Advanced Language Model-based Translator for English-Vietnamese Translation

Hoai Nam Nguyen
hoainam1001.nhn@gmail.com

Thanh Trong Nguyen
trongnt2002@gmail.com

Quoc Bao Nguyen
baonq@thinkprompt.com

Vu Anh Tran
tranvuanh.cs@gmail.com

Doctranslate.io, ThinkPrompt CO., LTD

Abstract

We introduce a transformative approach to English-Vietnamese translation, leveraging the cutting-edge capabilities of the **Gemma-7B-IT** (Gemma Team et al. 2024) model. Enhanced by the Advanced Language Model-based Translator (ALMA) (Xu et al. 2023) methodology, our system significantly advances beyond the conventional Transformer models in handling complex linguistic contexts. This research details our robust training framework, experimental validations, and the rigorous evaluation processes that establish a new state-of-the-art for Vietnamese translation tasks.

Our results emphatically surpass those of well-known systems such as VinAI Translate (Nguyen et al. 2022) and Google Translate (Google 2024b), demonstrating an improvement of over 12 BLEU scores against the previously top-performing systems. These achievements highlight the superior flexibility and contextual understanding capabilities of Large Language Models (LLMs) (Zhao et al. 2023) integrated within our ALMA framework, which excel in adapting to varied translation nuances and complexities.

Capitalizing on these remarkable advancements, we have also introduced a user-centric translation product, available at *https://www.doctranslate.io* (Doctranslate 2023) [1] [2]. This tool embodies our commitment to merging technological innovation with practical utility, offering users a seamless and high-quality translation experience.

1 Introduction

Vietnam has dramatically transformed, becoming a crucial participant in global trade and investment. This rapid economic evolution has spurred increased demand for adept Vietnamese-English translation services to support text and speech communications. Foreign investors and business partners predominantly utilize automatic translation systems to overcome linguistic barriers and maintain current with regional developments.

Amidst the advancements in Natural Language Processing (NLP), Large Language Models (LLMs) like GPT-3.5 and GPT-4 OpenAI 2023 have shown remarkable competencies across a range of NLP tasks. These models stand as unique challengers to traditional supervised encoder-decoder models in translation tasks. To further refine the quality of English-Vietnamese (EN-VI) and Vietnamese-English (VI-EN) translations, we have developed an innovative system integrating the Gemma model with the Advanced Language Model-based Translator (ALMA) method. This system exploits the robust contextual comprehension and generative response capabilities of LLMs.

One of the distinctive strengths of LLMs is their exceptional adaptability in response generation, which enables them to produce contextually accurate translations that align closely with the original speaker's intent. Importantly, LLMs offer extensive customization in prompts, allowing users to specify instructions or desired response styles—be it scientific, literary, or other formats. This flexibility significantly enhances the human-like quality of the translations and aids in providing specific outputs tailored to diverse professional needs. Additionally, the inherent ability of LLMs to generalize effectively across languages promotes their utility in multilingual translation tasks, setting new standards in the field and demonstrating why LLMs are an optimal choice for our docstranslate initiative.

[1] Ours code, model, and dataset are available at *https://github.com/doctranslate-io/viet-translation-llm*
[2] Ours website: *https://www.doctranslate.io*

2 Preliminary

2.1 Task Definition

This research is dedicated to enhancing the quality of translations between English and Vietnamese by employing state-of-the-art machine translation models. Our primary objective is to deliver translations that are not only highly accurate but also exhibit natural fluency in both English-to-Vietnamese and Vietnamese-to-English directions.

The study further investigates the application of Large Language Models (LLMs) augmented with advanced zero-shot learning techniques. Zero-shot learning enables these models to perform translation tasks without having been directly trained on specific instances of the English-Vietnamese language pair. This innovative approach allows the LLMs to apply and adapt knowledge gleaned from various languages and contexts to the task of translating between English and Vietnamese.

