Vietnamese–English Medical Domain Machine Translation with LLMs and GRPO Optimization Using Verified Rewards

Dang Sy Duy[1] **Nguyen Duy Chien**[2] **Chu Ngoc Vuong**[1]
[1]NTQ Solutions [2]Independent Researcher
dduy193.cs@gmail.com duychien.work@gmail.com chungocvuong3@gmail.com

Abstract

Machine translation in the medical domain requires high accuracy and precise handling of specialized terminology, but this is difficult under limited-resource conditions. Challenges include the scarcity of bilingual medical data, frequent out-of-vocabulary terms, and the limited capacity of smaller models. We investigate English–Vietnamese and Vietnamese–English medical translation with base models from the Qwen 2.5 and Qwen 3 families, constrained to a maximum of 3B parameters. Our experiments include QLoRA fine-tuning for efficient domain adaptation and GRPO alignment to better match human translation preferences. Among the evaluated systems, Qwen2.5-3B-Instruct achieves the best overall performance on shared task datasets. Results show that compact instruction-tuned models, when adapted with efficient fine-tuning and alignment strategies, can deliver accurate and reliable medical translations, offering a practical solution for high-stakes domains under strict resource constraints.

1 Introduction

Machine Translation has advanced rapidly with the rise of Large Language Models (LLMs), which capture multilingual structures and contextual nuances with high accuracy. While these systems achieve strong results for high-resource language pairs supported by large parallel corpora, their effectiveness is reduced in specialized domains with limited bilingual data, where terminology coverage is incomplete and translation quality often diverges between directions.

English–Vietnamese medical translation illustrates this challenge. Despite the demand for accurate systems to support healthcare communication, high-quality bilingual data is scarce and fragmented. General-purpose systems such as Google Translate or NLLB-200 provide functional coverage but often misinterpret clinical terminology, leading to literal, inconsistent, or even misleading translations that are unsuitable for sensitive contexts.

To address this, we evaluate compact open-source LLMs from the Qwen 2.5 and Qwen 3 families, all constrained to a maximum of 3B parameters. We investigate multiple adaptation strategies. First, we explore continued pretraining (CPT) with QLoRA, where the model is further exposed to large amounts of unlabeled medical text, enabling efficient integration of domain-specific knowledge without requiring extensive resources. As a separate line of experimentation, we apply supervised fine-tuning (SFT) combined with GRPO alignment, which leverages limited bilingual data and preference-based optimization to improve translation adequacy, fluency, and alignment with human judgments. In addition, we incorporate back-translation to expand the training data with synthetic bilingual pairs and apply a length penalty to encourage more natural sentence outputs.

Our results show that Qwen2.5-3B-Instruct, trained with supervised fine-tuning and GRPO and further improved through back-translation with length penalty, achieves the strongest overall performance on the shared task datasets. This configuration yields higher terminology fidelity, more fluent sentence structure, and more reliable translations compared to smaller models and alternative strategies such as CPT with QLoRA. These findings suggest that compact instruction-tuned models, when adapted with carefully selected methods, can be effectively specialized for medical translation under resource constraints, offering a practical solution for high-stakes domains such as healthcare.

2 Related Works

2.1 Low-Resource MT Techniques

Neural machine translation (NMT) models tend to underperform on language pairs with limited par-

allel data, prompting strategies such as **data augmentation** (e.g., back-translation (Sennrich et al., 2016)) and **transfer learning** from high-resource "parent" models (Ekle and Das, 2025). These approaches leverage shared linguistic representations and have yielded substantial gains in many low-resource settings (Zoph et al., 2016).

For the medical domain, Vo et al. (2024) improved BLEU by 4.94% over Google Translate using vinai-translate with MedEV—a high-quality Vietnamese–English parallel dataset constructed specifically for the medical domain. This shows that carefully curated in-domain corpora can yield measurable gains over general-purpose systems, but also highlights the dependence of supervised approaches on costly and labor-intensive dataset construction.

2.2 Multilingual Pretrained MT Models

Large multilingual models such as **mBART** (Liu et al., 2020), **mT5** (Xue et al., 2021), **M2M-100** (Fan et al., 2020), and **NLLB-200** (Team et al., 2022) achieve strong low-resource performance via extensive multilingual pretraining. Xiaomi's **GemmaX** family (Cui et al., 2025) follows this paradigm with a decoder-only architecture (2B–7B), translation-specific instruction tuning, and over 100 language pairs, providing competitive BLEU in low-resource MT while remaining deployable on modest hardware.

