# VN-MTEB: Vietnamese Massive Text Embedding Benchmark

Anonymous ACL submission


## Abstract

Vietnam ranks among the top countries in terms of both internet traffic and online toxicity. As a result, implementing embedding models for recommendation and content control duties in applications is crucial. However, a lack of large-scale test datasets, both in volume and task diversity, makes it tricky for scientists to effectively evaluate AI models before deploying them in real-world, large-scale projects. To solve this important problem, we introduce a Vietnamese benchmark, VN-MTEB for embedding models, which we created by translating a large number of English samples from the Massive Text Embedding Benchmark using our new automated framework, thereby contributing an extension of the Massive Multilingual Text Embedding Benchmark with our additional Vietnamese tasks and datasets. We leverage the strengths of large language models (LLMs) and cutting-edge embedding models to conduct translation and filtering processes to retain high-quality samples, guaranteeing a natural flow of language and semantic fidelity while preserving named entity recognition (NER) and code snippets. Our comprehensive benchmark consists of 41 datasets from six tasks specifically designed for Vietnamese text embeddings. In our analysis, we find that bigger and more complex models using Rotary Positional Embedding outperform those using Absolute Positional Embedding in embedding tasks.


## 1 Introduction

Recent advancements in Large Language Models (LLMs) (Grattafiori et al., 2024; DeepSeek-AI et al., 2025; Team et al., 2025) have led to significant improvements in various Natural Language Processing (NLP) tasks. To the best of our knowledge, numerous benchmarks have been established for NLP tasks; they predominantly focus on widely spoken languages such as English and Chinese (Muennighoff et al., 2023). In contrast, low-resource languages like Vietnamese, which is spoken by over 100 million people [1], have yet to benefit from the creation of large-scale benchmarks. Although several datasets have been published, including ViQuAD (Nguyen et al., 2020), ViMMRC (Van Nguyen et al., 2020), and UIT-VSFC (Nguyen et al., 2018), these resources are often limited to a single task and domain, with a noticeable scarcity in their publication.

Text embedding methods (Cao, 2024) have become increasingly popular in both industrial and academic fields due to their critical role in a variety of natural language processing tasks. The significance of universal text embeddings has been further highlighted with the rise of LLMs applications such as Retrieval-Augmented Systems (RAGs) (Lewis et al., 2021). Consequently, researchers who seek to evaluate models must often resort to manually collecting datasets and converting them into formats suitable for model evaluation, a process that is both time-consuming and labor-intensive. The Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2023) was created to collect data and standardize ways to evaluate and score different text embedding models. Later the MMTEB: Massive Multilingual Text Embedding Benchmark (Enevoldsen et al., 2025) introduced more dataset for many language, including low-resource like Vietnamese. However, in MMTEB, Vietnamese has only 18 datasets, while English has more than 300, German has 80, and Mandarin Chinese has over 44. This work aims to increase the number of Vietnamese datasets by adding 41 more, thereby creating a larger, more reliable, and more challenging benchmark that enables more accurate conclusions about embedding model performance across a wide range of tasks and domains.

Machine translation methods often require hu-

[1] https://www.macrotrends.net/global-metrics/countries/vnm/vietnam/population

man intervention for quality verification (Qian et al., 2024), sample collection for benchmarks, and overall evaluation, leading to a significant increase in effort. To address this challenge, our approach integrates translation with additional quality assurance to ensure that our translated datasets satisfy key criteria. By utilizing the latest state-of-the-art models in text embedding, language detection, and LLMs for automatic translation and filtering of low-quality samples, However, to ensure the benchmark's reliability and quality, we acknowledge the importance of human evaluation. Including a human evaluation of translation quality, even on a small subset, will further strengthen the claim that the resulting benchmark is both high-quality and a valuable resource for the community. This approach strikes a balance between high resource consumption (time, infrastructure) and high-quality output, with a significantly reduced human effort.

Recognizing the need for a standardized benchmark, this paper introduces VN-MTEB (Vietnamese Massive Text Embedding Benchmark). The scope and key contributions of this work are as follows.

- We introduce **VN-MTEB** - a substantial benchmark consisting of **41 datasets** from **6 tasks** (retrieval, reranking, classification, clustering, pair classification, and semantic textual similarity), designed to evaluate text embeddings for the Vietnamese language. This is an extension of MMTEB for the Vietnamese subset.

- We contribute to and integrate with MTEB[2] and make the source code used in the experiments available to the public.

- We evaluate a collection of embedding models, including both multilingual and monolingual variants, on the VN-MTEB benchmark, and provide insights into the correlation between model types and their performance across various tasks.

- We propose a translation method that enables strict control over the fidelity of synthesized samples by considering multiple evaluation criteria. The goal of this approach is to facilitate translation tasks requiring minimized human involvement in either the translation or the quality assurance process.

[2]https://huggingface.co/spaces/mteb/leaderboard

## 2 Related Works

### 2.1 Benchmarks, MTEB and MMTEB

GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019), Big-BENCH (Srivastava et al., 2023), and evaluation frameworks (Gao et al., 2024) play a crucial role in driving NLP progress. However, they are not suitable for evaluating text embedding, so dedicated benchmarks such as SentEval (Conneau and Kiela, 2018), often known as a benchmark for semantic textual similarity (STS), USEB (Wang et al., 2021), introduced with additional reranking tasks, and Beir (Thakur et al., 2021) have become the standard for embedding evaluation for zero-shot information retrieval. The MTEB (Muennighoff et al., 2023) incorporates the above benchmarks and consists of 58 datasets covering 112 languages from 8 embedding tasks: bitext mining, classification, pair classification, clustering, reranking, retrieval, semantic textual similarity (STS), and summarization. Our work follows the structure and is compatible with the current working source of MTEB.

Our VN-MTEB integrates a wide range of datasets, including clustering, classification, BEIR (retrieval) (Thakur et al., 2021), and others from various tasks, to provide a comprehensive and reliable performance assessment of text embedding models in Vietnamese.

### 2.2 Translation Pipeline

In Beir-PL (Wojtasik et al., 2024), the verification process involved randomly selecting 100 query-passage pairs, assessed by a linguist in a strict setting and a researcher in a semantic setting. Additionally, an automated comparison was conducted using the multilingual LaBSE model (Feng et al., 2022), as in the original paper, to compare source texts and translations automatically. The paper applied machine translation with a large language model (Yang et al., 2023), where the LLM first generates a draft translation. The pipeline then retrieves similar translation pairs and feedback from the database as in-context examples, allowing the model to refine the draft based on these domain-specific revisions. Furthermore, LLM can be used with various prompt templates to predict human-annotated direct assessment for translation quality (Qian et al., 2024). They also explored different prompting techniques, including chain-of-thought (CoT) (Wei et al., 2022), which involves a two-step process where the LLM first analyzes the differ-

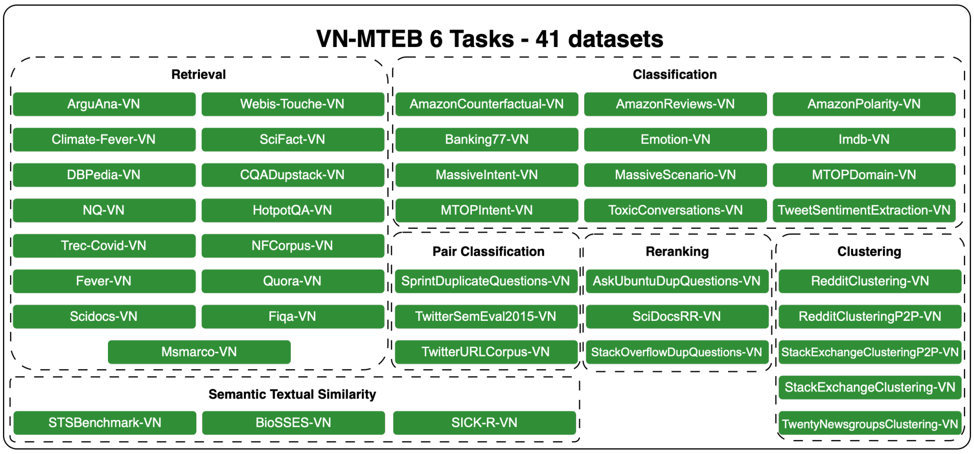


Figure 1: An overview of tasks and datasets in VN-MTEB.

ences between the machine translation output and the reference and then scores the translations based on its analysis. In our method, we utilize the embedding model to compare the equivalence between the original text and its translation, while the LLM analyzes and scores the translation quality, allowing us to create a high-quality translated dataset without relying on human effort.

### 2.3 Embedding models

Embedding models create vector representations for tokens, with a key challenge being how they handle positional information in sequences. Our paper extends the foundation laid by (Zhu et al., 2024) on classifying embedding models. It explores architectures like Absolute Positional Encoding (APE) and Rotary Positional Encoding (RoPE), alongside tuning strategies including Instruct-tuned and Non-Instruct-tuned methods. To incorporate positional embeddings into token embeddings, most encoder-based text embedding models, such as the BERT architecture (Devlin et al., 2019), adopt the APE approach. In contrast, the RoPE method (Su et al., 2023) encoded positional information through rotational transformations applied directly to the query and key vectors within the attention mechanism. This approach adopted positional encoding strategies in the age of LLMs, with its use seen in models like LLaMA (Touvron et al., 2023) and Qwen (Bai et al., 2023).

The Instruct-tuned model refers to models that were trained with the natural language descriptions of the embedding tasks. Instructions can better inform embedding models about the task at hand, thereby enhancing the quality of the embeddings.

## 3 Methodology

Our goal is to create a large-scale benchmark that serves as a reference point for comparing different text embedding models in Vietnamese. To achieve this, we focus on a language with a substantial volume of data instances available in the MTEB benchmark and translate its dataset into Vietnamese. For each criterion, we explore the flexible use of embedding models or the application of CoT prompting techniques (Wei et al., 2022) in large language models to perform evaluation. The objective is to select high-quality synthesized samples while maintaining performance and ensuring resource efficiency.

The Figure 2 illustrates our pipeline for generating a synthesized dataset by transforming a source dataset into a low-resource language. Our pipeline consists of three main stages:

- **Stage 1:** The purpose of this stage is to filter out only the samples in the desired source language. Supposing the original dataset is multilingual, we employ language detection using a LLM to detect the language in the original dataset, keeping only samples in the desired source language. Future studies aiming to translate the entire dataset may omit this stage.

- **Stage 2:** This stage employs the LLM to translate the dataset. The result is a set of Vietnamese sequences that exhibit high similarity to the original texts while preserving semantic fidelity, named entity recognition (NER), code snippets, and other critical aspects, which will be further examined and evaluated in the subsequent stage.

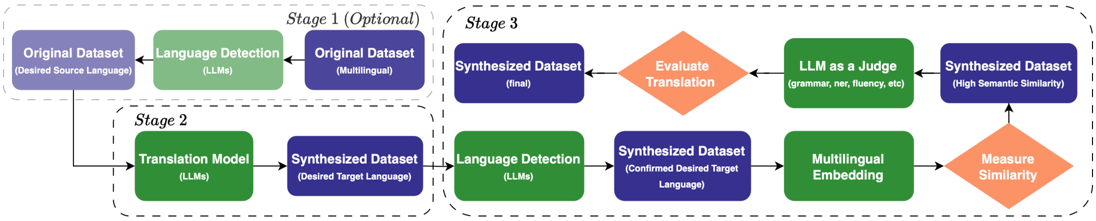


Figure 2: An overview of translation pipeline.