Translation is inherently a conditional and context-dependent problem (Buden et al. 2009). Traditional models such as encoder-decoder architectures or sequence-to-sequence (seq2seq) models often struggle to fully capture the complexities and subtleties required for high-quality translation. These models typically require extensive training data specific to the language pair and can falter when dealing with nuanced or contextually rich texts. In contrast, LLMs excel in contextual understanding and learning, which allows them to interpret and reflect the intended meaning and stylistic nuances of the source text more effectively. This capability makes LLMs particularly suited to meet the practical demands of translation, where understanding the context and the conditions under which the text is generated is crucial for maintaining accuracy and fluency.

```
Translate this from [source language] in to [target language]
[source language]: <source sentences>
[target language]:
```

Above are the prompts used for training and evaluation. *[source language]* and *[target language]* represents the full name of the language, e.g. Translate this language from English to Vietnamese.

2.2 Zero-shot Learning and Instruction Prompt in Large Language Models

In the realm of machine translation, zero-shot learning is a pivotal technique that empowers Large Language Models (LLMs) to tackle translation tasks without explicit prior training on the target language pair. This approach leverages the extensive pre-training on diverse datasets, enabling the models to generalize their learned linguistic patterns to new, unseen languages and contexts.

Zero-shot learning in LLMs is particularly effective when combined with precise instruction prompting. By formulating clear and context-specific prompts, researchers can guide the models to apply their generalized knowledge more accurately and efficiently. For example, prompts can be designed to specify not only the translation task but also the desired tone, formality level, and inclusion of cultural nuances, which are critical for producing translations that are not only linguistically correct but also contextually appropriate.

Instruction prompting harnesses the inherent flexibility of LLMs to adapt their output to meet specific requirements. This is crucial in translation tasks where the context, such as legal documents or literary works, demands a particular stylistic approach or technical accuracy. Through instruction prompts, LLMs can be directed to prioritize certain aspects of the translation, like maintaining the original sentiment or adhering to industry-specific terminology, enhancing both the relevance and quality of the translated content.

The instruction prompt to include custom dictionary and style specifications for enhanced translation with Large Language Models (LLMs):

```
Translate this from [source language] into [target language],
ensuring to use the provided custom dictionary for specified terms
and maintaining the poetic style as described.

Custom Dictionary: <List of specific terms and their translations>

Desired Style: <Specify style, e.g., formal, poetic, technical>
Maintain a poetic and elegant tone, enhancing the translation with
stylistic elements that resonate with the original's intellectual
depth.

[source language]: <source sentences>
[target language]:
```

This revised prompt structure enables the translation task to be clearly defined with precise instructions on using a custom dictionary for specific terms and maintaining a particular style throughout the translation. This approach ensures that the output not only remains linguistically accurate but also stylistically consistent with the desired tone and terminology.

Overall, the combination of zero-shot learning and tailored instruction prompting forms a robust framework within which LLMs can perform highly adaptive and context-sensitive translations. This methodology not only expands the utility of LLMs across various languages and domains but also sets a new standard in the precision and adaptability of automated translation systems.

2.3 The Advanced Language Model-based Translator

The Advanced Language Model-based Translator (ALMA) (Xu et al. 2023) method represents a significant stride in addressing the limitations of moderate-sized generative Large Language Models (LLMs) in translation tasks. While LLMs such as those with 7B or 13B parameters have shown remarkable capabilities in various NLP tasks, they often lag behind conventional supervised encoder-decoder models in translation accuracy. Traditional methods heavily rely on extensive parallel data to achieve high-quality translation. However, ALMA circumvents this need by employing a novel fine-tuning approach that leverages both monolingual and high-quality parallel data, thus minimizing the dependency on large-scale bilingual corpora.

This two-stage fine-tuning process begins with the initial adaptation on monolingual data to grasp the linguistic nuances of each language independently. It is then followed by a targeted fine-tuning on a select set of high-quality parallel data. This strategy not only enhances the translation capabilities of LLMs but also significantly boosts their performance

2.4 The backbone LLM

2.4.1 Mistral

Mistral 7B (Jiang et al. 2023), a decoder-only transformer model, stands out for its architectural choices aimed at revolutionizing language comprehension and generation. Key features include:

- Sliding Window Attention: With an 8k context length and a fixed cache size, Mistral 7B achieves a theoretical attention span of 128K tokens, attributed to a 4k sliding window size across its 32-layered architecture. This expansive attention mechanism allows for nuanced understanding and generation of lengthy text sequences by effectively managing computational resources.
- Grouped Query Attention (GQA): This innovation enables faster inference times and reduced cache sizes by grouping queries before computing attention, streamlining the process without compromising the model's depth of understanding.
- Byte-fallback BPE tokenizer: By ensuring that characters are never mapped to out-of-vocabulary (OOV) tokens, this tokenizer eliminates the occurrence of unknown tokens, thereby preserving the integrity of the input data.