2.3 Adapting General-Purpose LLMs for Translation

General-purpose **LLMs** such as GPT-4, LLaMA, Qwen, and GemmaX have shown strong translation capabilities when used with prompting or fine-tuning (Sizov et al., 2024; Vashee, 2024). However, in specialized domains such as medicine, these models often struggle with terminology fidelity and domain-specific semantics due to limited exposure during pretraining. To address these issues, compact open-source models can be adapted with efficient methods such as **parameter-efficient fine-tuning (e.g., QLoRA)** (Dettmers et al., 2023), which enables domain adaptation even under strict computational constraints. Another promising line of work uses preference-based reinforcement learning methods such as **Direct Preference Optimization (DPO)** (Rafailov et al., 2024), **Relative Preference Optimization (RPO)**, or exploration-based algorithms like **GRPO**, which optimize model outputs to align more closely with human translation quality. Beyond fine-tuning and alignment, data augmentation techniques such as **back-translation** and decoding strategies including **length penalties** have also been widely applied to improve fluency and coverage in low-resource or domain-specific MT scenarios.

3 Background

3.1 Neural Machine Translation Fundamentals

Neural Machine Translation (NMT) casts translation as a conditional sequence generation task. Given a source sentence $\mathbf{x} = (x_1, \ldots, x_m)$, the model generates a target sequence $\mathbf{y} = (y_1, \ldots, y_n)$ according to:

$$P(\mathbf{y} \mid \mathbf{x}; \theta) = \prod_{t=1}^{n} P(y_t \mid y_{<t}, \mathbf{x}; \theta) \qquad (1)$$

where θ are the model parameters. Modern systems predominantly use Transformer architectures (Vaswani et al., 2023) with an encoder-decoder design, in which the encoder produces contextual representations of the source and the decoder autoregressively generates target tokens while attending to both source and past outputs. This architecture effectively captures long-range dependencies and complex linguistic relationships.

3.2 Low-Resource Learning Principles

In low-resource scenarios, parallel data scarcity limits the ability to train high-capacity models without overfitting. Limited samples increase variance in parameter estimation, leading to unstable predictions and degraded translation quality. To mitigate this, prior work has explored:

- **Multilingual pretraining:** initializing from models trained on large, diverse multilingual corpora to inherit broad cross-lingual representations.
- **Data augmentation:** expanding training coverage with synthetic bitext, such as back-translation or pivot-based generation.
- **Alignment-based optimization:** directly refining model outputs toward desired fluency and adequacy using preference-based objectives, without requiring a large supervised fine-tuning stage.

3.3 Policy Gradient and GRPO

Reinforcement Learning from Human Feedback (RLHF) frames text generation as a policy π_θ producing a sequence y given input x, with a scalar reward $R(y)$ reflecting quality. The learning goal is to maximize the expected reward via policy gradients:

$$\nabla_\theta J(\theta) \approx \frac{1}{K}\sum_{k=1}^{K}\big(R(y^{(k)}) - b\big) \times \nabla_\theta \log \pi_\theta\big(y^{(k)}\big), \quad (2)$$

where b is a baseline to reduce gradient variance.

PPO. In Proximal Policy Optimization (PPO) (Schulman et al., 2017), this baseline comes from a learned *critic* network estimating token-level advantages $\hat{A}_t$:

$$\mathcal{J}_{\text{PPO}} = \mathbb{E}\left[\frac{1}{|y|}\sum_{t=1}^{|y|}\min\Big(w_t\hat{A}_t, \text{clip}(w_t, 1-\epsilon, 1+\epsilon)\hat{A}_t\Big)\right], \quad (3)$$

where

$$w_t = \frac{\pi_\theta(y_t \mid y_{<t}, x)}{\pi_{\theta_{\text{old}}}(y_t \mid y_{<t}, x)}, \quad (4)$$

$$\hat{A}_t = \text{Advantage from critic at token } t. \quad (5)$$

GRPO. Group Relative Policy Optimization (GRPO) (Shao et al., 2024) completely eliminates the critic. Instead, it sets the baseline b as the *group average reward* over K candidates:

$$b = \frac{1}{K}\sum_{j=1}^{K} r(y_j), \quad \text{(group baseline)} \quad (6)$$

$$\hat{A}_i = \frac{r(y_i) - b}{\text{std}(\{r(y_j)\}_{j=1}^{K})}, \quad \text{(normalized advantage)} \quad (7)$$

where $r(y)$ is the scalar reward for a complete translation.

