# MT Shared Task 2025: Medical domain Machine Translation with Limited-Pretraining models

**Tran Hong Viet[1], Nguyen Tien Khoi[1], Tran Duy Long[1], Nguyen Minh Quy[1], Nguyen Van Vinh[1]**

[1]University of Engineering and Technology,
Vietnam National University, Hanoi, Vietnam

thviet@vnu.edu.vn,nguyentienkhoi210@gmail.com,longtd.179@gmail.com,,minhquy1624@gmail.com,vinhnv@vnu.edu.vn

## Abstract

In this paper, we present the summary of the MT Shared Task 2025 challenge on medical English–Vietnamese machine translation with small language models, as part of the Vietnamese Language and Speech Processing Workshop. In this shared task, 14 teams have participated in building and optimizing English-Vietnamese Machine Translation Systems using parallel medical corpora. Most of the teams used pre-trained models in a shared task for the competition with fine-tuning and data augmentation techniques to improve translation efficiency (prompt engineering with instruction templates; training to reinforce bilingual consistency; full-parameter fine-tuning of compact Qwen models with limited parameters). However, the difference in the high-achieving teams was that they developed effective fine-tuning techniques that combined data features in the medical domain, including small language models (SLMs), and appropriate pre-trained models. These carefully designed methods achieved high performance, outperforming the baseline methods. Despite the good results achieved, the inherent complexity of medical data presents significant opportunities for further development.

## 1 Introduction

Machine Translation (MT) is an essential tool, serving as a bridge to overcome language barriers, and promoting communication and the exchange of information. The role of machine translation is crucial, especially in the field of medicine, where the accuracy and timeliness of information can directly impact research, diagnosis, and treatment. The rapid translation of medical documents, research papers, and literature summaries helps healthcare professionals worldwide to access the latest knowledge while also improving the quality of patient care. However, translation of documents in the medical field is one of the most challenging problems for Machine Translation (MT) systems, especial for English-Vietnamese bilingual system. The main problems include strict accuracy requirements, domain-specific terminology, and scarce bilingual resources.

To address these challenges, the MT Shared Task 2025 has been established to foster advancements in machine translation research for the medical documents. The task release the medical parallel corpus for the medical domain which is collected from reliable sources including hospitals, medical centers. The corpus is then preprocessed to retain the most high quality pairs, result in a collection of 500,000 pairs of English and Vietnamese. The task also promotes the utilization of pretrained small language models (SLMs) to further investigate and advance their capabilities in machine translation.

We present a comprehensive overview of the Medical MT shared task, including the task definition, training dataset, developed systems, and the evaluation framework. In total, 14 teams participated, all constrained to using small language models (SLMs), though their training and inference strategies varied considerably, ranging from domain adaptation with medical corpora to different fine-tuning and decoding techniques. System performance was evaluated using SacreBLEU scores and human judgments, ensuring both quantitative and qualitative assessment. Beyond reporting results, this shared task highlights the importance of advancing bilingual English–Vietnamese medical MT research and its potential for developing accessible translation tools for all Vietnamese speakers, not only doctors or medical researchers.

## 2 Medical MT shared task

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:

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

where $\mathcal{V}_{\text{src}}$ is the source vocabulary, $T_x$ is the sentence length and $x_i$ is the token, 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 $\theta$ 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 $x$.

### 2.1 Datasets for the Share Task

The training corpus is collected mainly from MedEV (Vo et al., 2024), which is a high-quality English-Vietnamese medical dataset sourced from professional medical resources. We also collect ICD-10, the 10th edition of the International Statistical Classification of Diseases and Related Health Problems (ICD), a list of medical classifications published by the World Health Organization (WHO). Next, we filter the samples using sample length filtering and deduplication. The process results in 500,000 training samples and 3000 validation samples and test dataset with 1000 samples each language.

| 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 |

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

### 2.2 Evaluating machine translation systems

We follow (Post, 2018) to evaluate the translation results using SacreBLEU. The participating teams are allowed to use the validation dataset for model optimization. The final submissions are further judged by human doctors. Each translation is assigned a score based on the following scale: SacreBLEU is computed as a brevity-penalized geometric mean of $n$-gram precisions. The formula is given by:

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

Table 2: Human evaluation score scale for translations.

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

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:

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

with $c$ the candidate length and $r$ the reference length.

### 2.3 SLM constrain

The competition encourages teams to use SLMs from Qwen (Yang et al., 2025; Qwen et al., 2025) family to for the task. The teams are allowed to use several training and decoding strategies (Han et al., 2024) to improve models performance on medical translation task.

## 3 System Submissions

The competition was hosted on Codabench (Xu et al., 2021), 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.