- **Stage 3:** We evaluate the generated translations used in the official VN-MTEB through a three-step process, with each step reflecting an increasing level of rigor. First, we assess whether the data contains any contamination from other languages. Second, we ensure that the data preserves high semantic similarity with the original content. Finally, we score each synthesized sample based on a combination of multiple evaluation criteria. We discard all data samples whose scores fall below the predefined threshold.

**Translation.** The generated sequences must achieve high quality to minimize the likelihood of being filtered out during the validation stage. Therefore, selecting an appropriate LLM is crucial. In this stage, we recommend using an LLM with at least a medium-sized model and support for maximum token lengths in the tens of thousands. Additionally, we consider utilizing models that demonstrate strong performance on the target language by consulting relevant leaderboards, such as SEA-HELM[3].

Evaluating the quality of model-generated translations is crucial, as embedding models require high-quality datasets for both training and testing. While human evaluation can ensure the quality of translations on a small subset, the sheer volume of data presents a significant challenge. To address this, beside human manual evaluation step, we propose a series of data filtering steps to ensure that the final synthesized dataset preserves essential NLP properties while optimizing the framework's execution efficiency.

**Language Detection.** We employ a lightweight LLM for language detection to identify samples in the desired source language for translation (Stage 1). While LLMs are generally proficient at translating text, they may misidentify the language when multiple languages are present or when the text includes uncommon phrases, regional dialects, or jargon (Qian et al., 2024). Additionally, translations may not always capture contextual nuances, idioms, or cultural subtleties. In (Qian et al., 2024), the shortcomings noted in the LLM's initial translation output are primarily related to domain-specific nuances, terminology, and sometimes word order or structure. Therefore, we also leverage the same language detection model used in Stage 1 to verify whether the translated outputs are entirely in Vietnamese in Stage 3.

**Semantic Similarity.** The translated text must maintain semantic equivalence with the original sentence. Therefore, we consider using multilingual embeddings to compute similarity scores between sentence pairs and subsequently filter the data based on a predefined threshold. A key factor in selecting an evaluation model is ensuring that the inferred score distributions for similar and unrelated sentence pairs are well separated. Additionally, the model's maximum sequence length should be relatively large (preferably greater than or equal to 8192 tokens) to fully encode the content of each sequence. To determine the optimal threshold for specific models, we need to balance the separation of similarity scores between semantically related and contradictory pairs while minimizing the number of incorrectly filtered samples. (See Section 5 for a more detailed discussion.).

**LLM as a Judge.** In addition to ensuring consistency in the target language and maintaining semantic similarity to the input sequence, other criteria should also be considered to guarantee that the synthesized samples are of high quality and aligned with human knowledge. Since translation is fundamentally about generating text that is both accurate and aligned with human linguistic expectations in a different language, the findings of (Zheng et al., 2023) are directly relevant to and encouraging for the application of LLM-as-a-Judge for quality assurance in LLM-based translation. The advantages discussed in the paper include scalability and explainability, which support the

[3]https://leaderboard.sea-lion.ai/

reason why we are using LLM to judge a large-scale dataset's translation quality. In this paper, we leverage LLMs at this stage to evaluate the following criteria: grammar, named entity recognition (NER), numbers/links/special characters, fluency, and meaning preservation. The following generalized formula computes the final score for each output:

$$\text{score}_{\text{LLM\_judge}} = \frac{\sum_{i \in S} \alpha_i \cdot \text{score}_i}{|S|}, \quad (1)$$

where $S$ is the set of evaluation criteria, $\sum_{i \in S} \alpha_i = 1$, $\alpha_i$ and $\text{score}_i \in [1, 5]$ denote the importance weight and the score of criterion $i$, respectively. Synthesized translations whose score $score_{LLM\_judge}$ exceeds the threshold $\xi_{LLM\_judge}$ are selected.

# 4 VN-MTEB

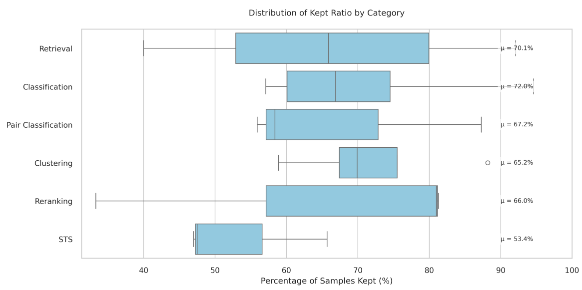


Figure 3: Kept Ratio by Tasks.

In Figure 1 and Table 1, we present an overview of the sample collection and count with multi-step filtering, comparing the original dataset (labeled as "Before") with the final set of samples obtained after processing through the translation pipeline, which utilizes semantic similarity and a LLM-based judge filter. In our approach, we treat each sequence as an individual sample for the purpose of Stage 3, which is translation validation. Consequently, the sample count may differ from that of the original dataset (Muennighoff et al., 2023) and the dataset statistic D after formatting to be compatible with MTEB code. To the best of our knowledge, our research release large-scale datasets, which cover the diverse set of tasks for benchmarking Vietnamese embedding models, comprising 41 datasets across 6 tasks. This is an extension of MMTEB for the Vietnamese subset. Full detail of comparison between VN-MTEB and MMTEB is at E

Table 1: The overview of **VN-MTEB**.

| Dataset Name | # Samples (Original) | # Filter 1 (Semantic Similarity) | # Filter 2 (LLM as a judge) | % Kept (Final/Before) |
|---|---|---|---|---|
| **Retrieval** | | | | |
| ArguAna-VN | 1,406 | 1,209 | 1,295 | 92.1% |
| Touche2020-VN | 2,214 | 2,190 | 1,138 | 51.4% |
| ClimateFEVER-VN | 4,681 | 4,088 | 3,401 | 72.6% |
| CQADupstack-*-Retrieval-VN | 19,938 | 17,567 | 13,140 | 65.9% |
| DBPedia-VN | 49,188 | 45,561 | 39,551 | 80.4% |
| FEVER-VN | 16,016 | 14,224 | 12,739 | 79.5% |
| FiQA2018-VN | 1,706 | 1,829 | 1,021 | 59.8% |
| HotpotQA-VN | 25,704 | 23,156 | 21,956 | 85.5% |
| MSMARCO-VN | 16,697 | 12,089 | 8,019 | 48.0% |
| NFCorpus-VN | 12,334 | 10,201 | 6,819 | 55.2% |
| NQ-VN | 4,201 | 3,091 | 2,283 | 54.4% |
| QuoraRetrieval-VN | 23,301 | 20,077 | 17,135 | 73.5% |
| SCIDOCS-VN | 29,928 | 25,101 | 11,969 | 40.0% |
| SciFact-VN | 339 | 205 | 155 | 45.7% |
| TRECCOVID-VN | 66,336 | 61,624 | 57,358 | 86.4% |
| **Classification** | | | | |
| EmotionVNClassification | 4,000 | 3,469 | 2,570 | 64.3% |
| Banking77VNClassification | 13,083 | 12,989 | 12,378 | 94.6% |
| ToxicConversationsVNClassification | 50,000 | 31,299 | 28,560 | 57.1% |
| ImdbVNClassification | 25,000 | 24,721 | 22,081 | 88.3% |
| TweetSentimentExtractionVNClassification | 3,534 | 3,145 | 2,065 | 58.5% |
| AmazonCounterfactualVNClassification | 1,005 | 802 | 711 | 70.7% |
| MTOPDomainVNClassification | 30,517 | 28,129 | 20,414 | 66.9% |
| MTOPIntentVNClassification | 30,517 | 28,129 | 20,414 | 66.9% |
| AmazonReviewsVNClassification | 9,990 | 8,792 | 6,766 | 67.8% |
| MassiveIntentVNClassification | 5,005 | 4,128 | 3,005 | 60.1% |
| MassiveScenarioVNClassification | 5,006 | 3,892 | 3,006 | 60.1% |
| AmazonPolarityVNClassification | 400,000 | 389,124 | 344,197 | 86.0% |
| **Pair Classification** | | | | |
| SprintDuplicateQuestions-VN | 202,000 | 189,224 | 176,259 | 87.3% |
| TwitterSemEval2015-VN | 16,777 | 12,144 | 9,374 | 55.9% |
| TwitterURLCorpus-VN | 51,534 | 40,829 | 30,111 | 58.4% |
| **Clustering** | | | | |
| TwentyNewsgroupsClustering-VN | 59,436 | 49,891 | 45,034 | 58.9% |
| RedditClustering-VN | 190,653 | 151,128 | 133,217 | 69.9% |
| RedditClusteringP2P-VN | 438,322 | 404,290 | 331,020 | 75.5% |
| StackExchangeClustering-VN | 35,052 | 29,824 | 23,618 | 67.4% |
| StackExchangeClusteringP2P-VN | 73,577 | 67,525 | 64,869 | 88.2% |
| **Reranking** | | | | |
| AskUbuntuDupQuestions-VN | 375 | 349 | 305 | 81.3% |
| StackOverflowDupQuestions-VN | 2,992 | 2,787 | 2,421 | 81.0% |
| SciDocsRR-VN | 7,959 | 5,912 | 2,656 | 33.3% |
| **Semantic Textual Similarity** | | | | |
| STSBenchmark-VN | 2,879 | 2,329 | 1,891 | 65.7% |
| BIOSSES-VN | 100 | 60 | 47 | 47.0% |
| SICK-R-VN | 9,927 | 7,485 | 4,716 | 47.5% |

**Kept ratio.** The percentage of retained samples (% Kept) is determined by the ratio of the final sample count to the original sample count. The varying kept ratios suggest different levels of data quality and filtering requirements across tasks. Some datasets have a kept ratio lower than 50%, indicating that half of the translations were invalid due to complexities in grammar and semantics, which are difficult to translate, as well as issues with passing quality control in Stage 3 of our pipeline. Further implementation detail please refer to section 5.

**Word length.** Since both English and Vietnamese originate from Latin roots, analyzing the distribution of word lengths between original and synthesized samples has the potential to reflect translation quality. We conduct a statistical analysis over a word length range that covers the majority of samples in the VN-MTEB dataset. Figure 4 compares the distributional trends over a dataset consisting of millions of sample pairs. The results reveal a strong correlation between Vietnamese and English word lengths. This observation serves as supporting evidence for translation quality assessment, in addition to the evaluation criteria discussed in Section 3.

For more detailed statistics, please refer to our Table 13 for information on the train, dev, and test split samples, and see G for further details about GPU usage and the time spent creating all datasets.