The GRPO objective then mirrors PPO but shares the same $\hat{A}_i$ for all tokens in candidate y_i:

$$\mathcal{J}_{\text{GRPO}} = \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min\Big(w_{i,t}\hat{A}_i, \text{clip}(w_{i,t}, 1-\epsilon, 1+\epsilon)\hat{A}_i\Big)\right], \quad (8)$$

$$w_{i,t} = \frac{\pi_\theta(y_{i,t} \mid y_{i,<t}, x)}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid y_{i,<t}, x)}, \quad (9)$$

(likelihood ratio for token t in candidate i).

This design reinforces translations scoring above the group mean and suppresses those below, yielding stable optimization without the complexity of training a critic.

3.4 Efficient Finetuning with QLoRA

QLoRA (Dettmers et al., 2023) enables fine-tuning large models under limited hardware by combining low-rank adaptation (LoRA) with 4-bit quantization. Given pretrained weights $W_0 \in \mathbb{R}^{d\times k}$, LoRA introduces a low-rank update

$$\Delta W = BA, \quad A \in \mathbb{R}^{r\times k},$$

$$B \in \mathbb{R}^{d\times r}, \quad r \ll \min(d, k)$$

so that the adapted weights become

$$W = W_0 + \Delta W.$$

During training, W_0 is frozen and quantized, while only A, B are updated in higher precision. For input $x \in \mathbb{R}^k$, the forward pass is

$$h = W_0 x + BAx,$$

where $W_0 x$ uses quantized matrix multiplication. This reduces memory usage while retaining performance, enabling domain-specific adaptation of compact LLMs.

4 Methodology

4.1 Overview

We follow a three–stage Vi–En medical MT pipeline—SFT on parallel data, BT to augment with synthetic pairs, and GRPO for preference-based alignment—starting GRPO from the SFT + BT checkpoint.

Central to this process is the *reward function*, which combines automatic translation metrics with penalty terms to discourage undesirable behaviors such as excessive length deviation or semantic drift. This scalar reward reflects overall translation quality, balancing fluency, adequacy, and stylistic consistency. For each source sentence, the model generates K candidate translations, each scored by the reward function. GRPO computes a normalized advantage for each candidate relative to the group mean, reinforcing outputs above average and suppressing those below.

By bypassing an explicit supervised adaptation stage, our pipeline focuses on exploiting the base model's multilingual prior and directly steering its

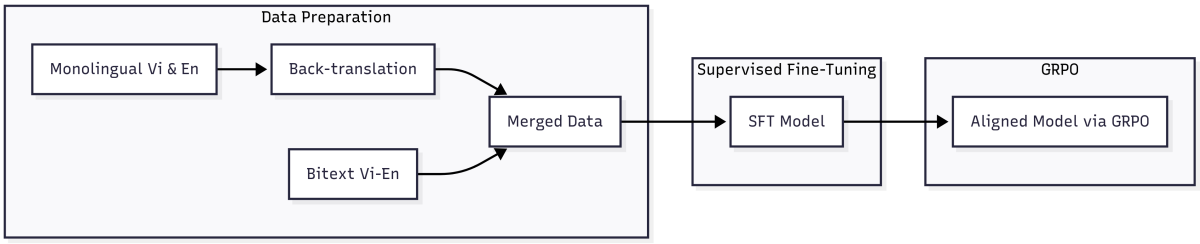

Figure 1: The pipeline consists of three main stages. First, monolingual Vietnamese and English corpora are augmented through back-translation and combined with available bilingual data to form a merged training set. Second, this data is used to perform supervised fine-tuning (SFT), grounding the model in domain-specific bilingual mappings. Finally, the SFT model is refined with Group Relative Preference Optimization (GRPO), where preference-based rewards such as BLEU and length penalties align outputs for greater fluency, structural fidelity, and reliability in medical translation.

outputs toward human-preferred translations. This strategy allows exploration of diverse phrasings while maintaining source fidelity, resulting in a translation system tuned for both quality and deployability in low-resource settings.