### 3.1 Baseline model

We measure the SacreBleu of Qwen3-0.6B (Yang et al., 2025) before any training on our corpus. The model only rely on reasoning ability and chain of thought few-shot prompting. The evaluation result on both validation and test is provided to teams through the platform.

| Direction | Public Test | Private Test |
|---|---|---|
| EN → VI | 0.1874 | 0.2300 |
| VI → EN | 0.1830 | 0.1600 |

Table 3: Translation performance (SacreBleu scores) on public and private test sets.

### 3.2 Approaches adopted by participants

The task has attracted 15 teams In this section, we present an overview of the approaches taken by the teams that submitted papers describing their methods.

**Team Bosch@AI** applied a bidirectional training strategy to fine-tune small language models (Zoph et al., 2016). The motivation behind this approach comes from how humans learn foreign languages: by using translation examples in both directions (e.g., English to Vietnamese and Vietnamese to English) to better master bilingual knowledge.

They employed both full-parameter fine-tuning and Low-Rank Adaptation (Hu et al., 2021) (LoRA) to adapt the Qwen3-1.7B model (Yang et al., 2025) for bilingual machine translation tasks, where supervised fine-tuning minimizes the negative log-likelihood:

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

and LoRA updates a frozen weight $W_0$ by injecting a low-rank decomposition:

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

The training data consisted of parallel corpora in English–Vietnamese (en–vi) and Vietnamese–English (vi–en). The team also provided their evaluation of GPT-4o and GPT-4.1 on the validation dataset. Their best model was trained using full-parameter fine-tuning. After training, the model improved significantly in both directions and surpassed the GPT-4.1 baseline (7.13 and 8.21 BLEU on EN–VI and VI–EN, respectively).

**BenignRhythms** also using SFT with QLoRA (Dettmers et al., 2023) to finetuning qwen3-1.7B (Yang et al., 2025) model on both translation directions. Their innovation came from combining finetuned model with RAG (retrieval augmented generation) (Lewis et al., 2021) when inference.

**Algorithm 1** **Glossary Term Retrieval**

```
Input: Source sentence $x$, dictionary $D$, top-$k$,
       threshold $\tau$
Output: Glossary $G(x)$
    Step 1: Encode source.
      $u \leftarrow f(x)$ ;    // SBERT embedding of $x$
    Step 2: Compute similarities.
      $s \leftarrow uE_D^\top$ ;    // Dot product with
      dictionary embeddings
    Step 3: Select top-$k$ terms.
      $T' \leftarrow \text{TopK}(s, k)$
    Step 4: Filter terms.
      $T(x) \leftarrow \{t \in T' \mid s_t \geq \tau\}$
    Step 5: Build glossary.
      $G(x) \leftarrow \{(t, D(t)) \mid t \in T(x)\}$
return $G(x)$
```

They deploy MiniLM-L6-v2 (Reimers and Gurevych, 2019) for English translation and vietnamese-sbert (vie) to retrieve similar documents that satisfied cosine similarity of their embedding and the input is bigger than a predefined threshold.

**Team ZERO** implemented a robust deduplication pipeline based on the MinHash (Broder, 1997) LSH (Indyk and Motwani, 1998) algorithm. After their deduplication, the training and testing corpus dropped by 30.8%, resulting in 346,000 samples. The new corpus was then split into a 331K-pair set for SFT, a 15K-pair set for second training phase, and the remaining 3K pairs were used for model evaluation. The team's training approach has two phases using Qwen2.5-3B (Qwen et al., 2025) as base model. The first phase is similar to Bosch@AI team's strategy of applying supervised fine-tuning on bidirectional pairs. The second phase applies GRPO, a widely adopted reinforcement learning method (Shao et al., 2024), to further boost performance.

In GRPO (Shao et al., 2024), the policy parameters $\theta$ are updated to maximize expected rewards where $R(x, y)$ is a weighted reward based on BLEU and ChrF++ (Popović, 2017). ChrF++ measures character $n$-gram precision/recall with an $F$-score:

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

where $P$ and $R$ denote character $n$-gram precision and recall, and $\beta$ is a weighting factor.

In this phase, they trained each translation direction separately. The final model is the SFT model from phase 1 equipped with two LoRA adapters (one for EN–VI and one for VI–EN), both optimized using GRPO. These innovations led the team to achieve the highest evaluation score on the private test.

The team **JustGraduate** also did data preprocessing, SFT training as the first phase and then GRPO similar to ZERO team. The key difference is the team apply length penalty with BLEU score as reward signal. This design avoids outputs that are too long or short and can guide model toward greater fluency, structural fidelity, and robustness in medical translation.

They also describe how they curated the data for GRPO training. They 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, they obtained a balanced subset of approximately 50,000 sentence pairs (25,000 Vi–En and 25,000 En–Vi). The GRPO phase showed a notable improvement from the SFT phase, which is 2.46 points on En-Vi and 4.17 points on Vi-En.