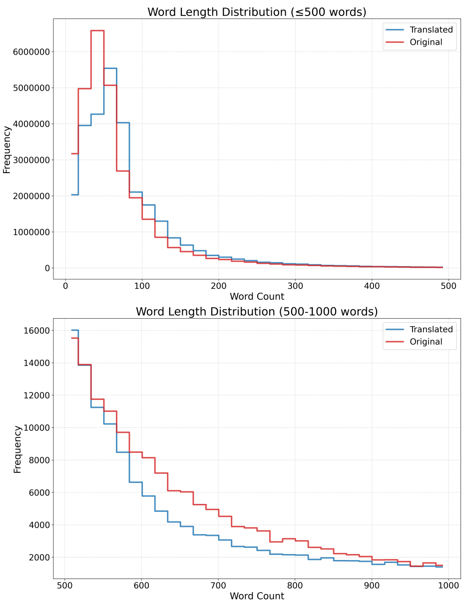


Figure 4: Word Length Distribution between Original and Translated in overall dataset.

## 5 Experiments

### 5.1 Implementation Details

In this part, we provide a detailed report on the models and hyperparameters used for dataset translation and verification. In our pipeline, we refer to the Seahelm leaderboard[4] and select Qwen/Qwen2.5-3B-Instruct [5] to perform detecting language, which was the top model with the relatively small size compared to the time our experiment was conducted. The choice of model at translation stage is guided by a trade-off between translation quality and the computational cost of processing large-scale resources, potentially involving millions of documents. Throughout the course of this research, we evaluated a diverse set of machine translation models, including pre-trained multilingual models such as SeamlessM4T (Communication et al., 2023), M2M100 (Fan et al., 2020), and NLLB-200 (Team et al., 2022), all of which represent significant advancements in cross-lingual representation learning. Additionally, we also evaluated state-of-the-art bilingual translation models tailored specifically for English–Vietnamese translation, including EnViT5-Translation (Ngo et al., 2022) and VinAI-Translate-En2Vi (Nguyen et al., 2022). There are limitations of prior machine translation works such as VinAI-Translate-En2Vi (Nguyen et al., 2022), which is short context length (1024) and limitation of domain trained. API-based models like OpenAI's GPT-4, Google's Gemini, etc are costly to translate on a massive dataset. At the time the experiment and translation were conducted, we chose the best model according to SouthEast Asian Holistic Evaluation of Language Models (SEA Healms) [6] that time (May 23, 2024), we used Coherence AI's Aya-23-35B (Aryabumi et al., 2024), which has relatively good performance on Vietnamese, and the model size is relatively feasible (35 billion parameters). We utilize the embedding model Alibaba-NLP/gte-Qwen2-7B-instruct [7]text to compute semantic similarity for embedding-based evaluations. The advantage of deploying this model lies in its ability to encode long sequences (up to 32,768 tokens). For the "LLM-as-a-Judge" evaluation framework, we adopt aisingapore/Llama-SEA-LION-v3-70B-IT as the scoring model. According to the SEA Healms benchmark, this model currently demonstrates the strongest performance for Vietnamese. To enhance judgment quality, we further incorporate chain-of-thought (CoT) prompting techniques in the evaluation process.

In our research, we used 4 NVIDIA H100 GPUs to run our pipeline. For a full estimate about the resource usage, please refer to Appendix G for GPU usage , and for LLMs hyperparameters in translation, please refer to Appendix Table 4 .

### 5.2 Experimental Results

**Language Detection.** A conventional approach for language detection on text sequences is to employ FastText (Joulin et al., 2017). However, synthesized texts often contain interleaved characters from multiple languages, as discussed in Section 3. Through our experiments, we demonstrate that FastText frequently yields inaccurate predictions in such cases. Consequently, leveraging a lightweight large language model (LLM) in conjunction with the CoT technique proves to be a more effective solution for detecting the language of generated samples. Visual results are presented in Table 2.

**Translation.** Table 1 presents the results obtained using the selected translation model Aya-23-35B (Aryabumi et al., 2024). Our pipeline demonstrates strong translation performance across most datasets, achieving a relatively high reten-

[4]https://leaderboard.sea-lion.ai
[5]https://huggingface.co/Qwen/Qwen2.5-3B-Instruct
[6]https://leaderboard.sea-lion.ai
[7]https://huggingface.co/Alibaba-NLP/gte-Qwen2-7B-instruct

| Dataset Name | Translated Text | True Label | Qwen2.5-7B-Instruct | Qwen2.5-3B-Instruct | FastText |
|---|---|---|---|---|---|
| cqadupstack-mathematica-vn | Dựa trên một tập dữ liệu, có cách nào để thay đổi một giá trị? Ví dụ (*data = First@Import["dataset.xlsx"];*) data= {{"Supplier", "Material", "Geography", "Quantity"}, {"Acme", "A", "United States", 676.}... | vie_Latn | vie_Latn | vie_Latn | krc_Cyrl |
| webis-touche2020-vn | 2007 Hall of Fame BBWAA *(98,5%) Được chọn vào HOF năm 2007 bởi BBWAA All-Star Games 1983 * 1984 (SS) 1985 (SS) 1986 (SS) 1987 (SS) 1988 (SS) 1989 (SS) 1990...* | vie_Latn | vie_Latn | vie_Latn | kor_Hang |
| msmarco-vn | Ga Amtrak gần Buena Park: 1 5 dặm: FULLERTON (120 E. SANTA FE AVE.) . 2 8 dặm: ANAHEIM (2150 KATELLA AVE.) . 3 12 dặm: SANTA ANA (1000 E. SANTA ANA BLVD.)... | vie_Latn | vie_Latn | vie_Latn | kor_Hang |

Table 2: Comparison of Vietnamese Language Identification: Qwen2.5-7B-Instruct vs Qwen2.5-3B-Instruct vs. FastText.

tion rate and satisfactory quality in terms of preserving semantic meaning, named entities, and other key elements. Although some datasets, such as SciDocsRR-VN, SCIDOCS-VN, and Scifact-VN, exhibit retention rates below 50%, these belong to the scientific domain, which poses particular challenges for translation.

**Semantic Similarity.** Figure 5 illustrates the percentage distribution of semantic similarity score regions (binned in intervals of 0.1) for different sentence pairs, including original English sentences with their corresponding Vietnamese labels, semantically similar English sentences, contradictory Vietnamese sentences, and unrelated Vietnamese sentences. We evaluate 500 samples from the FLoRes [8] dataset, which provides pre-aligned English-Vietnamese sentence pairs. The remaining sentence categories for semantic comparison are manually curated by bilingual experts. The results presented in Figure 5 indicate a clear separation in the semantic similarity score distribution between original English sentences paired with their Vietnamese labels and semantically similar English sentences, compared to the other sentence pairs. Based on these results, we discard generated texts that scores fail to satisfy the minimum threshold of 0.8.

**LLM as a Judge.** This step involves evaluating translations based on criteria such as grammar, named entities, fluency, and more. Since translation is essentially about producing text that is both accurate and conforms to human linguistic standards in another language, the findings from (Zheng et al., 2023) are relevant and encouraging for using LLM-as-a-Judge in quality assurance for LLM-based translations. The paper highlights advantages such as scalability and explainability, which justify using LLM to assess translation quality across large datasets. Although the LLM as a Judge has limited reasoning, with Chain-of-Thought (CoT) prompting techniques (Wei et al., 2022), CoT guides LLMs in evaluation tasks by breaking down the entire evaluation process into smaller steps with detailed definitions and constraints for each step in the prompts. We used this technique to design the prompt guiding the LLM to step-by-step generate an explanation and then scoring the translation. We're using a prompt that is described in Figure 6.

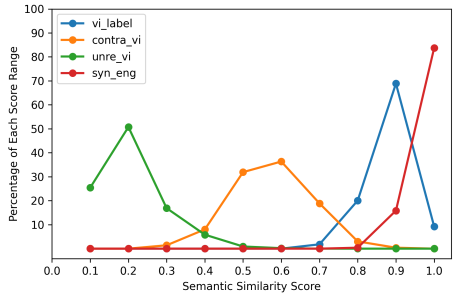


Figure 5: The distribution of semantic similarity score using Alibaba-NLP/gte-Qwen2-7B-instruct. vi_label, contra_vi, unre_vi, and syn_eng respectively represent the semantic similarity scores between the original English sequences and the corresponding labeled Vietnamese sequences, contrastive Vietnamese sequences, unrelated Vietnamese sequences, and synonymous English sequences.

```
LLM_AS_A_JUDGE = """
You are an expert in English-to-Vietnamese translation
evaluation, specializing in linguistic accuracy, natural fluency, and
computational assessment.
You will be provided with an original English sentence
and its Vietnamese translation.
Your task is to evaluate the translation based on the following
criteria (0-5 for each):
Grammar (30%) - Correct sentence structure, word order, and verb agreement.
NER Accuracy (25%) - Proper translation or retention of names, places, brands.
Numbers, Links, Special Characters (20%) -
Ensure correct handling of numbers, URLs, emails, and symbols.
Fluency & Naturalness (15%) - Smooth, natural Vietnamese phrasing.
Meaning Preservation (10%) - No loss or distortion of meaning.

Return the result in strict JSON format with the following structure,
with additional explanation:
{
"explanation": <reason>,
"grammar": <score>,
"ner_accuracy": <score>,
"numbers_links_special_chars": <score>,
"fluency": <score>,
"meaning_preservation": <score>,
"final_score": <weighted_average_score>
}
Output:
"""
```

Figure 6: LLM as a Judge prompt.

The VN-MTEB dataset is the result of considerable efforts in translation and evaluation. Given the constraints of time and resources, we opted to outsource the scoring of translation samples to a large language model (LLM).