4.2 Supervised Fine-Tuning (SFT) Stage

The first stage applies **supervised fine-tuning** (SFT) to adapt the base model for bidirectional Vietnamese–English (**Vi↔En**) medical translation. Let x be a source sentence and $y = [y_1, \ldots, y_{|y|}]$ its target translation. Each x is wrapped in a prompt template $I(x)$, ensuring consistent instruction-following in both *Vi* $\rightarrow$ *En* and *En* $\rightarrow$ *Vi*.

Training uses bilingual *bitext*, supplemented with **backtranslation** (Sennrich et al., 2016): English monolingual sentences are translated into Vietnamese and vice versa. This augmentation broadens domain coverage, increases stylistic diversity, and balances both translation directions.

Objective. Let $\mathcal{D}^{\text{bi}}_{\text{bitext}}$ be the merged bidirectional dataset, containing both *Vi* $\rightarrow$ *En* and *En* $\rightarrow$ *Vi* pairs. The SFT objective is the negative log-likelihood:

$$\mathcal{L}_{\text{SFT}}(I(x), y; \theta) = -\log P_\theta(y \mid I(x)) \quad (10)$$

(full sequence loss)

$$= -\sum_{k=1}^{|y|} \log P_\theta\big(y_k \mid y_{<k}, I(x)\big) \quad (11)$$

(token-level form),

where:

- $I(x)$ — instruction-formatted prompt containing the source sentence x,
- y_k — the k-th target token,
- $y_{<k}$ — preceding target tokens,
- θ — model parameters.

This stage focuses purely on maximizing translation fidelity and fluency in both directions, providing the model with a strong, instruction-aligned translation capability before applying preference-based optimization.

4.3 Reward Function Design

To guide the model toward human-preferred translations, we define a scalar reward that balances accuracy and output length. Given an instruction-formatted source $I(x)$ and its reference translation y_{ref}, the model generates K candidates $\{y^{(1)}, \ldots, y^{(K)}\}$. Each candidate is scored as:

$$R(y^{(k)}) = \text{BLEU}\big(y^{(k)}, y_{\text{ref}}\big) - \lambda \cdot \frac{\big|\,|y^{(k)}| - |y_{\text{ref}}|\,\big|}{|y_{\text{ref}}|}, \quad (12)$$

where the first term measures n-gram precision and the second penalizes deviations in length, with λ controlling penalty strength. We also experimented with a BLEU-only reward, which improved performance for short, formulaic sentences but failed to control verbosity consistently — motivating the inclusion of the length penalty in the final formulation.

The *group baseline* for variance reduction is the mean reward across all K candidates:

$$b = \frac{1}{K} \sum_{j=1}^{K} R\big(y^{(j)}\big). \quad (13)$$

4.4 GRPO Optimization Process

We adopt **Group Relative Policy Optimization (GRPO)** (Shao et al., 2024) to directly align model outputs with our reward function, bypassing the need for a critic. GRPO computes the relative advantage $R(y^{(k)}) - b$ for each candidate and updates the model parameters via:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x,\, y^{(k)} \sim \pi_\theta} \left[\frac{1}{K} \sum_{k=1}^{K} \big(R(y^{(k)}) - b\big) \times \nabla_\theta \log \pi_\theta\big(y^{(k)} \mid I(x)\big) \right], \quad (14)$$

where π_θ is the model policy. Candidates scoring above the baseline are reinforced, while those below are suppressed. Rewards are computed for both *Vi→En* and *En→Vi* directions for balanced improvements.