### 3.3 Results of the team

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 |

On all of the evaluation methods, all of the models appeared to have better translation results on EN-VI direction than the quality of VI-EN. We also deployed GPT-4o as a judge, the result showed that GPT-4o has the same translation preference with human on EN-VI direction but it scored translations on VI-EN lower than human did. While team ZERO leading the SacreBleu leaderboard, the human evaluation prefer translation result of team Bosch@AI. However, the gap between these team on human leaderboard is not significant. The same positions changes happened with team BenignRhythms and team JustGraduate but there is big different with 2 previous team, the gap between these tem grow bigger on human preference leaderboard. JustGraduate achieved only 11.2 SacreBleu score but this is the only team that has vi-en translation quality better than en-vi quality. The final ranking is computed using the formula:

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 6: GPT-4o Leaderboard

| Team | AVG | EN-VI | VI-EN |
|---|---|---|---|
| **Bosch@AI** | 83.53 | 88.41 | 78.66 |
| **ZERO** | 80.02 | 88.87 | 71.17 |
| **BenignRhythms** | 74.40 | 84.19 | 64.60 |
| **JustGraduate** | 70.21 | 68.91 | 71.52 |

$$\text{FinalScore} = 0.5 \cdot \text{SacreBLEU} + 0.5 \cdot \text{Human}$$

where SacreBLEU and Human is the average score on both translation directions.

Table 7: Final Ranking

| Rank | Team | Final Score |
|---|---|---|
| **1st** | **ZERO** | 63.3 |
| 2nd | **Bosch@AI** | 61.7 |
| 3rd | **JustGraduate** | 54.5 |
| 4th | **BenignRhythms** | 54.0 |

## 4 Further Analysis

According to our evaluation, team Zero achieved the highest performance. Most of the teams outperform baseline. While the SacreBleu results of teams are low, the human still judged the translations as good, which showed the variety of translations and the potential to apply models to real world application. However, none of the team achieves perfect score, this show that only training on parallel corpus makes model can improved the translation, yet there is still a gap between human preference and models translated output.

All of the team use SFT training combined with customized prompts. It also showed each team has an unique way to engineer the prompt for training, result in the variety of the competition result. Due to each team's resources limitation varies, it is hard to conclude which team's method is the best, however it is showed that full finetuned model helps Bosch@AI achieved high score compare to other teams.

During the inference, the BenignRhythms team showed that enhancing the model's context with reference translations may not help model achieve high result on SacreBleu but it can achieve comparable result to other methods on human preference.

The Zero team and JustGraduate have shown GRPO can boost the translation quality. However, the improvement is minor on the result of team ZERO, which doubt the effectiveness of the second phase.On the other hand, JustGraduate have shown that the model improved after the GRPO phase but their performance on private test dropped and the human evaluation also show that the outputs are not preferred over other teams. While GRPO is a method rely on verifiable reward and the team also formulated a novel reward function for training, the Machine Translation task in general and even in medical domain rely heavily human preferences, which optimize model with n-grams signals only without human preferences signal can be discussed more.

## 5 Conclusion and Future work

The Medical Machine Translation Shared Task highlights the potential of small language models (SLMs), which leverage both pretrained knowledge and instruction-following capabilities. The results suggest that translation quality can be further optimized not only through continued training but also by applying advanced reinforcement learning techniques and carefully engineered prompts.

Despite recent progress, a noticeable performance gap remains between the two translation directions. This indicates that, although medical machine translation has significantly benefited from advances in natural language processing and large language models, it still presents open challenges. Future research should continue to explore methods for improving translation accuracy, expanding domain coverage, and addressing context sensitivity in medical texts.

## 6 Limitations

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

First, although the translations are of high quality, the dataset primarily represents sentence-level equivalences and lacks broader contextual information. In practice, the same source text may yield different valid translations depending on the intended audience or situational context, which the corpus does not capture.

Second, the coverage of the corpus is restricted to medical documents. Although this ensures domain specificity, it narrows the diversity of text types. Important sources such as clinical dialogues, patient–doctor communications, electronic health records, and instructional materials are not represented, which may limit the applicability of models trained solely on this dataset.

## References

Vietnamese-sbert-v2. https://huggingface.co/thang1943/vietnamese-sbert-v2. Accessed: 2025-09-17.

Andrei Z. Broder. 1997. On the resemblance and containment of documents. Min-wise hashing / MinHash idea (original work for near-duplicate detection).

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

Lifeng Han, Serge Gladkoff, Gleb Erofeev, Irina Sorokina, Betty Galiano, and Goran Nenadic. 2024. Neural machine translation of clinical text: An empirical investigation into multilingual pre-trained language models and transfer-learning. *Preprint*, arXiv:2312.07250.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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)*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation (WMT)*, pages 612–618.

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.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

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.

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

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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 (EMNLP)*, pages 1568–1575.