An overview of the final dataset, along with

[8]https://github.com/facebookresearch/flores

| | Size | Dim | Type | Retr. | Class. | PairClass. | Clust. | Rerank. | STS | Avg. ↑ |
|---|---|---|---|---|---|---|---|---|---|---|
| Num. Datasets (→) | (Params) | (Dim) | | 15 | 12 | 3 | 5 | 3 | 3 | 41 |
| gte-Qwen2-7B-instruct* | 7B | 3584 | RoPE | **46.05** | 70.76 | 72.09 | **53.15** | 74.28 | 78.73 | 65.84 |
| e5-Mistral-7B-instruct* | 7B | 4096 | RoPE | 41.73 | 72.21 | 84.01 | 51.71 | **75.15** | 81.20 | 67.67 |
| bge-multilingual-Gemma2* | 9B | 3584 | RoPE | 20.52 | 71.78 | 66.97 | 40.13 | 64.21 | 66.11 | 54.95 |
| gte-Qwen2-1.5B-instruct* | 1.5B | 1536 | RoPE | 42.01 | 67.14 | 72.70 | 47.64 | 71.37 | 79.97 | 63.47 |
| m-e5-large-instruct* | 560M | 1024 | APE | 40.88 | **73.39** | **84.47** | 52.96 | 73.28 | **82.94** | **67.99** |
| m-e5-large | 560M | 1024 | APE | 37.65 | 65.03 | 83.70 | 45.78 | 70.40 | 80.65 | 63.87 |
| bge-m3 | 568M | 1024 | APE | 39.84 | 69.09 | 84.43 | 45.90 | 71.28 | 78.84 | 64.90 |
| Vietnamese-Embebedding | 568M | 1024 | APE | 34.18 | 69.06 | 82.84 | 45.61 | 70.89 | 77.48 | 63.34 |
| KaLM-embedding-m-mini-v1 | 494M | 896 | RoPE | 35.07 | 62.84 | 79.95 | 46.85 | 68.85 | 78.54 | 62.02 |
| LaBSE | 471M | 768 | APE | 17.77 | 60.93 | 77.57 | 34.59 | 65.65 | 72.04 | 54.76 |
| gte-multilingual-base | 305M | 768 | APE | 38.38 | 64.99 | 84.42 | 50.25 | 71.78 | 81.51 | 65.22 |
| m-e5-base | 278M | 768 | APE | 34.50 | 63.29 | 82.51 | 45.70 | 69.07 | 79.45 | 62.42 |
| halong-embedding | 278M | 768 | APE | 34.45 | 63.33 | 81.20 | 43.42 | 69.83 | 77.39 | 61.60 |
| m-e5-small | 118M | 384 | APE | 34.12 | 60.27 | 81.18 | 43.16 | 67.69 | 77.56 | 60.66 |
| vietnamese-bi-encoder | 135M | 768 | APE | 25.37 | 58.92 | 77.40 | 34.13 | 64.95 | 68.58 | 54.89 |
| sup-SimCSE-VN-phobert-base | 135M | 768 | APE | 12.03 | 59.69 | 71.31 | 33.05 | 58.86 | 68.61 | 50.59 |
| MiniLM-L12 | 33.4M | 384 | APE | 14.14 | 45.57 | 69.46 | 24.36 | 60.44 | 62.34 | 46.05 |
| MiniLM-L6 | 22.7M | 384 | APE | 9.65 | 45.19 | 66.13 | 20.40 | 59.46 | 58.25 | 43.18 |

Table 3: Average performance of the main metric (in percentage) per task and per model on VN-MTEB subsets. The symbol * indicates that the model is **Instruct-tuned**. Bold values highlight the best results for each specific task. The column "Avg." represents the mean of the average scores across all tasks.

the corresponding Kept ratio, is presented in Table 1, and Figure 3. The mean Kept ratio for the various tasks is as follows: Retrieval (15 datasets) – 66.03%, Classification (13 datasets) – 70.11%, Pair Classification (3 datasets) – 67.2%, Clustering (5 datasets) – 71.98%, Re-ranking (3 datasets) – 65.2%, and Semantic Textual Similarity (3 datasets) – 53.4%.

### 5.3 Benchmark Result

In this paper, we select open-source embedding models to perform benchmarking. In our benchmark, we classified two types of models: APE-based, RoPE-based, and Instruct-tuned models. Our benchmark results collected from 18 models and averaged from 41 datasets from 6 tasks are represented in Table 3. For more detail of model scoring on each dataset, please refer to Appendix J for results on all of the models we experimented with.

**Comparison of models**: As visualized in Figure 7, there is a clear correlation between the number of parameters in a model and its overall average VN-MTEB score. Larger models tend to achieve higher scores. Specifically, RoPE-based models, such as e5-Mistral-7B-Instruct and e5-Qwen2-7B-Instruct, generally outperform APE-based models like gte-multilingual-base, bge-m3, and m-e5-large. As mentioned in the preliminary section 2, instruct-tuned models were trained with task descriptions. This training approach typically results in higher overall performance, as evidenced by the significant performance improvement of the instruct-tuned m-e5-large-instruct compared to its non-instruct counterpart, m-e5-large. In the model evaluation process, we adhere to the methodology outlined in the MTEB task (Muennighoff et al., 2023). Specifically, we employ the model to embed both the queries and the corpus documents for the Retrieval task. Cosine similarity is then used to compute the similarity scores between each query and document. Next, we rank the corpus documents for each query based on their respective similarity scores and calculate various evaluation metrics. It is noteworthy that models with higher-dimensional representations tend to yield improved results in the retrieval task.

## 6 Conclusion

We utilize our proposed translation pipeline for translating 41 datasets from 6 tasks to create a massive text embedding benchmark from English to a low-resource language—Vietnamese. Through extensive experiments on our translation pipeline, we show that with LLMs we can delegate lots of effort from humans to translate a massive dataset with quality. Additionally, we evaluated 18 text embeddings and revealed the superiority of RoPE-based embedding models over APE-based ones in some tasks, giving an overview of choices to consider when selecting types of models to put in production and further research.

## Limitations

**Language variability** While this pipeline can be applied to any source language and translated into various low-resource languages, further research and analysis are required to determine the most suitable model for translation. In our study, we have selected LLMs and embeddings based on their performance with English and Vietnamese. For application to other languages, additional experiments must be conducted to identify the most appropriate model for each target language.

**Cultural context** Although our work comes from machine translation, datasets are still limited about the cultural context of the translation, such as formal, informal, or the specific dialect used.

**Absent of re-generation** Our pipeline does not guarantee the retention of all samples, resulting in some datasets being reduced by nearly half. Therefore, future research should consider incorporating a regeneration mechanism after the evaluation stage to improve the kept ratio.

**Insufficient analysis of synthetic data bias and contamination** During the research progress, we acknowledge this problem and thus, applying the quality filtering to minimize the error in translation, that introducing substantial data loss, and we also state this in our limitation **Absent of re-generation**. We recommend applying regeneration method to the quality filtering that ensure the quality of the translation and resolve the data loss.

**Long context** The VN-MTEB dataset encompasses a range of text lengths, including sequence-to-sequence, sequence-to-paragraph, and paragraph-to-paragraph formats. However, it lacks datasets comprising very long documents.

## References


Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, and 2 others. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Hongliu Cao. 2024. Recent advances in text embedding: A comprehensive review of top-performing methods on the mteb benchmark. *Preprint*, arXiv:2406.01607.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, and 46 others. 2023. Seamless: Multilingual expressive and streaming speech translation. *Preprint*, arXiv:2312.05187.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Veysel Çağatan, and 63 others. 2025. MMTEB: Massive multilingual text embedding benchmark. In *The Thirteenth International Conference on Learning Representations*.

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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

878–891, Dublin, Ireland. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. A framework for few-shot language model evaluation.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

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.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Chinh Ngo, Trieu H Trinh, Long Phan, Hieu Tran, Tai Dang, Hieu Nguyen, Minh Nguyen, and Minh-Thang Luong. 2022. Mtet: Multi-domain translation for english and vietnamese. *arXiv preprint arXiv:2210.05610*.

Kiet Nguyen, Vu Nguyen, Anh Nguyen, and Ngan Nguyen. 2020. A Vietnamese dataset for evaluating machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2595–2605, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kiet Van Nguyen, Vu Duc Nguyen, Phu X. V. Nguyen, Tham T. H. Truong, and Ngan Luu-Thuy Nguyen. 2018. Uit-vsfc: Vietnamese students' feedback corpus for sentiment analysis. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 19–24.

Thien Hai Nguyen, Tuan-Duy H. Nguyen, Duy Phung, Duy Tran-Cong Nguyen, Hieu Minh Tran, Manh Luong, Tin Duy Vo, Hung Hai Bui, Dinh Phung, and Dat Quoc Nguyen. 2022. A Vietnamese-English Neural Machine Translation System. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association: Show and Tell (INTERSPEECH)*.

Shenbin Qian, Archchana Sindhujan, Minnie Kabra, Diptesh Kanojia, Constantin Orasan, Tharindu Ranasinghe, and Fred Blain. 2024. What do large language models need for machine translation evaluation? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3660–3674, Miami, Florida, USA. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, and et al Adrià Garriga-Alonso. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. Roformer: Enhanced transformer with rotary position embedding. *Preprint*, arXiv:2104.09864.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

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.

Nandan Thakur, Nils Reimers, Andreas Ruckl'e, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *ArXiv*, abs/2104.08663.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Kiet Van Nguyen, Khiem Vinh Tran, Son T Luu, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen.

2020. Enhancing lexical-based approach with external knowledge for vietnamese multiple-choice machine reading comprehension. *IEEE Access*, 8:201404–201417.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. *SuperGLUE: a stickier benchmark for general-purpose language understanding systems*. Curran Associates Inc., Red Hook, NY, USA.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Konrad Wojtasik, Kacper Wołowiec, Vadim Shishkin, Arkadiusz Janz, and Maciej Piasecki. 2024. BEIR-PL: Zero shot information retrieval benchmark for the Polish language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2149–2160, Torino, Italia. ELRA and ICCL.

Xinyi Yang, Runzhe Zhan, Derek F. Wong, Junchao Wu, and Lidia S. Chao. 2023. Human-in-the-loop machine translation with large language model. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 88–98, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2024. LongEmbed: Extending embedding models for long context retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 802–816, Miami, Florida, USA. Association for Computational Linguistics.

## A Hyperparameters for Translation

In our translation pipeline, we used this configuration,

Table 4: Translation Hyperparameters

| **Hyperparameter** | **Value** |
|---|---|
| temperature | 0.0 |
| max_new_tokens | 4096 |
| tensor_parallel_size | 4 |
| max_model_len | 8192 |
| max_num_seqs | 256 |
| vllm_gpu_memory_utilization | 0.95 |

## B Model Translation Selection

We've tested other translation models and created a preference translation from human translations, randomly selecting 100 samples from 41 datasets based on document length and number of named entities. We present some of these samples as qualitative comparisons. As shown in Table 5, Aya-23-35B aligns more with human references than other models.
We use BLEU scoring metrics to measure the model outputs with the preference translation, as in the table below. We collect and represent some samples as the quantitative comparisons between models in Table 6. The Aya-23-35B gives a highest BLEU score on all tasks.

## C Examples

Tables 7-12 provide examples for each dataset for each task.

## D Dataset Statistics

Table 13 provides statistics of all VN-MTEB dataset (after processed and formatted). In our pipeline only the split test is considered to run on the translation verification.