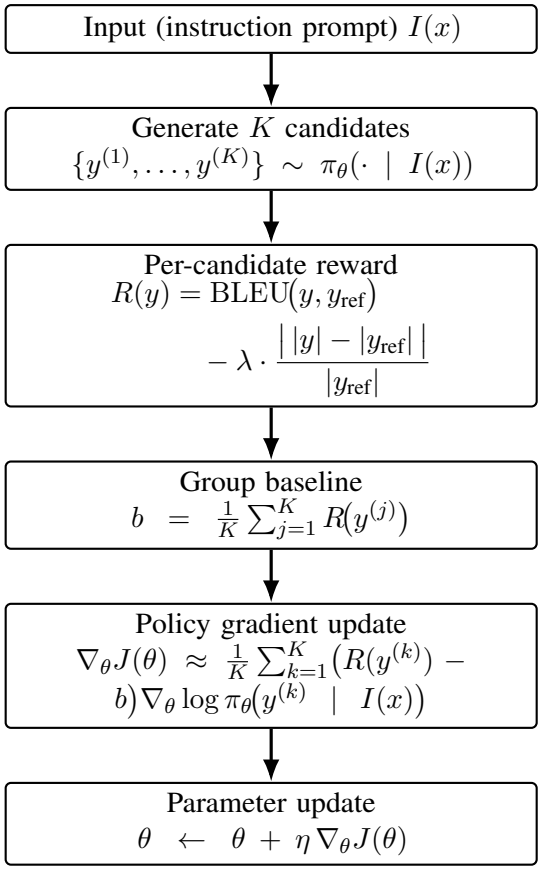

Figure 2: GRPO pipeline for translation optimization. The model generates K candidates, computes BLEU + length-penalized rewards, applies a group baseline, and updates parameters based on relative advantages.

5 Datasets Resource

5.1 Data Collection

To develop a robust Vietnamese–English medical domain translation system, we curated a multi-stage dataset pipeline entirely based on the resources released by the shared task organizers.[1] The dataset preparation followed two key phases: supervised fine-tuning and reinforcement learning with GRPO.

Supervised Fine-tuning:
The supervised fine-tuning utilized a primary Vietnamese–English parallel corpus, provided by VLSP 2025. To ensure data quality and consistency, datasets were processed through a comprehensive multi-step cleaning pipeline.

- *Punctuation Filtering:* We excluded sentence pairs with abnormal punctuation patterns to minimize noise in the training data.
- *Deduplication:* Exact duplicates were identified and removed through hashing, ensuring data diversity and reducing overfitting risks.
- *Quality-Based Filtering* We applied sentence-level filtering using SacreBLEU (Post, 2018). Each source sentence was translated with a preliminary baseline model, and the hypothesis was compared against the provided target side as reference. Sentence pairs with BLEU scores below a defined threshold were removed, improving the overall quality of the parallel data.[2]

Each step progressively refined the datasets by discarding low-quality or problematic data. The final dataset sizes and the detailed statistics of sentence pairs remaining and removed at each stage for both corpora are summarized in Table 1.

GRPO-based Reinforcement Learning:
In this stage, we did not introduce additional datasets such as ViPubmed. Instead, we reused the bilingual corpus from the SFT + BT stage and applied a semantic embedding model to filter for the most semantically relevant pairs. This ensured that GRPO optimization focused on high-quality, domain-relevant translations while discarding noisy or weakly aligned examples. From this filtering, we obtained a balanced subset of approximately

[1] We strictly adhered to the competition rules and did not use any external parallel corpora beyond the official dataset.

[2] The baseline system was trained only on organizer-provided data; no external resources were used.

Step	VLSP2025	
	Remaining	Removed
Initial Dataset	500,000	–
Punctuation Filtering	500,000	0
De-duplication	369,224	130,776
Quality-Based Filtering	358,601	10,623
Final Dataset	358,601	141,399

Table 1: Data cleaning pipeline statistics showing sentence pair counts and removal percentages at each processing stage for VLSP2025 and ViPubmed datasets.

50,000 sentence pairs (25,000 Vi–En and 25,000 En–Vi). This curated subset was then employed for GRPO-based fine-tuning, where preference-based rewards such as BLEU and length penalties guided the model toward greater fluency, structural fidelity, and robustness in medical translation.

5.2 Data Augmentation

We augmented the Vietnamese–English medical corpus through bidirectional backtranslation, using only the monolingual datasets released by VLSP 2025. English medical texts were translated into Vietnamese to create synthetic En→Vi pairs, while Vietnamese texts were translated into English for Vi→En. Synthetic data was generated with our intermediate SFT model trained exclusively on organizer-provided data.[3] This approach increases coverage of domain-specific terminology such as drug names, diseases, and clinical procedures that are often underrepresented in parallel corpora. Backtranslation also improves balance between translation directions, broadens stylistic variety, and strengthens generalization to unseen inputs. To reduce noise, synthetic pairs were further filtered using automatic quality metrics to retain only high-confidence translations.

6 Experiments

6.1 Experiment Setup

To validate our two-stage Vietnamese–English (Vi↔En) medical translation pipeline, we executed each component sequentially and evaluated performance at every step. This staged evaluation isolates the contribution of each method. We first established a baseline using Supervised Fine-Tuning (SFT) on bilingual and backtranslated data. Separately, we explored Parameter-Efficient Fine-Tuning (PEFT) with QLoRA to efficiently integrate domain-specific knowledge. Finally, we applied Group Relative Policy Optimization (GRPO) to align model outputs with preference-based reward signals. This setup allows comparison between likelihood-based and preference-aligned training, highlighting gains in fluency, adequacy, and style.