Mistral-7B-Instruct-v0.1, the instruction-tuned model optimized for chat and interactive purposes through Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO). SFT enhances model

response to specific instructions, while DPO adjusts model preferences based on outcome desirability. An enhanced version, Mistral-7B-Instruct-v0.2, builds on the instruction-tuned model's capabilities, offering refined interaction quality and comprehension.

2.4.2 Gemma

Trained on up to 6 trillion tokens, Gemma (Gemma Team et al. 2024) models leverage architecture and training strategies from the Gemini family (Gemini Team et al. 2023) to achieve state-of-the-art capabilities. They have been rigorously evaluated across automated and human benchmarks in fields such as question answering, commonsense reasoning, and technical domains like mathematics and coding.

The Gemma models represent a significant advancement in open language models, designed to enhance language understanding, reasoning, and safety across diverse applications.

2.5 PhoMT Dataset

PhoMT (Doan et al. 2021) is a high-quality and large-scale Vietnamese-English parallel dataset introduced to improve machine translation tasks. It contains 3.02 million sentence pairs, making it significantly larger than the previous benchmark Vietnamese-English machine translation corpus, IWSLT1

The dataset consists of a total of 3.02 million sentence pairs covering a diverse range of domains, including News, Blogspot, TED-Talks, MediaWiki, WikiHow, and OpenSub. The dataset is divided into three subsets: the training set, which contains 2.98 million pairs; the validation set, which includes 18,719 pairs; and the test set, which comprises 19,151 pairs.

Domain	Total Pairs	Training Pairs	Validation Pairs	Test Pairs
News	41,504	40,990	257	257
Blogspot	93,956	92,545	597	814
TED-Talks	320,802	316,808	1,994	2,000
MediaWiki	496,799	490,505	3,024	3,270
WikiHow	513,837	507,379	3,212	3,246
OpenSub	1,548,971	1,529,772	9,635	9,564
Total	**3,015,869**	**2,977,999**	**18,719**	**19,151**

Table 1: Domain Breakdown of PhoMT Dataset

The chart 1 shows the average number of characters per sentence for English (en/s) and Vietnamese (vi/s) across the Training, Validation, and Test sets.

3 Experimentals

3.1 Training Setup

Fine-tuning is essential for the enhancement of models. By integrating additional translation data, it improves the model's ability to handle diverse linguistic structures and contextual nuances, leading to greater accuracy. Fine-tuning also ensures the model adheres to specific instructions more precisely, thereby reducing errors and misunderstandings. In our approach, we utilize a many-to-many multilingual translation framework with *gemma-7b-it* as our backbone model due to its outstanding zero-shot performance.

Gemma was chosen over Mistral as our base model because of its superior performance. While Mistral-7B-v1.0 achieved respectable zero-shot scores (BLEU: 23.51), Gemma-7B-IT significantly outperformed it (BLEU: 34.41). This clear advantage in performance made Gemma the preferred choice for our training setup.

Our fine-tuning process, conducted on the PhoMT training data, involves two stages, resulting in two distinct types of models.

1. **ALMA-Gemma-7B-IT** [3]: LoRA fine-tuning on PhoMT training data for *gemma-7b-it* model.
2. **ALMA-Gemma-7B-IT-ST**: Full-weight fine-tuning on PhoMT training data, then continue LoRA fine-tuning on high-quality parallel synthetic data.

[3]Model are available at: *https://github.com/doctranslate-io/viet-translate-llm*