## E Compare VN-MTEB and MMTEB

As previously discussed, the VN-MTEB is an extension of the MMTEB specifically designed for the Vietnamese language track. The details regarding the domain, subtask, and task of each dataset are provided in Table 14 and Table 15.

| **Aya-23-35B** | **NLLB** | **SeamlessM4T** | **Envit** | **Human_reference** | **dataset-task** | **original text** |
|---|---|---|---|---|---|---|
| **Một người đang đi xe đạp một bánh** | Một người đang đi xe đạp trên một bánh xe | Một người đang cưỡi xe đạp trên một bánh xe | Một người đang đi xe đạp trên một bánh xe | **Một người đang lái một chiếc xe đạp một bánh.** | SICK-R-VN | A person is riding the bicycle on one wheel |

Table 5: Comparison of model translation.

| Dataset | **Aya-23-35B** | **NLLB** | **SeamlessM4T** | **Envit** |
|---|---|---|---|---|
| AmazonCounterfactualVNClassification | 0.259626 | 0.120453 | 0.142945 | 0.149178 |
| MassiveIntentVNClassification | 0.177365 | 0.0893522 | 0.105255 | 0.0785961 |
| MassiveScenarioVNClassification | 0.187124 | 0.110192 | 0.147087 | 0.0762865 |
| AmazonPolarityVNClassification | 0.310677 | 0.153601 | 0.141143 | 0.194886 |
| AmazonReviewsVNClassification | 0.220978 | 0.0831676 | 0.0970089 | 0.116723 |
| ArguAna-VN | 0.329003 | 0.185319 | 0.156948 | 0.227969 |
| AskUbuntuDupQuestions-VN | 0.21701 | 0.126919 | 0.156549 | 0.133957 |
| Banking77VNClassification | 0.221967 | 0.167266 | 0.172289 | 0.145763 |
| BIOSSES-VN | 0.351678 | 0.215462 | 0.222327 | 0.241592 |
| ClimateFEVER-VN | 0.396737 | 0.155072 | 0.108626 | 0.254626 |
| CQADupstackMathematicaRetrieval-VN | 0.471219 | 0.247966 | 0.180818 | 0.257275 |
| DBPedia-VN | 0.347674 | 0.219082 | 0.239175 | 0.269178 |
| EmotionVNClassification | 0.201493 | 0.121039 | 0.127727 | 0.112674 |
| FEVER-VN | 0.386533 | 0.225274 | 0.168575 | 0.292802 |
| FiQA2018-VN | 0.327639 | 0.1806 | 0.10177 | 0.232315 |
| HotpotQA-VN | 0.429049 | 0.238127 | 0.242271 | 0.34017 |
| ImdbVNClassification | 0.356003 | 0.113464 | 0.0628686 | 0.185273 |
| MSMARCO-VN | 0.340505 | 0.196092 | 0.188766 | 0.22888 |
| MTOPDomainVNClassification | 0.202678 | 0.0759699 | 0.0741559 | 0.0196606 |
| MTOPIntentVNClassification | 0.205115 | 0.0901129 | 0.0713109 | 0.043651 |
| NFCorpus-VN | 0.452488 | 0.13407 | 0.0493039 | 0.231121 |
| NQ-VN | 0.415465 | 0.231762 | 0.202107 | 0.291572 |
| QuoraRetrieval-VN | 0.162906 | 0.152699 | 0.171029 | 0.158951 |
| RedditClusteringP2P-VN | 0.35617 | 0.131064 | 0.0958325 | 0.179491 |
| RedditClustering-VN | 0.273405 | 0.150356 | 0.176406 | 0.161824 |
| SciDocsRR-VN | 0.253023 | 0.177177 | 0.209065 | 0.183504 |
| SCIDOCS-VN | 0.406158 | 0.194238 | 0.104031 | 0.210699 |
| SciFact-VN | 0.412818 | 0.119744 | 0.0529717 | 0.145655 |
| SICK-R-VN | 0.177573 | 0.108461 | 0.116953 | 0.109692 |
| SprintDuplicateQuestions-VN | 0.404384 | 0.25737 | 0.271373 | 0.247144 |
| StackExchangeClusteringP2P-VN | 0.439518 | 0.139462 | 0.0987495 | 0.23968 |
| StackExchangeClustering-VN | 0.273632 | 0.148539 | 0.170056 | 0.194071 |
| StackOverflowDupQuestions-VN | 0.302429 | 0.182474 | 0.187633 | 0.189299 |
| STSBenchmark-VN | 0.123917 | 0.130112 | 0.153599 | 0.124351 |
| ToxicConversationsVNClassification | 0.324964 | 0.150357 | 0.141163 | 0.195978 |
| TRECCOVID-VN | 0.373649 | 0.150463 | 0.0838845 | 0.213836 |
| TweetSentimentExtractionVNClassification | 0.26379 | 0.0806592 | 0.125064 | 0.114902 |
| TwentyNewsgroupsClustering-VN | 0.206904 | 0.112983 | 0.109645 | 0.106614 |
| TwitterSemEval2015-VN | 0.116634 | 0.0433268 | 0.0547665 | 0.0461487 |
| TwitterURLCorpus-VN | 0.189587 | 0.114558 | 0.18553 | 0.163423 |
| Touche2020-VN | 0.241901 | 0.0837427 | 0.0882266 | 0.160647 |

Table 6: BLEU scores for different models

## F Dataset Licenses

Table 16 provides publicly available model checkpoints used for VN-MTEB evaluation.

## G GPU usage for translation

In our experiment, we utilized 4 H100 GPUs, each GPU electricity consumption is about 700W. As shown in Table 17, we measured an output token rate of 3,800 tokens per second. Since the entire process requires counting both input and output tokens, we multiply this rate by 2 to accurately estimate the time and energy consumption for each dataset as well as the overall workload. To summary, the estimated time to translate all VN-MTEB dataset is

$$\begin{aligned}\text{Total time} \times 2 &= 1,215,981.64 \text{ seconds} \times 2 \\ &= 2,431,963.28 \text{ seconds} \\ &\approx 675.54 \text{ hours} \\ &\approx 28.14 \text{ days}\end{aligned} \quad (918)$$

| Dataset | Query | Relevant-Document |
|---|---|---|
| ArguAna-VN | Trong mắt công chúng, chính phủ dường như nghi ngờ tất cả mọi người. | <*Title*> Nhà triết học chính trị cho rằng các quyền dân sự nên bị hy sinh <*Paragraph*> Đây chỉ là một cuộc điều tra như bất kỳ cuộc điều tra nào khác. Chính phủ rõ ràng phải có cách tiếp cận rộng rãi bởi vì bất kỳ lỗ hổng nào cũng có thể bị lợi dụng bởi những kẻ khủng bố vô đạo đức. Đó là một sự cần thiết, mặc dù cùng với những hậu quả không may, nhưng vẫn là sự cần thiết. Còn về đàm phán với những kẻ khủng bố, theo quan điểm của đề xuất này thì lựa chọn này không tồn tại khi đối phó với những kẻ khủng bố có nền tảng chủ nghĩa nguyên lý, vốn theo định nghĩa là không sẵn lòng thỏa hiệp và do đó không thể đàm phán được... |
| ClimateFEVER-VN | "Nếu bạn loại bỏ băng giá, có tiềm năng không chỉ là sự bất ổn định của vách băng sẽ bắt đầu xảy ra, nhưng một quá trình được gọi là sự bất ổn định của tấm băng biển", Matthew Wise, một nhà khoa học cực địa tại Đại học Cambridge nói. | <*Title*>Nam Cực <*Paragraph*> Nam Cực là lục địa phía Nam nhất trên Trái Đất. Nó bao gồm cực Nam Địa lý và nằm ở vùng Nam Cực của Bán cầu Nam, hầu hết về phía nam của Vòng Bắc Cực, và được bao quanh bởi Đại Dương Nam Cực. Với diện tích $14000000km^2$, đây là lục địa lớn thứ năm trên thế giới. So sánh với Úc thì diện tích của nó gấp đôi nước Úc . Khoảng 98% lãnh thổ bị băng tuyết che phủ với độ dày trung bình 1,9 km, kéo dài từ những nơi xa nhất về phía bắc đến Bán đảo Tây Nam Cực... |
| CQADupstack-*-Retrieval-VN | Làm thế nào để tôi có thể sử dụng Mathematica để tạo ra mã Fortran tốt hơn? | <*Title*>Tạo mã C/Java hiệu quả giảm thiểu các phép toán <*Paragraph*> Có thể dùng Mathematica để tạo ra mã C/Java nhằm tối thiểu hóa số lượng phép toán thực hiện không? Ví dụ, đối với ma trận nghịch đảo hay định thức? Với biến lưu trữ tốt? |
| DBPedia-VN | American sinh đôi nổi tiếng là vận động viên quần vợt chuyên nghiệp người Mỹ | <*Title*>Giải quần vợt chuyên nghiệp Nam Natomas <*Paragraph*> Giải quần vợt chuyên nghiệp Nam Natomas là một giải đấu quần vợt được tổ chức tại Sacramento, California, Hoa Kỳ từ năm 2005. Sự kiện này là một phần của ATP Challenger Tour và được chơi trên sân cứng ngoài trời. |
| FEVER-VN | Bee Gees đã viết ba bài hát cho các nghệ sĩ khác. | <*Title*> Bee Gees <*Paragraph*> Bee Gees là một nhóm nhạc pop được thành lập vào năm 1958. Thành viên của họ bao gồm ba anh em Barry, Robin và Maurice Gibb. Nhóm đã có những thành công lớn trong nhiều thập niên thu âm nhạc, nhưng họ cũng có hai giai đoạn đặc biệt nổi bật; đó là thời kỳ ca khúc tại vị trí số một trên bảng xếp hạng cuối thập niên 60 và đầu thập niên 70... |
| FiQA2018-VN | Các hình thức thay thế cho lương của nhân viên | <*Paragraph*> Có một vài sáng kiến tiền tệ địa phương ở danh sách Mỹ ở đây. Hầu hết là những nỗ lực để chuẩn bị một giá trị như một mức lương sống, hoặc khuyến khích mạng lưới tiêu thụ địa phương. Nếu bạn ở trong khu vực thu hút của một trong những điều này, hãy xem nếu bạn có thể có được một khoản trợ cấp hoặc vay để bắt đầu (nếu bạn sẵn sàng mua vào triết lý của nhóm như là một mức lương $10 tối thiểu) |
| HotpotQA-VN | Năm nào thì phim hoạt hình Barbie Thumbelina và Barbie and the Three Musketeers được phát hành? | <*Title*> Barbie Thumbelina <*Paragraph*> Barbie Thumbelina, hay còn gọi là B̈arbie Presents: Thumbelina,̈ là một bộ phim Barbie năm 2009 do Conrad Helten và Nishpeksh Mehra đạo diễn. Đây là tập thứ 15 trong loạt phim hoạt hình của Barbie, với sự lồng tiếng của Kelly Sheridan cho nhân vật chính Barbie. Tên gọi của câu chuyện giống như truyện cổ tích T̈humbelina(̈Cô bé ngón tay) của Hans Christian Andersen nhưng nội dung lại khác nhau. |
| MSMARCO-VN | chuyển oz sang gallon | <*Paragraph*> Có 0.007812500004244 gallon trong một ounce. Một Ounces bằng 0, 078125 Gallon. Định nghĩa của Ounces . Được biết đến với tên gọi là US fluid ounce, đơn vị thể tích cho các chất lỏng được sử dụng như ounce ở Mỹ và các nước khác thực hành hệ thống US Customary. |
| NFCorpus-VN | Chất béo bão hòa | <*Title*> LDL và HDL cholesterol và nồng độ LDL oxy hóa thay đổi ở người bình thường và tăng cholesterol sau khi sử dụng các mức khác nhau của canxi <*Paragraph*> Bột ca cao giàu polyphenols như catechin và procyanidins, đã được chứng minh trong nhiều nghiên cứu trên động vật về tác dụng ức chế LDL oxy hóa và tạo mảng xơ vữa. Nghiên cứu của chúng tôi đánh giá nồng độ LDL và LDL oxy hóa trong huyết thanh sau khi dùng các lượng khác nhau của bột ca cao (13, 19,5 và 26 g/ngày) ở những người bình thường và tăng nhẹ cholesterol. Trong nghiên cứu so sánh này... |
| NQ-VN | phim Silver Linings Playbook được quay ở đâu? | <*Title*> Silver Linings Playbook<*Paragraph*> Những địa điểm là Upper Darby, Ridley Park và Lansdowne, những cộng đồng nhỏ nằm ngay bên ngoài Philadelphia, Pennsylvania. Mặc dù không được nhắc tên trong phim, nhưng Ridley Park đã được ghi chú ở cuối, và một cảnh sát viên có thể được nhìn thấy đang đeo chữ viết tắt R̈PPDẗrên cổ áo của mình. |
| QuoraRetrieval-VN | Những ý tưởng kinh doanh tốt với mức đầu tư thấp ở Ấn Độ là gì? | <*Paragraph*> Những ý tưởng kinh doanh nhỏ là gì? |
| SCIDOCS-VN | Một Phương pháp hai bước để phân cụm dữ liệu hỗn hợp với các thể loại và số học | <*Title*> Forensics mạng WhatsApp: Giải mã và hiểu các thông điệp tín hiệu cuộc gọi WhatsApp <*Paragraph*> WhatsApp là một ứng dụng nhắn tin di động phổ biến với hơn 800 triệu người dùng. Gần đây, một tính năng gọi điện thoại đã được thêm vào ứng dụng và chưa có phân tích kỹ thuật số toàn diện nào được thực hiện về tính năng này vào thời điểm viết bài báo này. Trong tác phẩm này, chúng tôi mô tả cách chúng tôi có thể giải mã lưu lượng mạng và thu thập các bằng chứng pháp y liên quan đến tính năng gọi điện thoại mới này bao gồm: a) Số điện thoại WhatsApp, b) địa chỉ IP máy chủ WhatsApp, c) mã hóa âm thanh WhatsApp (Opus), d) thời gian gọi điện thoại WhatsApp và e) chấm dứt cuộc gọi điện thoại WhatsApp. Chúng tôi giải thích các phương pháp và công cụ sử dụng để giải mã lưu lượng truy cập cũng như trình bày chi tiết các phát hiện của chúng tôi liên quan đến các thông điệp điều khiển WhatsApp. Hơn nữa, chúng tôi cũng cung cấp cho cộng đồng một công cụ giúp hình dung các thông điệp giao thức WhatsApp. |
| SciFact-VN | Sự kích hoạt NFAT4 đòi hỏi sự di chuyển Ca2+ được trung gian bởi IP3R. | <*Title*> Điều khiển kích hoạt NFAT isoform và biểu hiện gen phụ thuộc NFAT thông qua hai tín hiệu Ca2+ trong tế bào trùng hợp và phân tách không gian <*Paragraph*> Sự kết hợp kích thích-chuyển tự, liên kết kích thích tại bề mặt tế bào với sự thay đổi biểu hiện gen nhân, được bảo tồn trong tất cả các sinh vật nhân thực. Làm thế nào các yếu tố chuyển tự đồng thời được biểu hiện có liên quan chặt chẽ vẫn chưa rõ ràng. Ở đây, chúng tôi cho thấy hai isoform yếu tố chuyển tự phụ thuộc canxi NFAT1 và NFAT4 đòi hỏi các tín hiệu InsP3 và Ca2+ phân biệt để kích hoạt bền vững về mặt sinh lý. ... |
| Touche2020-VN | Khuynh hướng tình dục có được xác định khi sinh ra? | <*Paragraph*> Khuynh hướng tình dục được xác định khi sinh ra. Làm thế nào? Bạn có thể dễ dàng nhìn thấy một em bé là nam hay nữ bằng cách nhìn bộ phận sinh dục của nó. Bộ phận sinh dục nam là dương vật và bộ phận sinh dục nữ là âm đạo. Đơn giản. |
| TRECCOVID-VN | Những chiếc mặt nạ nào là tốt nhất để phòng ngừa nhiễm Covid-19? | <*Title*> Sự lây lan của virus corona chủng mới (SARS-CoV-2): Mô hình hóa và mô phỏng các chiến lược kiểm soát <*Paragraph*> Bệnh dịch viêm đường hô hấp cấp do virus corona đang lan rộng khắp thế giới và tất cả các hệ thống y tế đều bị quá tải. Virus này được đặt tên là SARS-CoV-2. Trong tình hình này, cần phải đưa ra những quyết định hợp lý về cách chăm sóc bệnh nhân bị COVID-19. Báo cáo tỷ lệ mắc bệnh, các triệu chứng chung và các bộ dụng cụ thử nghiệm sẵn có, các chiến lược kiểm soát khác nhau, mô hình phân ngăn cơ bản và một số nghiên cứu hiện tại về dịch tễ học của bệnh được thảo luận và các mô hình đã công bố trước đó được xem xét. ... |