For bidirectional translation, the *prompt format* is key. Instead of a single fixed template, we created multiple instruction-style prompts explicitly specifying source and target languages. Variations prevent overfitting to phrasing and improve robustness. Prompts use the `messages` format, with `role` and `content` fields, matching the LLM's chat-based training. Each source sentence x is embedded into a prompt $I(x)$ paired with reference y, ensuring consistency between training and inference and enabling both translation directions within a single model.

In GRPO, the reward combines BLEU with a length penalty to avoid outputs that are too long or short. Ablation with BLEU-only shows improvements over SFT, but adding the length penalty yields more concise, fluent, and structurally faithful translations, particularly in En→Vi.

6.2 Implementation Detail

All experiments using standard Supervised Fine-Tuning (SFT) were implemented with the HuggingFace `Trainer` API for sequence-to-sequence training. For QLoRA experiments, we employed the `PEFT` library to perform parameter-efficient fine-tuning on the base model.

Training was conducted on a single GPU with 48GB of VRAM using mixed-precision (`fp16`) to reduce memory usage and accelerate computation. Batch sizes and sequence lengths were tuned to fully utilize GPU memory while maintaining sta-

[3] No external translation models or corpora were used for backtranslation.

Model	En→Vi	Vi→En
envit5-translation†	42.86	31.33
ChatGPT (zero-shot)†	34.38	29.79
SeamlessM4T-medium†	31.04	21.57
QLoRA	37.80	23.35
QLoRA + GRPO (BLEU-only)	41.96	27.44
QLoRA + GRPO (BLEU + length penalty)	42.55	28.94
SFT	40.92	27.32
SFT + GRPO (BLEU-only)	42.54	29.78
SFT + GRPO (BLEU + length penalty)	**43.38**	**31.49**

Table 2: BLEU scores for external baselines and our experimental models on the Vi↔En test set. †External baseline scores were reported on the same larger dataset but with a different evaluation split; therefore, results are not strictly comparable to ours.

bility. For GRPO, reward signals combined BLEU with a length penalty and were computed during training. Backtranslated data were filtered for quality before inclusion, and early stopping based on validation BLEU was applied to prevent overfitting.

7 Results

Model Variant	En→Vi	Vi→En
SFT	40.92	27.32
+ GRPO (BLEU-only)	42.54	29.78
+ GRPO (BLEU+len)	**43.38**	**31.49**

Table 3: Incremental BLEU improvements from GRPO over the SFT baseline.

We evaluate translation quality of the proposed Vietnamese–English (Vi↔En) MT system against strong external baselines. All experiments are conducted on a held-out test set, with evaluation in *both* directions (*Vi→En* and *En→Vi*).

7.1 Evaluation Metrics

We report translation quality using **BLEU** (Papineni et al., 2002), computed with SacreBLEU (Post, 2018) for reproducibility.

7.2 Baselines and Stages

We compare against:

- **envit5-translation** — a transformer-based Vietnamese–English translation model from VietAI.
- **ChatGPT (zero-shot)** — evaluated without task-specific fine-tuning.
- **SeamlessM4T-medium** — a multilingual translation model released by Meta.

and evaluate our models across the following stages:

- **QLoRA** — parameter-efficient continual pretraining on medical text.
- **QLoRA + GRPO (BLEU-only)** — preference optimization using BLEU as reward.
- **QLoRA + GRPO (BLEU + length penalty)** — adds a length-regularization term to encourage natural outputs.
- **SFT** — supervised fine-tuning on parallel and backtranslated data (full configuration in the appendix table 4).
- **SFT + GRPO (BLEU-only)** — SFT model further aligned with BLEU-based preference optimization.
- **SFT + GRPO (BLEU + length penalty)** — extends BLEU optimization with length penalty for fluency.

7.3 Observations

Table 2 shows clear improvements across the pipeline stages. Among external baselines, `envit5-translation` achieves the strongest BLEU scores (42.86 En→Vi, 31.33 Vi→En), while ChatGPT and SeamlessM4T-medium perform substantially lower, highlighting the challenge of medical-domain translation for general-purpose systems.