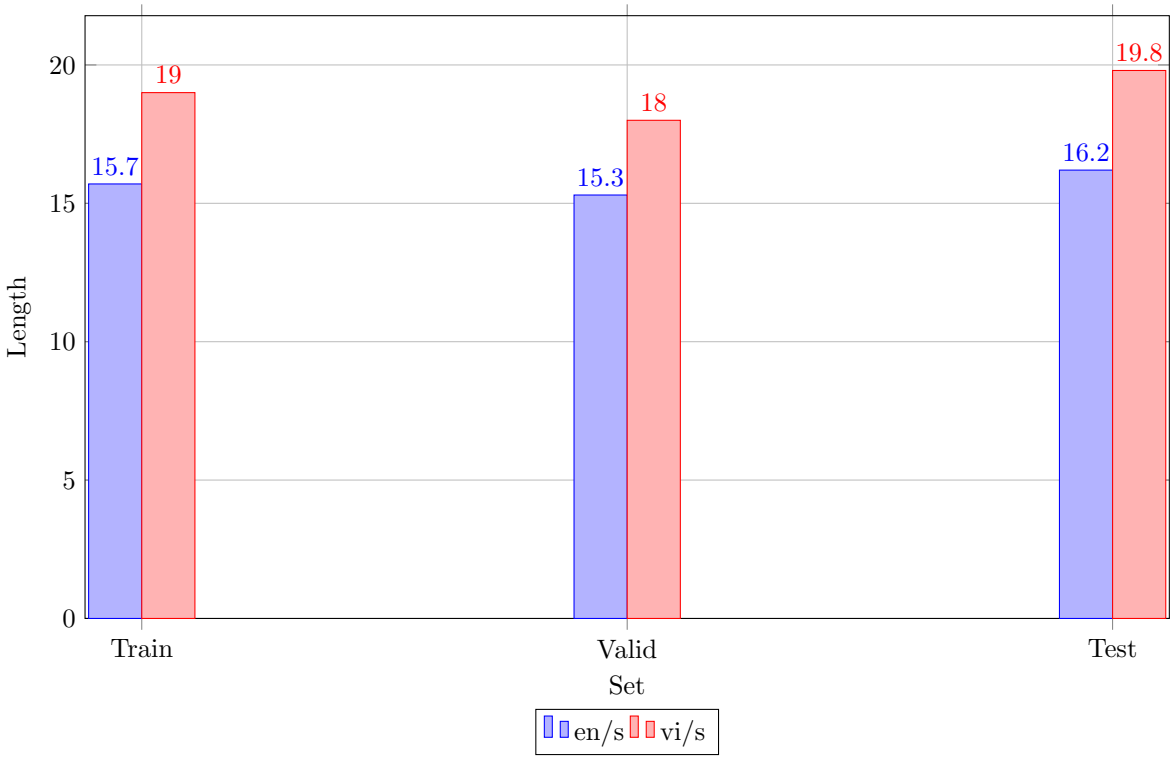


Figure 1: Average Length per Sentence on PhoMT Dataset

In this study, we utilize Low-Rank Adaptation (LoRA) with a rank of 16, updating only 0.1% of the parameters. Specifically, the gemma-7b-it model is fine-tuned using a batch size of 256, a warm-up ratio of 0.01, and sequences capped at 512 tokens.

For monolingual data fine-tuning, we subject gemma-7b-it to training on the PhoMT dataset across two epochs, a duration found sufficient for achieving clear model convergence. We select the iteration exhibiting the lowest validation loss as the optimal model configuration.

Furthermore, the ALMA-Gemma-7B-IT-ST model undergoes additional training not only on the PhoMT dataset but also on internally labeled data. This supplementary training is aimed at enhancing its capability to comprehend instructions and improving the quality of translations. The ALMA-Gemma-7B-IT-ST model has been commercialized and is available at the *Doctranslate.io* website.

3.2 Results

Models	**en-vi**	**vi-en**
Mistral-7B-v1.0, zero-shot	23.51	22.35
Gemma-7B-IT, zero-shot	34.41	32.78
Google Translate	39.86	35.76
Madlad400 7B	40.77	39.86
VinAI Translate	44.29	40.42
GPT-3.5-Turbo, zero-shot	39.3	34.1
Gemini-1.0-Pro, zero-shot	41.23	39.09
ALMA-Gemma-7B-IT (ours)	**52.70**	**50.80**
ALMA-Gemma-7B-IT-ST (ours)	**56.21**	**57.32**

Table 2: The overall results with BLEU Score on PhoMT Testset

The results clearly show that our models, particularly the ALMA-Gemma-7B-IT-ST, deliver outstanding performance, setting a new benchmark in the English to Vietnamese (en-vi) translation with a BLEU score of 56.21, the highest in our evaluations. This score significantly surpasses those achieved by