Table 7: Examples of queries and relevant documents for all datasets included in VN-MTEB. (<*Title*>) and (<*Paragraph*>) are used to distinguish the title separately from the paragraph within a document in the table above. These tokens were not passed to the respective models.

## H Model performance with size

Figure 7 represent an overview of model performance along with size and model type.

## I Model

Table 18 provides publicly available model checkpoints used for MTEB evaluation.

| Dataset | Text | Label |
|---|---|---|
| AmazonCounterfactualVNClassification | Quintus tiên tri rằng họ sẽ trở thành những vị tử đạo một ngày nào đó, nhưng không phải là ngày hôm đó. | not-counterfactual |
| AmazonPolarityVNClassification | Chúc mừng năm mớiPat yêu quý của tôi có một trong những giọng ca tuyệt vời nhất của thế hệ cô ấy. Tôi đã nghe đĩa CD này trong nhiều NĂM và tôi vẫn YÊU nó. Khi tôi có tâm trạng tốt, nó khiến tôi cảm thấy tốt hơn. Tâm trạng xấu chỉ tan biến như đường trong mưa. Đĩa CD này tràn đầy sự sống. Giọng ca thật tuyệt vời và lời bài hát thật tuyệt vời... | positive |
| AmazonReviewsVNClassification | Không xứng đáng với giá cả và thiết kế nắp rất tệ. Thiết kế nắp vô cùng kém. Không phù hợp để sử dụng hàng ngày. Nắp đậy quá chặt đến nỗi chúng ta phải vật lộn với chai mỗi ngày để mở nắp. Khi bế em bé trong một tay, việc mở nắp là một cơn ác mộng. Ngoài những tính năng siêu an toàn của nắp, chúng còn rất đắt so với các thương hiệu khác. Hãy tránh xa những sản phẩm này cho đến khi họ cải thiện những vấn đề về nắp. Chúng tôi đã nhiều lần làm tổn thương bản thân khi cố gắng mở nắp vì chúng có những cạnh sắc ở cả cạnh trong và ngoài. Không xứng đáng với giá cả. | 0 |
| Banking77VNClassification | Làm sao tôi có thể tìm thấy thẻ của mình | card_arrival |
| EmotionVNClassification | Tôi cảm thấy mình vẫn đang nhìn vào một tấm vải vẽ trống hoặc một tờ giấy trắng | sadness |
| ImdbVNClassification | Tôi yêu khoa học viễn tưởng và sẵn sàng chấp nhận nhiều điều. Phim/phim truyền hình khoa học viễn tưởng thường bị thiếu kinh phí, không được đánh giá cao và hiểu lầm. Tôi đã cố gắng thích điều này, tôi thực sự đã cố gắng, nhưng nó giống như so sánh phim truyền hình khoa học viễn tưởng tốt với Babylon 5 và Star Trek... | negative |
| MassiveIntentVNClassification | Hãy đánh thức tôi lúc 5 giờ sáng trong tuần này | alarm_set |
| MassiveScenarioVNClassification | Ai là người đang chơi bản nhạc này? | music |
| MTOPDomainVNClassification | Gọi Nicholas và Natasha | calling |
| MTOPIntentClassification | Tôi còn những nguyên liệu nào? | GET_INFO_RECIPES |
| ToxicConversationsVNClassification | Bingo: Mọi thứ luôn liên quan đến sự tăng trưởng dân số. Nếu chúng ta hạn chế nhập cư, chúng ta sẽ có mức tăng trưởng dân số xấp xỉ KHÔNG. Điều đó thật tuyệt vời cho chất lượng cuộc sống và môi trường! | not toxic |
| TweetSentimentExtractionVNClassification | Tôi rất thích bài hát Love Story của Taylor Swift | positive |

Table 8: Classification examples

| Dataset | Text | Cluster |
|---|---|---|
| RedditClustering-VN | Một người Úc đích thực là ai? | australia.txt |
| RedditClusteringP2P-VN | Những chiến thắng không được ghi lại chính xác Hôm nay tôi đã có 5 chiến thắng trong chế độ solo, nhưng hồ sơ của tôi lại hiển thị 0 chiến thắng ở chế độ solo và 5 chiến thắng ở LTM tôi có thể đảm bảo rằng tôi không chơi LTM và chưa bao giờ chơi chế độ này vì đây là tài khoản mới. Có ai gặp phải vấn đề này không? Tôi chơi trên PC. | FortNiteBR |
| StackExchangeClustering-VN | Thuật ngữ nào tốt hơn cho "front-end" và "back-end" của cơ sở dữ liệu dành cho người dùng phi kỹ thuật? | ux.stackexchange.com.txt |
| StackExchangeClusteringP2P-VN | Có ai có ví dụ về Dual Contouring trong C# không? Tôi đang cố gắng phát triển một phương pháp tạo địa hình sử dụng Perlin. Tôi đã theo dõi rất nhiều hướng dẫn của Minecraft và đã khiến chúng hoạt động. Tôi đã thử nghiệm với MarchingSquares, nhưng tôi không thích nó. Bây giờ, tôi đang cố gắng tạo ra một phương pháp dual contouring và tôi cũng đang cố gắng nắm bắt khái niệm về Octrees. Tôi từng phân đoạn mảng dữ liệu của mình thành những phần nhỏ, nhưng việc thu gọn và tạo một "phần" lớn hoạt động giống như một bộ phân đoạn nhỏ hơn không hiệu quả.Tôi hy vọng ai đó có thể chia sẻ một số mã C#, tốt nhất là dành cho Unity nhưng bất cứ điều gì để tôi có thể phân tích và hiểu cũng sẽ hữu ích. | unity |
| TwentyNewsgroupsClustering-VN | Windows 3.1 mới bán với giá $35 | 6 |

Table 9: Clustering examples

| Dataset | Sentence 1 | Sentence 2 | Label |
|---|---|---|---|
| SprintDuplicateQuestions-VN | Tại sao tôi không thể tìm ra cách dễ dàng nào để gửi một hình ảnh có văn bản trên Kyocera DuraCore của tôi? | Gửi hoặc nhận hình ảnh có văn bản Kyocera DuraCore | 1 |
| TwitterSemEval2015-VN | Kết thúc của phim 8 Mile là phần yêu thích nhất của bộ phim. | Đó chỉ là lời bài hát rap trong phim 8 Mile | 0 |
| TwitterURLCorpus-VN | Làm thế nào những ẩn dụ chúng ta sử dụng để miêu tả sự khám phá ảnh hưởng đến nam và nữ trong lĩnh vực khoa học | Những ý tưởng lớn đòi hỏi phải có những nỗ lực to lớn, và cách chúng ta nói về chúng cũng rất quan trọng. | 0 |