For QLoRA experiments, parameter-efficient continual pretraining alone yields modest gains (37.80 BLEU En→Vi), showing that domain adaptation helps but is insufficient by itself. Adding GRPO improves performance, and length penalty provides a slight additional boost, particularly in Vi→En.

Supervised Fine-Tuning (SFT) is a much stronger baseline, already surpassing QLoRA alone. Table 3 further breaks down the incremental effect of GRPO over SFT: GRPO adds +1.6 BLEU (En→Vi) and +2.5 BLEU (Vi→En), while the length penalty contributes a smaller +0.8 and +1.7 BLEU respectively. This indicates that most of the gain arises from reinforcement-based alignment, with length penalization acting mainly as a stabilizer.

Overall, the results demonstrate that combining SFT with GRPO achieves the best performance (43.38 BLEU En→Vi, 31.49 Vi→En). The progression across experiments indicates that supervised signals and preference optimization are complementary, addressing issues such as literalism, verbosity, and phrasing quality.

8 Discussion

8.1 Limitations

Despite these promising results, several limitations remain. First, the evaluation is restricted to BLEU scores, which primarily capture n-gram overlap but do not fully reflect semantic adequacy or terminology correctness. This makes it difficult to assess translation safety in critical medical contexts. Second, while the models demonstrate consistent gains, the absolute improvements in Vi→En remain modest, suggesting that reverse-direction translation is more sensitive to data coverage and alignment strategies. Third, GRPO training introduces additional complexity in terms of hyperparameter sensitivity and computational overhead, which may limit reproducibility or scalability to larger settings. Finally, the reliance on synthetic backtranslation data could bias the model toward overfitting generated patterns, reducing robustness to real-world clinical text.

8.2 Future Work

While our pipeline demonstrates strong performance, several directions remain for improvement. First, exploring larger or more diverse medical corpora could further enhance domain coverage. Second, extending GRPO reward functions to incorporate semantic similarity metrics or terminology accuracy could reduce errors in clinical terms. Third, investigating multi-task or multilingual adaptation might allow simultaneous translation across multiple language pairs without compromising quality. Finally, integrating uncertainty estimation could help flag translations that require human verification in high-stakes medical contexts.

Another promising direction is to explore hybrid evaluation frameworks that combine automatic metrics with domain-expert feedback. Incorporating physician or translator annotations could expose systematic weaknesses in terminology fidelity, pragmatic correctness, or ambiguity handling. Additionally, scaling the pipeline to other low-resource medical languages could test the generalizability of the approach and highlight whether the synergy between SFT and GRPO is universally beneficial across linguistic and cultural boundaries.

8.3 Conclusion

We present a resource-efficient, two-stage Vietnamese–English medical translation pipeline combining Supervised Fine-Tuning, parameter-efficient QLoRA adaptation, and preference-based GRPO optimization. Experiments show that SFT with GRPO and length-penalized rewards achieves the strongest performance across both translation directions. Our results demonstrate that compact LLMs, when carefully adapted and aligned with preference-based signals, can provide reliable, fluent, and domain-accurate translations suitable for medical applications under computational and data constraints.

In summary, this work confirms that the combination of data-driven supervised training and preference-guided reinforcement represents a practical and scalable solution for domain-specific translation. While challenges remain in evaluation depth and robustness, the demonstrated gains suggest that compact architectures with tailored adaptation strategies can serve as a viable alternative to general-purpose large-scale models in high-stakes environments such as healthcare.

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. *Preprint*, arXiv:2502.02481.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Ocheme Anthony Ekle and Biswarup Das. 2025. Low-resource neural machine translation using recurrent neural networks and transfer learning: A case study on english-to-igbo. *Preprint*, arXiv:2504.17252.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *Preprint*, arXiv:2010.11125.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting bleu scores. *Preprint*, arXiv:1804.08771.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. 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. *Preprint*, arXiv:2402.03300.

Fedor Sizov, Cristina España-Bonet, Josef Van Genabith, Roy Xie, and Koel Dutta Chowdhury. 2024. Analysing translation artifacts: A comparative study of LLMs, NMTs, and human translations. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1183–1199, Miami, Florida, USA. Association for Computational Linguistics.

