Report 3

Introduction

This report describes the process of fine-tuning 2 models (envit5 and vinAI) using the training set and evaluate set from the MedEV dataset. Due to compute power constrains, we could only do a fraction of the fine-tuning. The different between the full tuning and the fraction tuning will be given first. After the fine-tuning, we reevaluate both 2 models using the same metrics as the previous report (BLEU, TER and METEOR score). The result shows that even with a fraction of the fine-tuning, the model performance get better at medical translation in both direction (Vietnamese to English and the other way around). The evaluation result using the test set from MedEV will also be given in the report. So from this experiment, we could expect that if we can make full use of the MedEV in fine-tuning, the English to Vietnamese medical translation performance of the LLM models will be improved.

Pre-training and Fine-tuning process in LLM

In large language models (LLMs) and neural machine translation (NMT) systems, training typically follows a two-stage paradigm: **pre-training** and **fine-tuning**. During pre-training, models are trained on large-scale general-domain text corpora to learn fundamental linguistic patterns, grammar, and cross-lingual representations. However, such general training is often insufficient for specialized domains such as medical translation, where terminology, style, and semantic precision differ significantly from everyday language.

To address this limitation, fine-tuning is performed using a domain-specific parallel corpus. In this process, a pre-trained sequence-to-sequence model is further optimized on sentence-aligned source–target pairs, allowing the model to adapt its parameters to domain-specific vocabulary, sentence structures, and translation conventions.

Use of the MedEV Dataset

In our experiment, the **MedEV dataset** is used as the domain-specific supervision signal for fine-tuning Vietnamese–English and English–Vietnamese translation models. MedEV is a high-quality medical parallel corpus consisting of approximately 360K sentence pairs collected from multiple medical resources, including journal abstracts, clinical manuals, and thesis summaries. The dataset is split into training, validation, and test sets at the document level to prevent information leakage.

During fine-tuning, the training portion of MedEV is used to update model parameters, while the validation set is employed for periodic evaluation and checkpoint selection based on translation quality metrics (e.g., BLEU). This process enables the models to

better capture medical terminology and domain-specific expressions, leading to substantial improvements in translation performance compared to non-fine-tuned baselines.

Fine-tuning using MedEV dataset

Due to computational constraints, we were unable to perform full-scale fine-tuning using the entire MedEV dataset as described in the original study. Instead, a lightweight fine-tuning configuration was adopted. Table 1 summarizes the key differences between our experimental setup and the original fine-tuning process reported in the paper.

Aspect	Original full fine-tuning	Light-weight fine-tuning (using Colab)
Model	Vinai-translate,envit5-translation	Vinai-translate,envit5-translation
Dataset (training)	Entire MedEV training split	First 40,000 sentences pairs from the training split
Validation data	Entire MedEV validation split	First 4000 sentecens pairs from the validation split
Max sequence lenght	256 tokens for source and target	128 tokens for source and target
Per-device batch size	4 examples /GPU	4 examples/GPU
Gradient accumulation	8 steps	8 steps
Effective global batch size	4GPU x 4 batch x 8 grad-accum = 128 examples/step	1GPU x 4 batch x 8 grad-accum = 32 examples/step
Warm-up step	1230	300
Learning rate	5e-5 (AdamW)	5e-5 (AdamW)
Precision	Mixed precision fp16	Mixed precision fp16
Hardware	4 x NVIDIA A100	1 x NVIDIA L4
Decoding for evaluation	Beam search, beam size = 5	Beam search, beam size = 3

Table 1: Comparison of Fine-Tuning Configurations Between the Original Study and the Lightweight Setup

As shown in Table 1, the proposed lightweight configuration is substantially smaller in scale in terms of training data size and computational resources. Consequently, the translation performance obtained after fine-tuning may differ from the results reported in the original work, and such differences should be interpreted in light of these experimental constraints.