Table 10: Pair classification examples. Labels are binary.

| Dataset | Query | Positive | Negative |
|---|---|---|---|
| AskUbuntuDupQuestions-VN | không thể khởi động từ USB | USB cài Windows 7 không khởi động sau khi cài Ubuntu | không thể khởi động từ liveusb được tạo với pendrivelinux |
| SciDocsRR-VN | Lý thuyết Lãnh đạo phức tạp: Chuyển đổi phong cách lãnh đạo từ thời kỳ công nghiệp sang kỷ nguyên tri thức | Lý thuyết lãnh đạo phức tạp: Một quan điểm tương tác về lãnh đạo trong các hệ thống thích ứng phức tạp. | MedRec: Sử dụng Blockchain cho Truy cập Dữ liệu Y tế và Quản lý Quyền truy cập |
| StackOverflowDupQuestions-VN | Sử dụng numpy.genfromtxt để đọc một tệp csv với các chuỗi chứa dấu phẩy | numpy genfromtxtpandas đọc csv bỏ qua dấu phẩy ; trong dấu ngoặc kép | Lời bình luận của đối số genfromtxt trong numpy |

Table 11: Reranking examples

## J Detail Model Result

Table 20 and table 19 represent detail model result. We split into 2 tables, each for RoPE-based and other one is for APE-based.

| Dataset | Sentence 1 | Sentence 2 | Score |
|---|---|---|---|
| BIOSSES-VN | Mutations của gen KRAS gây ung thư là những đột biến phổ biến trong ung thư. | Đáng chú ý, c-Raf gần đây đã được phát hiện là yếu tố thiết yếu cho sự phát triển của NSCLC do K-Ras gây ra. | 1.8 |
| SICK-R-VN | Một người đàn ông đang ở trong một bãi đậu xe và đang chơi quần vợt với một bức tường lớn. | Người trượt tuyết đang nhảy qua tuyết trắng một cách can đảm | 1.0 |
| STSBenchmark-VN | Người phát ngôn của vận động viên: Các cáo buộc sử dụng doping dường như là không có căn cứ. | Tin tức mới nhất về thời tiết khắc nghiệt: 1 người chết ở Texas sau cơn lốc xoáy | 0.0 |

Table 12: STS examples. Scores are continuous between 0 and 5 (included).

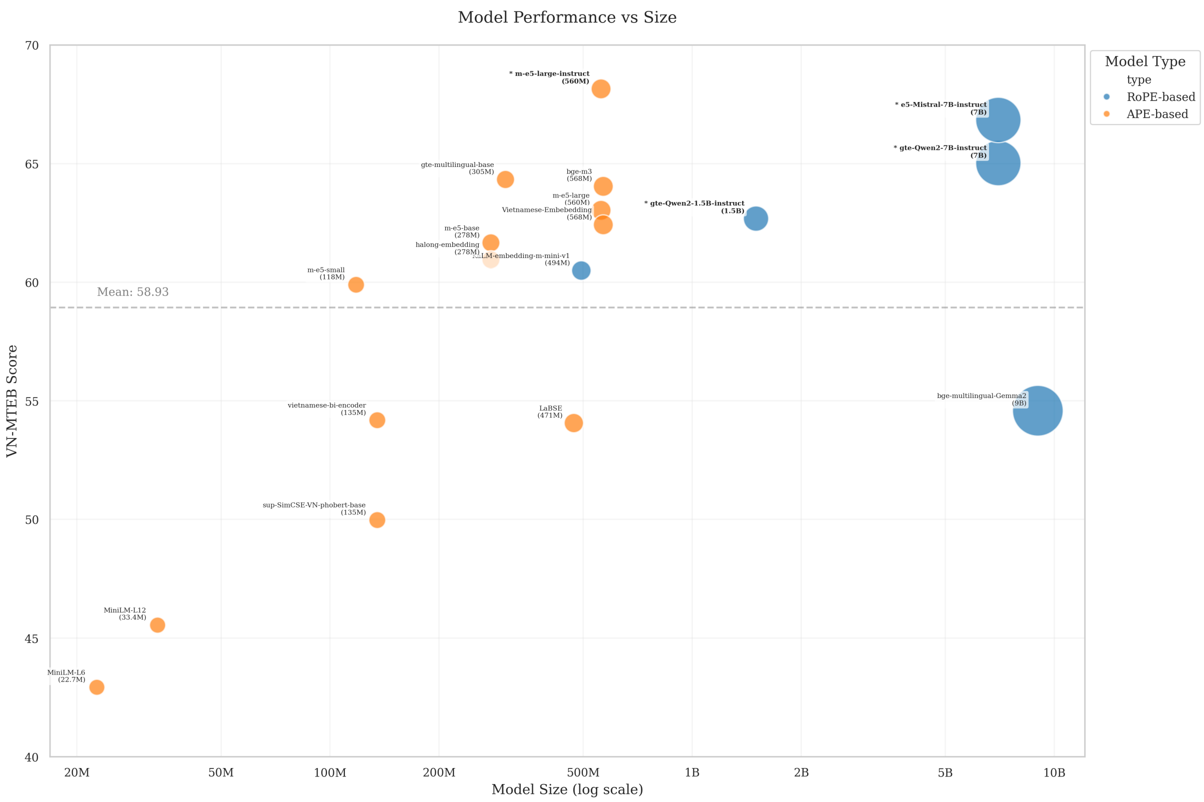


Figure 7: Model performance and size.

| Name | Type | Train Samples | Dev Samples | Test Samples |
|---|---|---|---|---|
| AmazonCounterfactualVNClassification | Classification | 0 | 0 | 466 |
| AmazonPolarityVNClassification | Classification | 0 | 0 | 344,197 |
| AmazonReviewsVNClassificat,ion | Classification | 0 | 0 | 3,424 |
| Banking77VNClassification | Classification | 0 | 0 | 2,378 |
| EmotionVNClassification | Classification | 0 | 0 | 1,290 |
| ImdbVNClassification | Classification | 0 | 0 | 22,081 |
| MassiveIntentVNClassification | Classification | 0 | 0 | 1784 |
| MassiveScenarioVNClassification | Classification | 0 | 0 | 2974 |
| MTOPDomainVNClassification | Classification | 0 | 0 | 13,291 |
| MTOPIntentVNClassification | Classification | 0 | 0 | 13,291 |
| ToxicConversationsVNClassification | Classification | 0 | 0 | 38,560 |
| TweetSentimentExtractionVNClassification | Classification | 0 | 0 | 2,065 |
| RedditClustering-VN | Clustering | 0 | 0 | 293,904 |
| RedditClusteringP2P-VN | Clustering | 0 | 0 | 346,846 |
| StackExchangeClustering-VN | Clustering | 0 | 0 | 251,974 |
| StackExchangeClusteringP2P-VN | Clustering | 0 | 0 | 66,150 |
| TwentyNewsgroupsClustering-VN | Clustering | 0 | 0 | 35,089 |
| SprintDuplicateQuestions-VN | PairClassification | 0 | 0 | 88,173 |
| TwitterSemEval2015-VN | PairClassification | 0 | 0 | 9,378 |
| TwitterURLCorpus-VN | PairClassification | 0 | 0 | 30,095 |
| AskUbuntuDupQuestions-VN | Reranking | 0 | 0 | 1,833 |
| SciDocsRR-VN | Reranking | 0 | 0 | 6,526 |
| StackOverflowDupQuestions-VN | Reranking | 0 | 0 | 2,808 |
| ArguAna-VN | Retrieval | 0 | 0 | 6,969 |
| ClimateFEVER-VN | Retrieval | 0 | 0 | 5,419,992 |
| CQADupstackAndroidRetrieval-VN | Retrieval | 0 | 0 | 24,505 |
| CQADupstackGisRetrieval-VN | Retrieval | 0 | 0 | 38,466 |
| CQADupstackMathematicaRetrieval-VN | Retrieval | 0 | 0 | 17,472 |
| CQADupstackPhysicsRetrieval-VN | Retrieval | 0 | 0 | 39,314 |
| CQADupstackProgrammersRetrieval-VN | Retrieval | 0 | 0 | 33,267 |
| CQADupstackStatsRetrieval-VN | Retrieval | 0 | 0 | 42,693 |
| CQADupstackTexRetrieval-VN | Retrieval | 0 | 0 | 71,313 |
| CQADupstackUnixRetrieval-VN | Retrieval | 0 | 0 | 38,666 |
| CQADupstackWebmastersRetrieval-VN | Retrieval | 0 | 0 | 18,597 |
| CQADupstackWordpressRetrieval-VN | Retrieval | 0 | 0 | 49151 |
| DBPedia-VN | Retrieval | 0 | 0 | 4,540,903 |
| FEVER-VN | Retrieval | 0 | 0 | 5,422,820 |
| FiQA2018-VN | Retrieval | 0 | 0 | 58,659 |
| HotpotQA-VN | Retrieval | 0 | 0 | 5,245,971 |
| MSMARCO-VN | Retrieval | 0 | 0 | 8,846,142 |
| NFCorpus-VN | Retrieval | 0 | 0 | 10,437 |
| NQ-VN | Retrieval | 0 | 0 | 2,683,751 |
| QuoraRetrieval-VN | Retrieval | 0 | 0 | 534,403 |
| SCIDOCS-VN | Retrieval | 0 | 0 | 37,626 |
| SciFact-VN | Retrieval | 0 | 0 | 5,338 |
| Touche2020-VN | Retrieval | 0 | 0 | 383,683 |
| TRECCOVID-VN | Retrieval | 0 | 0 | 228,690 |
| BIOSSES-VN | STS | 0 | 0 | 100 |
| SICK-R-VN | STS | 0 | 0 | 9927 |
| STSBenchmark-VN | STS | 0 | 0 | 1379 |