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, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Kirti Vashee. 2024. The evolving path to LLM-based MT. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)*, pages 18–18, Chicago, USA. Association for Machine Translation in the Americas.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Nhu Vo, Dat Quoc Nguyen, Dung D. Le, Massimo Piccardi, and Wray Buntine. 2024. Improving vietnamese-english medical machine translation. *Preprint*, arXiv:2403.19161.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

Stage	Configuration
SFT	Optimizer: AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay $= 0.01$) Learning rate: 2×10^{-5} with linear decay Batch size: 64 sequences (grad. accumulation $= 8$) Max sequence length: 512 tokens Epochs: 5 (early stopping, patience = 2)
QLoRA	Quantization: 4-bit NF4 with double quantization LoRA rank $r = 64$, $\alpha = 128$, dropout $= 0.05$ Target modules: linear projections in attention & MLP blocks Batch size: 128 sequences (enabled by quantization) Epochs: 5
GRPO	Candidate number $K = 4$ per source sentence Reward: BLEU – $\lambda\cdot$ length penalty (Eq. 12) Optimizer: Adam, learning rate 1×10^{-6} Training steps: 20,000 ($\approx$ 2 epochs over GRPO subset) Clip parameter $\epsilon = 0.2$ Batch size: 16 prompts (4 candidates each, total 64 samples/step)

Table 4: Detailed hyperparameter settings for SFT, QLoRA, and GRPO training.

Appendix

A Hyperparameter Settings

To ensure reproducibility, Table 4 summarizes the detailed configurations used in all training stages, including supervised fine-tuning (SFT), parameter-efficient fine-tuning (QLoRA), and GRPO alignment.

All experiments were run on a single NVIDIA A6000 GPU (48GB VRAM) with mixed-precision (fp16). QLoRA was adopted to increase effective batch size, enable longer sequence lengths (up to 1,024 tokens) during GRPO, and ensure reproducibility on more resource-constrained hardware.

B Length Penalty Sensitivity

In the reward function (Eq. 12), we set $\lambda = 1.0$ for all experiments, which we found to be a stable choice balancing BLEU gains with verbosity control.

To assess sensitivity, we ran an ablation varying $\lambda \in \{0.5, 1.0, 2.0\}$. Results (Table 5) show that $\lambda = 1.0$ consistently yields the best trade-off. Lower values reduce the effectiveness of verbosity control, while higher values over-penalize outputs, occasionally truncating translations.

λ	En→Vi BLEU	Vi→En BLEU
0.5	43.12	30.88
1.0	**43.38**	**31.49**
2.0	42.70	30.92

Table 5: Effect of λ on BLEU scores for the SFT+GRPO model.

C Additional Evaluation Metrics

To complement BLEU, we include further evaluation dimensions covering semantic similarity, terminology fidelity, and human validation.

C.1 BLEU and Semantic Similarity

We report BLEU together with cosine similarity computed from Qwen3-Embedding-4B(Zhang et al., 2025) to capture both surface-level overlap and semantic alignment. Results in Table 6 show that GRPO consistently improves over SFT. While length penalization contributes to modest gains by discouraging verbosity, the majority of improvements are attributable to the model's enhanced ability to phrase translations more effectively.

Metric	Model	En→Vi	Vi→En
BLEU	SFT	40.92	27.32
	SFT + GRPO (*BLEU-only*)	42.54	29.78
	SFT + GRPO (*BLEU+len*)	**43.38**	**31.49**
Cosine Similarity	SFT	0.894	0.862
	SFT + GRPO (*BLEU-only*)	0.901	0.869
	SFT + GRPO (*BLEU+len*)	**0.905**	**0.873**

Table 6: BLEU and semantic similarity (cosine similarity using Qwen3-Embedding-4B) for SFT and GRPO models.

C.2 Terminology Fidelity

To assess domain fidelity, we extracted 200 sentences containing medical terminology (e.g., drug

names, diagnoses, procedures) and measured exact-match accuracy. As shown in Table 7, GRPO with length penalty improves terminology consistency over SFT, reducing mistranslations of critical domain terms.

Model	**En→Vi**	**Vi→En**
SFT	87.2%	83.5%
SFT + GRPO (*BLEU+len*)	**90.1%**	**85.6%**

Table 7: Terminology-level accuracy on a medical subset.

D Additional Evaluation Metrics

To complement BLEU, we include further evaluation on semantic similarity and terminology fidelity. A human expert evaluation was not conducted due to the lack of available medical annotators, which we identify as an important direction for future work.