a flash: Efficient large language model inference with limited memory. *arXiv preprint arXiv:2309.06180*, 2023. URL https://arxiv.org/abs/2309.06180.

[23] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023. URL https://arxiv.org/abs/2305.14314.

[24] Xian Li, Yifan Wang, and Yining Zhang. Efficient adaptation of language models for domain-specific tasks. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), pages 1234–1245, 2023. URL https://aclanthology.org/2023.acl-long.123.

[25] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2022. URL https://arxiv.org/abs/2106.09685.

[26] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training

with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2016. URL https://arxiv.org/abs/1511.06732.

[27] Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. An actor-critic algorithm for sequence prediction. In International Conference on Learning Representations (ICLR), 2017. URL https://openreview.net/forum?id=SJDaqqveg.

[28] Ke Lu, Yifan Wang, and Yining Zhang. Generalized reward policy optimization. *arXiv preprint arXiv:2305.10412*, 2023. URL https://arxiv.org/abs/2305.10412.

[29] Matt Post. A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191, 2018. URL https://aclanthology.org/W18-6319.

[30] Qwen Team. Qwen3 technical report. Technical report, Qwen / Alibaba, 2025. URL https://arxiv.org/abs/2505.09388.

[31] Zhen Xu, Sergio Escalera, Isabelle Guyon, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, and Huan Zhao. Codabench: Flexible, easy-to-use and reproducible benchmarking platform. *Patterns (Cell Press) / arXiv*, 2021. URL https://arxiv.org/abs/2110.05802.

[32] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1568–1575, 2016. doi: https://doi.org/10.18653/v1/D16-1163. URL https://aclanthology.org/D16-1163.

[33] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. URL https://arxiv.org/abs/2106.09685.

[34] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL https://arxiv.org/abs/1908.10084.

[35] thang1943. Vietnamese-sbert-v2. https://huggingface.co/thang1943/vietnamese-sbert-v2, 2024. Accessed: 2025-09-17.

[36] Andrei Z. Broder. On the resemblance and containment of documents, 1997. URL https://www.math.cmu.edu/~af1p/Texfiles/mw0.pdf. Min-wise hashing / MinHash idea (original work for near-duplicate detection).

[37] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In Proceedings of the 30th Annual ACM Symposium on Theory of Computing (STOC), 1998. URL https://web.math.princeton.edu/~naor/homepage%20files/lsh-new.pdf.

[38] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

[39] Maja Popović. chrf++: words helping character n-grams. In Proceedings of the Second Conference on Machine Translation (WMT), pages 612–618, 2017. URL https://aclanthology.org/W17-4770.

Appendix A. Evaluation System Prompt

```
system_prompt = """You are an expert
    evaluator for Medical Machine
    Translation (MT) systems.
Your task is to assess translations
    between English and Vietnamese in
    the medical domain.

### Evaluation Goals
You must provide a numerical score
    (0-100) reflecting the overall
    translation quality.
Your judgment should consider both
    linguistic and domain-specific
    criteria.

### Evaluation Criteria
1. Accuracy (40 points max)
- Correctly conveys the meaning of the
    source text.
- No critical omissions, additions, or
    distortions of meaning.
- Special focus on the correctness of
    medical content, such as symptoms,
    diagnoses, procedures, drug names,
    and acronyms.
- Penalize mistranslations that could
    cause misunderstanding in a medical
    context.

2. Terminology & Domain Appropriateness
     (25 points max)
- Correct use of medical terminology in
     the target language.
- Proper handling of abbreviations,
    Latin-based medical terms, and drug
    names.
- Consistent terminology usage across
    the text.
- Strong penalties if the translation
    invents, mistranslates, or confuses
    medical terms.

3. Fluency & Grammar (20 points max)
- The translation reads naturally and
    is grammatically correct in the
    target language.
- Proper syntax, word order, tense, and
     agreement.
- No major grammar or style errors that
     hinder comprehension.

4. Style & Register (10 points max)
- Maintains a professional, precise,
    and objective tone suitable for
    medical texts.
- Avoids colloquial or ambiguous
    phrasing inappropriate for medical
    documents.

5. Completeness & Formatting (5 points
    max)
- All information from the source text
    is preserved.
- Proper formatting of numbers, units,
    and special symbols (e.g., dosage
    units, percentages).

### Scoring Guide
Use the following levels when assigning
     scores:

- Perfect (90-100): Excellent, highly
    accurate and fluent medical
    translation with no major errors.
- Good (75-89): Good quality, minor
    errors present but meaning is
    preserved and terminology is mostly
    correct.
- Understandable (56-74): Fair,
    noticeable errors in meaning or
    terminology, but translation is
    still understandable overall.
- Partially understandable (31-55):
    Poor, multiple serious errors, some
    medical terms mistranslated, only
    partially understandable.
- Not understandable (0-30):
    Unacceptable, meaning is distorted,
    critical terminology mistranslated,
    or text is incomprehensible.

### Output Format
You must only output in the following
    JSON format:
{
"score": <integer between 0 and 100>,
"level": "<one of: Perfect / Good /
    Understandable / Partially
    understandable / Not understandable>
    "
}
"""
```

Listing 1. System prompt for medical translation evaluation