Report 2

1. Introduction

In this report, following the previous evaluation of the two state-of-the-art Vietnamese–English translation models, VinAI Translate and EnViT5 Translation, we extend our analysis to include additional evaluation perspectives. Specifically, we continue to assess the models using the BLEU score, but this time with respect to the content categories of the sentences.

In the original MedEV paper, four different content categories were used for evaluation. However, these category labels were not included in the provided dataset. To reproduce a similar evaluation process, we employed a **one-shot classification model** to automatically classify the test set into the same four categories as defined in the paper. BLEU scores were then computed separately for each of these subsets.

To provide a more comprehensive and precise assessment, we further introduced two additional evaluation metrics — Translation Edit Rate **(TER)** and Metric for Evaluation of Translation with Explicit Ordering **(METEOR)** — alongside BLEU.

Finally, this report also discusses the challenges we encountered while preparing for the fine-tuning stage of the two models and outlines our proposed strategies to address these computational limitations in the next phase of the project.

2. Evaluation of evit5 translation model and VinAI Translate translation model with BLEU score with respect to context

Evaluation requirements

In the previous report, we evaluated the two models using BLEU scores with respect to sentence length. In the original MedEV paper, the authors further assess model performance based on sentence content, using the same BLEU metric. Specifically, the test set is divided into four content-based categories: **Article Abstracts**, **MSD Manuals**, **Thesis Summaries**, and **Article Translations**.

To replicate this experiment, it was necessary to partition our test set into the same four genre-based subsets. However, the publicly released MedEV test set does not include genre labels. To address this limitation, we employed an open-source multilingual transformer model, *joeddav/xlm-roberta-large-xnli*, to perform zero-shot classification of the English sentences into the four predefined genres. Each predicted genre label was then assigned to the corresponding Vietnamese sentence in the parallel pair. Finally, the classified sentence pairs were organized into four separate files corresponding to each genre, enabling genre-specific BLEU score evaluation consistent with the methodology of the original paper.

.

Evaluation results

The experiment results were shown below:

		Article Abstracts	MSD Manuals	Thesis Summaries	Article Translation
EnViT5 Translation (w/o FT)	Origin paper result	28.07	39.79	35.82	39.34
	Evaluation result on 100 random samples	**13.12**	**28.18**	**26.52**	**18.25**
VinAI Translate (w/o FT)	Origin paper result	24.51	42.86	33.65	38.13
	Evaluation result on 100 random samples	**38.63**	**41.63**	**39.85**	**34.04**

Table 1: BLEU scores for Vietnamese- English translation performance evaluation

		Article Abstracts	MSD Manuals	Thesis Summaries	Article Translation
EnViT5 Translation (w/o FT)	Origin paper result	37.99	53.46	37.03	48.74
	Evaluation result on 100 random samples	**36.89**	**41.42**	**32.69**	**43.35**
VinAI Translate (w/o FT)	Origin paper result	37.44	50.85	41.04	46.89
	Evaluation result on 100 random samples	**38.54**	**46.54**	**34.14**	**44.66**

Table 2: BLEU scores for English - Vietnamese translation performance evaluation

Result summaries

From the above results, several observations can be made regarding the Vietnamese–English and English–Vietnamese translation performance of the two models on medical texts:

- VinAI Translate consistently outperforms EnViT5 Translation across all four content categories, with the performance gap being particularly substantial in the Vietnamese–English translation direction.
- Although the results obtained from the 100 randomly sampled sentences exhibit the same performance pattern reported in the original paper, the absolute scores differ noticeably. This discrepancy is likely due to two factors: (i) the significantly smaller sample size used in our evaluation, and (ii) differences in genre classification, since our zero-shot classification approach may not perfectly replicate the genre-labelling method used by the authors.
- Both VinAI Translate and EnViT5 Translation achieve higher BLEU scores for English-to-Vietnamese translation than for Vietnamese-to-English translation.

- The results further show that translation performance on MSD Manuals and Article Translations is slightly higher than on Article Abstracts and Thesis Summaries. This may be attributed to the more standardized terminology and writing style found in MSD Manuals and professionally translated articles, which are more consistent and thus easier for the models to learn and translate.

3. Evaluation of EnViT5 Translation model and VinAI Translate translation model with METEROR and TER

In addition to BLEU, which primarily measures n-gram overlap between the system output and the reference translation, the origin paper also employed METEOR and **TER** to provide a more comprehensive evaluation of translation quality.