Evaluation after fine-tuning

After fine-tuning, the translation performance of each model was evaluated using three standard machine translation metrics: **BLEU**, **METEOR**, and **TER**. For a fair and comprehensive assessment, we conducted the evaluation on the entire MedEV test

split, consisting of 8,960 sentence pairs. This evaluation setting is consistent across all models, allowing a direct comparison of their post–fine-tuning performance.

envit5-translation

BLEU score on Vietnamese to English direction

Model		Envit5- translation vi2en					
		<10	[10,20]	[20,30]	[30,40]	[40,50]	>50
W/o fine-tuning		27.07	28.31	31.76	33.89	35.53	37.12
With fine-tuning	Origin paper	38.07	39.97	41.24	41.80	45.59	47.12
	Lightweight training result	32.38	34.86	35.52	35.96	36.60	35.19

Table 2 BLEU scores comparison between model before and after being fine-tuned

TER and METEOR on Vietnamese to English direction

Model		Envit5 translation vi2en	
		TER	METEOR
W/o fine-tuning		86.67	0.6158
W fine-tuning	Origin paper	67.63	0.627
	Lightweight training result	57.6320	0.6492

Table 3 TER and METEOR score comparison between model before and after being fine-tuned

BLEU score on English to Vietnamese direction

Model		Envit5-translation en2vi					
		<10	[10,20]	[20,30]	[30,40]	[40,50]	>50
W/o fine-tuning		38.72	41.77	42.75	43.73	44.08	42.59
With fine-tuning	Origin paper	49.97	50.50	50.30	50.81	51.27	51.99
	Lightweight training result	38.27	42.20	44.50	45.40	45.22	38.14

Table 4 BLEU scores comparison between model before and after being fine-tuned

TER and METEOR score on English to Vietnamese direction

Model		Envit5 translation en2vi	
		TER	METEOR

W/o fine-tuning		54.375	0.685
	Origin paper	42.23	0.733
W fine-tuning	Lightweight training	50.483	0.6858

Table 5 TER and METEOR score comparison between model before and after being fine-tuned

vinai-translate-vi2en

BLEU score on Vietnamese to English direction

Model		VinAI- translation vi2en					
		<10	[10,20]	[20,30]	[30,40]	[40,50]	>50
W/o fine-tuning		25.91	28.23	30.85	31.54	32.81	30.39
W fine-tuning	Origin paper	38.07	39.97	41.24	41.80	44.59	47.12
	Lightweight training result	32.85	32.56	34.04	34.42	34.90	32.41

Table 6 BLEU scores comparison between model before and after being fine-tuned

TER and METEOR score on Vietnamese to English direction

Model		VinAI translation vi2en	
		TER	METEOR
W/o fine-tuning		66.200	0.685
W fine-tuning	Origin paper	42.22	0.740
	Lightweight training result	62.023	0.6260

Table 7 TER and METEOR score comparison between model before and after being fine-tuned

Conclusion

Based on the experimental results, several observations can be made. First, after fine-tuning on the MedEV dataset, the models generally exhibit improvements across all three evaluation metrics—BLEU, TER, and METEOR—when compared to their pre–fine-tuning performance. Nevertheless, a noticeable performance gap remains between the models trained using the proposed lightweight fine-tuning setup and those reported in the original study, which employed full-scale fine-tuning.

Furthermore, while fine-tuning leads to overall performance gains, the improvements are not uniform across all sentence length categories. In particular, the BLEU score decreases for longer sentences (≥ 50 words) in certain translation directions. For example, in the EnViT5 English-to-Vietnamese setting, the BLEU score for long sentences drops from 42.59 to 38.14 after fine-tuning. A similar trend is observed in the VinAI translation Vietnamese-to-English direction, where the BLEU score decreases from 37.12 to 35.19. This suggests that the lightweight fine-tuning process may be insufficient for capturing long-range dependencies present in longer medical texts.

One plausible explanation for this behaviour is the limited scale of the fine-tuning process, as only a subset of the MedEV dataset was used due to computational constraints. Additionally, during training, model checkpoint selection was based solely on the overall BLEU score, without explicitly considering performance variations across sentence lengths or alternative metrics.

A similar pattern is also observed in the METEOR evaluation results, where performance degradation occurs after fine-tuning in some cases. This may be attributed to the fact that METEOR was not included as an optimization or checkpoint-selection criterion during the fine-tuning process. These findings highlight the importance of aligning training objectives and evaluation metrics, particularly when operating under constrained training settings.