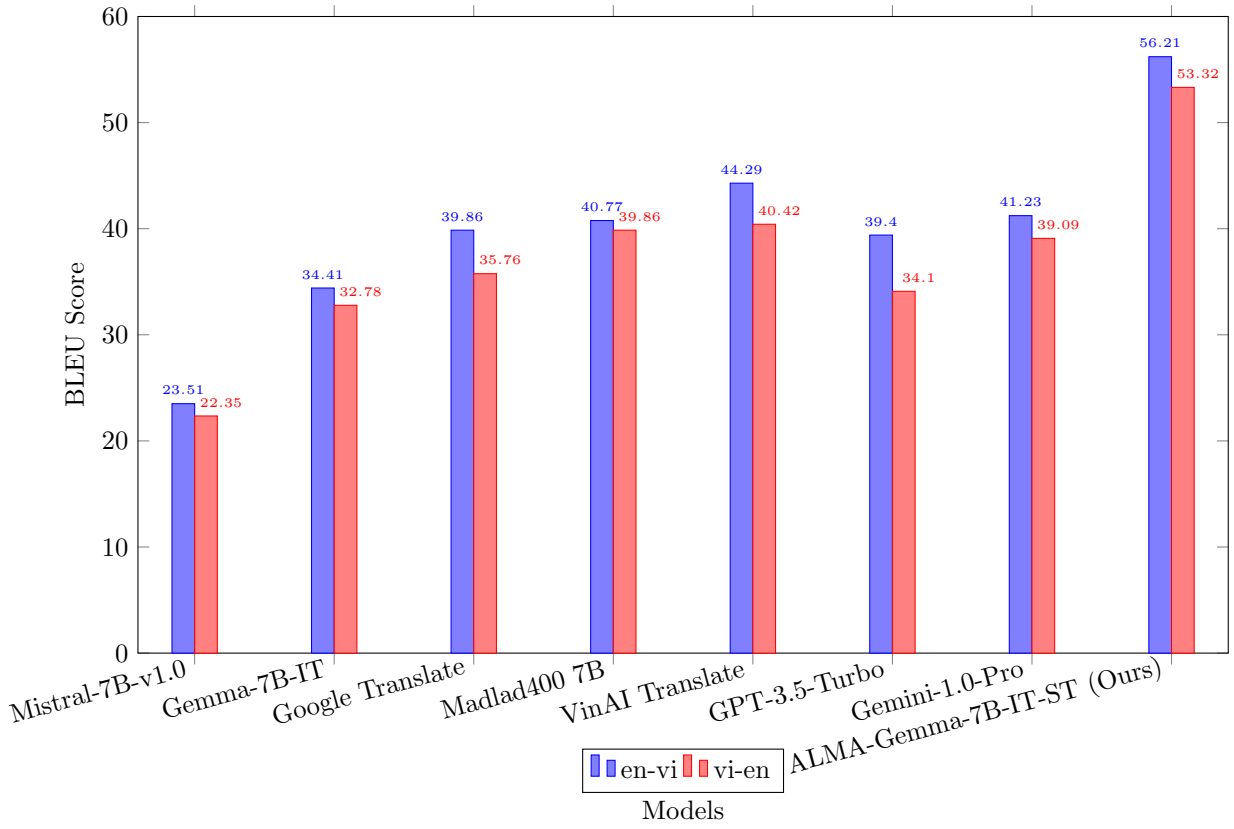


Figure 2: Comparison of BLEU Scores for VI-EN and EN-VI Translations

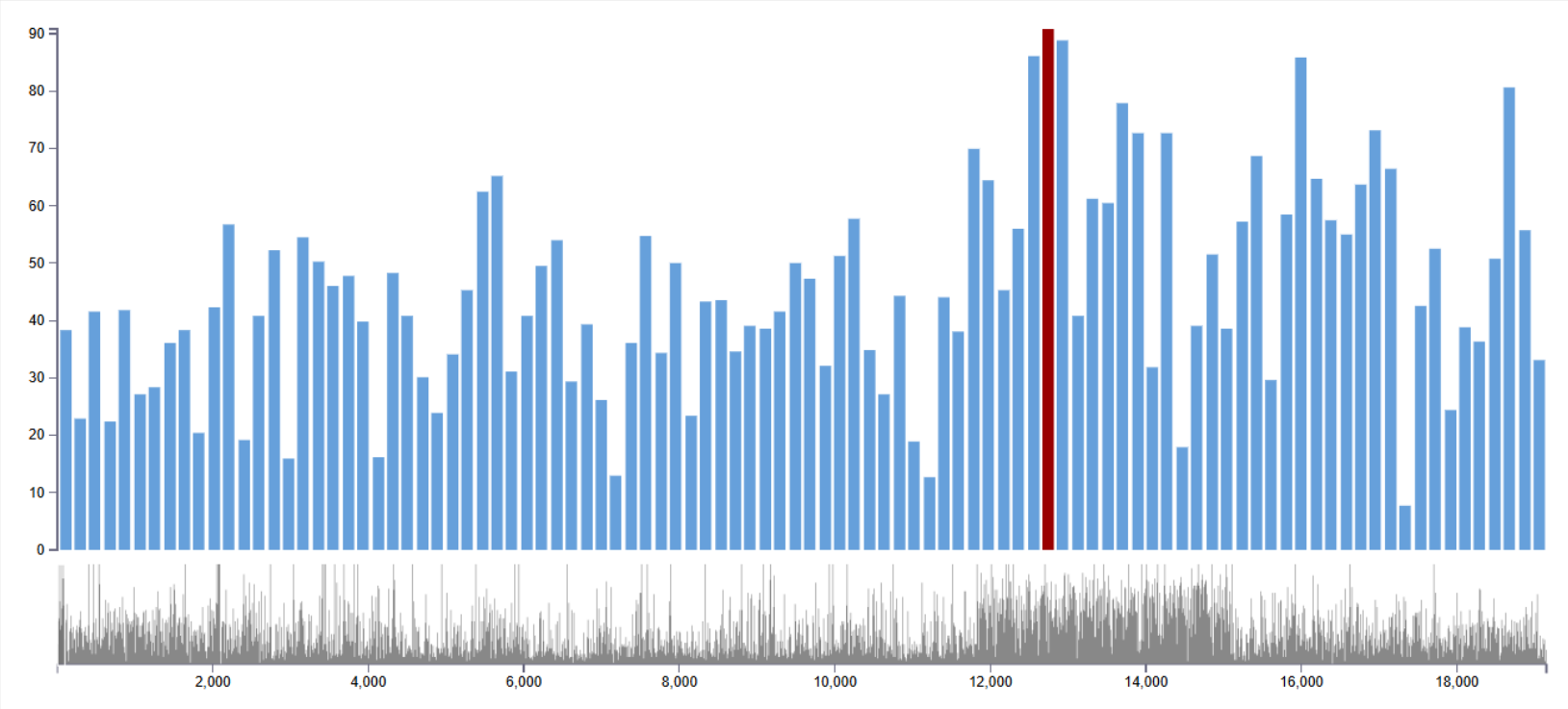
other models such as Mistral-7B, VinAI Translate, Google Translate, Madlad400 7B (Kudugunta et al. 2024), GPT-3.5-Turbo (OpenAI 2023), and Gemini-1.0-Pro (Google 2024a). Additionally, our ALMA-Gemma-7B-IT model also shows notable advancements with a BLEU score of 52.70, considerably higher than competitive models.

These impressive results highlight the technical excellence of our models and validate the effectiveness of our architectural and methodological enhancements. These improvements include leveraging zero-shot learning capabilities and advanced instruction-based tuning, which together enable a more nuanced and accurate translation of text between English and Vietnamese.

Human Label Errors

Line ID	Source	Human Label	Ours Model
140	This is more than double the number imprisoned during the previous year.	Con số này nhiều hơn gấp đôi năm trước.	Con số này cao hơn gấp đôi so với số người bị giam giữ trong năm trước.
3949	Let me help you.	Để tôi ra ngoài.	Hãy để tôi giúp bạn.
11036	Well, he might come out tonight.	Thay vì ngồi trong này với ông.	Vâng , anh ấy có thể ra ngoài tối nay.
11037	As a matter of fact,	Ông ấy cũng có thể sẽ tới.	Trên thực tế,
15095	Keep it simple	Nói lời cảm ơn một cách giản dị	Đơn giản thôi

The table above compares translations from a human label with those produced by our model. It highlights several key areas where the human label errs primarily due to reliance on contextual knowledge, rather than model performance. It's clear that the human label errors largely stem from misinterpretation of context, which is crucial for accurate translation. Our model, on the other hand, tends to maintain the integrity of the original meaning more consistently.