What is METEOR?

The **METEOR (Metric for Evaluation of Translation with Explicit Ordering)** score considers not only exact word matches but also synonyms, stemming, and word order, allowing it to better capture the semantic similarity between sentences. Higher METEOR scores indicate translations that are closer in meaning to the reference.

What is TER score?

On the other hand, the **TER (Translation Edit Rate)** measures the number of edits—insertions, deletions, substitutions, and shifts—required to transform a system translation into the reference. A lower TER score signifies that fewer corrections are needed, reflecting higher translation accuracy.

EnViT5 Translation translation model evaluation with METEOR

In our experiment, due to computation constrains, we could only set up the experiment to evaluate the METEOR and TER scores on the random 100 samples from the test set.

		METEOR	TER
EnViT5 Translation (w/o FT)	Origin paper result	0.627	61.93
	Evaluation result on 100 random samples	**0.6215**	**89.17**
VinAI Translate (w/o FT)	Origin paper result	0.626	67.63
	Evaluation result on 100 random samples	**0.6581**	**55.42**

Table 3: Comparation of METEOR and TER result on two models in Vienamese – English translation

		METEOR	TER
EnViT5 Translation (w/o FT)	Origin paper result	0.627	61.93
	Evaluation result on 100 random samples	**0.6797**	**52.76**
	Origin paper result	0.626	67.63

VinAI Translate (w/o FT)	Evaluation result on 100 random samples	**0.6879**	**45.54**

Table 4: Comparation of METEOR and TER result on two models in English - Vietnamese translation

Result summary

From the above evaluation results, we can make the following conclusion about the two models.

- The results indicate that both models achieve similar METEOR scores in both translation directions (Vietnamese–English and English–Vietnamese), suggesting comparable performance in terms of semantic similarity and alignment with reference translations.
- VinAI Translate demonstrates strong performance in TER, achieving lower edit distances in both translation directions. In contrast, EnViT5 Translation exhibits notably weaker TER performance, particularly in the Vietnamese–English direction, indicating that more edits are required to transform its outputs into the reference translations.
- A substantial discrepancy exists between our TER results and those reported in the original paper. This difference is likely attributable to the small sample size used in our experiment, which is considerably smaller than the full test set of 8,960 sentences. With limited samples, TER becomes more sensitive to sentence-level variability and may not reflect the true distribution of errors present in the larger dataset.

4. Fine tuning EnViT5 Translation and VinAI Translate translation with MedEV training set

After throughout analysis two state-of-the-art Vietnamese–English translation models with three different measurement metrics, in the next step, we will evaluate their performance after fine-tuned with MedEV training data set.

Fine-tuning requirements

VinAI Translate-translate, EnViT5 Translation-translation were fine-tuned on the MedEV training set using the HuggingFace *Transformers* library. Training was conducted for 5 epochs with the AdamW optimizer, which provides stable weight updates and reduces overfitting. An initial learning rate of 5×10^{-5} and a maximum sequence length of 256 tokens were used to balance training stability and computational cost.

To increase efficiency, the authors employed mixed-precision (fp16) training across 4 NVIDIA A100 GPUs. Each GPU used a batch size of 4, and 8 steps of gradient accumulation were applied to achieve a larger effective batch size without exceeding

GPU memory limits. Additionally, 1250 warm-up steps were used to gradually increase the learning rate at the start of training.

For translation generation, beam search with a beam size of 5 was used to improve output quality. Model performance was evaluated every 1000 training steps on the validation set using the standard metrics BLEU, TER, and METEOR. BLEU was computed using the reproducible SacreBLEU implementation with case sensitivity. The best checkpoint, based on the highest validation BLEU score, was selected for final evaluation on the test set.

Difficulties in doing the same fine-tuning process with current compute power

The original study used four NVIDIA A100 GPUs, which provide extremely high computational performance and large GPU memory, enabling efficient training of large transformer models such as VinAI Translate-translate and EnViT5 Translation. Multi-GPU training significantly accelerates training, supports larger batch sizes, and allows the models to fit into memory.

In contrast, our commercial computer typically does not have sufficient GPU memory or processing power to fine-tune these large models in a reasonable time. Training on a home GPU would either be extremely slow or may not run at all due to memory limitations.

Therefore, when equivalent hardware is not available, alternative approaches such as using cloud GPUs (e.g., A100 instances) or smaller translation models should be considered.