<< < 67 > >> From 1 to 100 of all **19151** sentences.

Sentence **67**	**BLEU**	**Length ratio**	**Text**
Source	-	-	The soundtrack to pop star Son Tung's documentary has topped Vietnam's iTunes charts, entering the top 10 in several other countries following domestic release.
Human	100.00	1.00	Bản nhạc phim tài liệu của ngôi sao nhạc pop Sơn Tùng đã đứng đầu bảng xếp hạng iTunes của Việt Nam , lọt vào top 10 ở một số quốc gia khác sau khi phát hành trong nước .
Machine	90.83	0.95	Nhạc phim tài liệu của ngôi sao nhạc pop Sơn Tùng đã đứng đầu bảng xếp hạng iTunes Việt Nam , lọt vào top 10 ở một số quốc gia khác sau khi phát hành trong nước .

Figure 3: Distribution of BLEU scores on PhoMT Testset with EN-VI translation

4 Conclusion

Our research presents a groundbreaking approach to machine translation for the English-Vietnamese (EN-VI) and Vietnamese-English (VI-EN) language pairs, leveraging the robust capabilities of the **gemma-7b-it** model and the Advanced Language Model-based Translator (ALMA) methodology.

Through extensive experimentation, we have demonstrated that our fine-tuned models not only exceed the performance of traditional Transformer-based models but also establish a new state-of-the-art in this domain. Our models have achieved substantial improvements in BLEU scores, significantly surpassing the performance of leading translation systems such as VinAI Translate and Google Translate.

The integration of zero-shot learning and advanced instruction prompting has proven to be highly effective, enabling our models to produce translations with superior contextual understanding and stylistic fidelity. These advancements highlight the potential of Large Language Models (LLMs) in transforming machine translation tasks, offering both high accuracy and natural fluency.

Moreover, the successful deployment of our model into a user-centric translation product, available at *https://www.doctranslate.io*, underscores the practical applicability and excellence of our approach. The positive reception of this tool further validates our commitment to delivering cutting-edge translation solutions that meet diverse user needs with exceptional quality.

Our achievements pave the way for future research and innovation in the field of language translation. We anticipate that continued advancements in LLMs and fine-tuning methodologies will yield even more refined and reliable translation systems, setting new benchmarks in multilingual communication and enhancing global connectivity.

References

[1] Boris Buden et al. "Cultural translation: An introduction to the problem, and responses". In: Translation Studies 2.2 (2009), pp. 196–219.

[2] Long Doan et al. "Phomt: A high-quality and large-scale benchmark dataset for vietnamese-english machine translation". In: arXiv preprint arXiv:2110.12199 (2021).

[3] Doctranslate. Doctranslate. 2023. URL: https://doctranslate.io/.

[4] Google. Google AI. 2024. URL: https://blog.google/technology/ai/google-gemini-ai.

[5] Google. Google Translate. 2024. URL: http://translate.google.com.

[6] Albert Q Jiang et al. "Mistral 7B". In: arXiv preprint arXiv:2310.06825 (2023).

[7] Sneha Kudugunta et al. "Madlad-400: A multilingual and document-level large audited dataset". In: Advances in Neural Information Processing Systems 36 (2024).

[8] Tuan-Duy H Nguyen et al. "A Vietnamese-English Neural Machine Translation System." In: INTERSPEECH. 2022, pp. 5543–5544.

[9] OpenAI. GPT-4 Technical Report. 2023. URL: https://openai.com/.

[10] Gemini Team et al. "Gemini: a family of highly capable multimodal models". In: arXiv preprint arXiv:2312.11805 (2023).

[11] Gemma Team et al. "Gemma: Open models based on gemini research and technology". In: arXiv preprint arXiv:2403 (2024).

[12] Haoran Xu et al. "A paradigm shift in machine translation: Boosting translation performance of large language models". In: arXiv preprint arXiv:2309.11674 (2023).

[13] Wayne Xin Zhao et al. "A survey of large language models". In: arXiv preprint arXiv:2303.18223 (2023).