Table 13: Tasks in VN-MTEB. Dataset already formatted and compatible with MTEB code

| **Data Name** | **Domain** | **Subtask** | **Task** |
|---|---|---|---|
| arguana-vn | **[Medical, Written]** | | Retrieval |
| touche2020-vn | **[Academic]** | Question answering | Retrieval |
| fever-vn | **[Encyclopaedic, Written]** | Claim verification | Retrieval |
| climate-fever-vn | **[Encyclopaedic, Written]** | Claim verification | Retrieval |
| scifact-vn | **[Academic, Medical, Written]** | | Retrieval |
| scidocs-vn | **[Academic, Written, Non-fiction]** | | Retrieval |
| dbpedia-entity-vn | **[Written, Encyclopaedic]** | | Retrieval |
| cqadupstack-*-vn | **[Written, Non-fiction]** | Question answering, Duplicate Detection | Retrieval |
| quora-vn | **[Written, Web, Blog]** | Question answering | Retrieval |
| nq-vn | **[Written, Encyclopaedic]** | Question answering | Retrieval |
| hotpotqa-vn | **[Web, Written]** | Question answering | Retrieval |
| fiqa-vn | **[Written, Financial]** | Question answering | Retrieval |
| trec-covid-vn | **[Medical, Academic, Written]** | | Retrieval |
| nfcorpus-vn | **[Medical, Academic, Written]** | | Retrieval |
| msmarco-vn | **[Encyclopaedic, Academic, Blog, News, Medical, Government, Reviews, Non-fiction, Social, Web]** | Question answering | Retrieval |
| EmotionVNClassification-VN | **[Social, Written]** | Sentiment/Hate speech | Classification |
| Banking77Classification-VN | [Written] | | Classification |
| ToxicConversationsClassification-VN | **[Social, Written]** | Sentiment/Hate speech | Classification |
| ImdbVNClassification-VN | [Reviews, Written] | Sentiment/Hate speech | Classification |
| TweetSentimentExtractionClassification-VN | **[Social, Written]** | Sentiment/Hate speech | Classification |
| AmazonCounterfactualClassification-VN | [Reviews, Written] | Counterfactual Detection | Classification |
| MTOPDomainClassification-VN | [Spoken] | | Classification |
| MTOPIntentClassification-VN | [Spoken] | | Classification |
| AmazonReviewsClassification-VN | [Reviews, Written] | | Classification |
| MassiveIntentClassification-VN | [Spoken] | | Classification |
| MassiveScenarioClassification-VN | [Spoken] | | Classification |
| AmazonPolarityClassification-VN | [Reviews, Written] | Sentiment/Hate speech | Classification |
| SprintDuplicateQuestions-VN | [Programming, Written] | Duplicate Detection | Pair-Classification |
| TwitterSemEval2015-VN | **[Social, Written]** | | Pair-Classification |
| TwitterURLCorpus-VN | **[Social, Written]** | | Pair-Classification |
| TwentyNewsgroupsClustering-VN | [News, Written] | Thematic clustering | Clustering |
| RedditClustering-VN | **[Web, Social, Written]** | Thematic clustering | Clustering |
| RedditClusteringP2P-VN | **[Web, Social, Written]** | Thematic clustering | Clustering |
| StackExchangeClustering-VN | [Web, Written] | Thematic clustering | Clustering |
| StackExchangeClusteringP2P-VN | [Web, Written] | Thematic clustering | Clustering |
| AskUbuntuDupQuestions-VN | [Programming, Web] | | Rerank |
| StackOverflowDupQuestions-VN | [Written, Blog, Programming] | Question answering | Rerank |
| SciDocsRR-VN | **[Academic, Non-fiction, Written]** | Scientific Reranking | Rerank |
| STSBenchmark-VN | [Blog, News, Written] | | Semantic Textual Similarity |
| BIOSSES-VN | **[Medical]** | | Semantic Textual Similarity |
| SICK-R-VN | [Web, Written] | Textual Entailment | Semantic Textual Similarity |

Table 14: Tasks in VN-MTEB. There are 6 task types and 41 datasets.

| **Data Name** | **Domain** | **Subtask** | **Task** |
|---|---|---|---|
| BelebeleRetrieval | [Web, News, Written] | Question answering | Retrieval |
| MLQARetrieval | [Encyclopaedic, Written] | Question answering | Retrieval |
| XQuADRetrieval | [Web, Written] | Question answering | Retrieval |
| WebFAQRetrieval | [Web, Written] | Question answering | Retrieval |
| PublicHealthQARetrieval | [Medical, Government, Web, Written] | Question answering | Retrieval |
| BibleNLPBitextMining | [Religious, Written] | | Bibtext Mining |
| FloresBitextMining | [Non-fiction, Encyclopaedic, Written] | | Bibtext Mining |
| NTREXBitextMining | [News, Written] | | Bibtext Mining |
| TatoebaBitextMining | [Written] | | Bibtext Mining |
| WebFAQBitextMiningQuestions | [Web, Written] | | Bibtext Mining |
| LanguageClassification | [Reviews, Web, Non-fiction, Fiction, Government, Written] | Language identification | Classification |
| MultilingualSentimentClassification | [Reviews, Written] | Sentiment/Hate speech | Classification |
| MassiveIntentClassification | [Spoken] | | Classification |
| MassiveScenarioClassification | [Spoken] | | Classification |
| SIB200Classification | [News, Written] | | Classification |
| VieStudentFeedbackClassification | [Reviews, Written] | Sentiment/Hate speech | Classification |
| XNLI | [Non-fiction, Fiction, Government, Written] | | Pair-Classification |
| SIB200ClusteringFast | [News, Written] | | Clustering |

Table 15: Tasks in MMTEB. There are 5 task types and 18 datasets

| Dataset | Type | Public Link | Translated Link | License |
|---|---|---|---|---|
| AmazonCounterfactualClassification | Classification | https://huggingface.co/datasets/mteb/amazon_counterfactual | - | cc-by-4.0 |
| AmazonPolarityClassification | Classification | https://huggingface.co/datasets/mteb/amazon_polarity | | apache-2.0 |
| AmazonReviewsClassification | Classification | https://huggingface.co/datasets/mteb/amazon_reviews_multi | - | - |
| Banking77Classification | Classification | https://huggingface.co/datasets/mteb/banking77 | - | mit |
| EmotionClassification | Classification | https://huggingface.co/datasets/mteb/emotion | - | - |
| ImdbClassification | Classification | https://huggingface.co/datasets/mteb/imdb | - | - |
| MassiveIntentClassification | Classification | https://huggingface.co/datasets/mteb/amazon_massive_intent | - | apache-2.0 |
| MassiveScenarioClassification | Classification | https://huggingface.co/datasets/mteb/amazon_massive_scenario | - | apache-2.0 |
| MTOPDomainClassification | Classification | https://huggingface.co/datasets/mteb/mtop_domain | - | - |
| MTOPIntentClassification | Classification | https://huggingface.co/datasets/mteb/mtop_intent | - | - |
| ToxicConversationsClassification | Classification | https://huggingface.co/datasets/mteb/toxic_conversations_50k | - | cc-by-4.0 |
| TweetSentimentExtractionClassification | Classification | https://huggingface.co/datasets/mteb/tweet_sentiment_extraction | - | - |
| RedditClustering | Clustering | https://huggingface.co/datasets/mteb/reddit-clustering | - | - |
| RedditClusteringP2P | Clustering | https://huggingface.co/datasets/mteb/reddit-clustering-p2p | - | - |
| StackExchangeClustering | Clustering | https://huggingface.co/datasets/mteb/stackexchange-clustering | - | - |
| StackExchangeClusteringP2P | Clustering | https://huggingface.co/datasets/mteb/stackexchange-clustering-p2p | - | - |
| TwentyNewsgroupsClustering | Clustering | https://huggingface.co/datasets/mteb/twentynewsgroups-clustering | - | - |
| SprintDuplicateQuestions | Pair-Classification | https://huggingface.co/datasets/mteb/sprintduplicatequestions-pairclassification | - | - |
| TwitterSemEval2015 | Pair-Classification | https://huggingface.co/datasets/mteb/twittersemeval2015-pairclassification | - | - |
| TwitterURLCorpus | Pair-Classification | https://huggingface.co/datasets/mteb/twitterurlcorpus-pairclassification | - | - |
| AskUbuntuDupQuestions | Reranking | https://huggingface.co/datasets/mteb/askubuntudupquestions-reranking | - | - |
| SciDocsRR | Reranking | https://huggingface.co/datasets/mteb/SciDocsRR | - | cc-by-4.0 |
| StackOverflowDupQuestions | Reranking | https://huggingface.co/datasets/mteb/stackoverflowdupquestions-reranking | - | - |
| ArguAna | Retrieval | https://huggingface.co/datasets/mteb/arguana | - | cc-by-4.0 |
| ClimateFEVER | Retrieval | https://huggingface.co/datasets/mteb/climate-fever | - | cc-by-4.0 |
| CQADupstackAndroid | Retrieval | https://huggingface.co/datasets/mteb/cqadupstack-android | - | apache-2.0 |
| CQADupstackGis | Retrieval | https://huggingface.co/datasets/mteb/cqadupstack-gis | - | apache-2.0 |
| CQADupstackMathematica | Retrieval | https://huggingface.co/datasets/mteb/cqadupstack-mathematica | - | apache-2.0 |
| CQADupstackPhysics | Retrieval | https://huggingface.co/datasets/mteb/cqadupstack-physics | - | apache-2.0 |
| CQADupstackProgrammers | Retrieval | https://huggingface.co/datasets/mteb/cqadupstack-programmers | - | apache-2.0 |
| CQADupstackStats | Retrieval | https://huggingface.co/datasets/mteb/cqadupstack-stats | - | apache-2.0 |
| CQADupstackTex | Retrieval | https://huggingface.co/datasets/mteb/cqadupstack-tex | - | apache-2.0 |
| CQADupstackUnix | Retrieval | https://huggingface.co/datasets/mteb/cqadupstack-unix | - | apache-2.0 |
| CQADupstackWebmasters | Retrieval | https://huggingface.co/datasets/mteb/cqadupstack-webmasters | - | apache-2.0 |
| CQADupstackWordpress | Retrieval | https://huggingface.co/datasets/mteb/cqadupstack-wordpress | - | apache-2.0 |
| DBPedia | Retrieval | https://huggingface.co/datasets/mteb/dbpedia | - | mit |
| FEVER | Retrieval | https://huggingface.co/datasets/mteb/fever | - | cc-by-sa-3.0 |
| FiQA2018 | Retrieval | https://huggingface.co/datasets/mteb/fiqa | - | cc-by-sa-4.0 |
| HotpotQA | Retrieval | https://huggingface.co/datasets/mteb/hotpotqa | - | cc-by-sa-4.0 |
| MSMARCO | Retrieval | https://huggingface.co/datasets/mteb/msmarco | - | cc-by-sa-4.0 |
| NFCorpus | Retrieval | https://huggingface.co/datasets/mteb/nfcorpus | - | cc-by-sa-4.0 |
| NQ | Retrieval | https://huggingface.co/datasets/mteb/nq | - | cc-by-nc-sa-3.0 |
| Quora | Retrieval | https://huggingface.co/datasets/mteb/quora | - | cc-by-sa-4.0 |
| SCIDOCS | Retrieval | https://huggingface.co/datasets/mteb/scidocs | - | cc-by-sa-4.0 |
| SciFact | Retrieval | https://huggingface.co/datasets/mteb/scifact | - | cc-by-sa-4.0 |
| Touche2020 | Retrieval | https://huggingface.co/datasets/mteb/touche2020 | - | cc-by-sa-4.0 |
| TRECCOVID | Retrieval | https://huggingface.co/datasets/mteb/trec-covid | - | cc-by-sa-4.0 |
| BIOSSES | STS | https://huggingface.co/datasets/mteb/biosses-sts | - | - |
| SICK-R | STS | https://huggingface.co/datasets/mteb/sickr-sts | - | cc-by-nc-sa-3.0 |
| STSBenchmark | STS | https://huggingface.co/datasets/mteb/stsbenchmark-sts | - | - |

Table 16: Dataset licenses for MTEB and VN-MTEB

| Name | Type | Total Number of tokens | Time Estimated (s) | GPU Electricity Consumption (kWh) |
|---|---|---|---|---|
| AmazonCounterfactualVNClassification | Classification | 910,364 | 239.57 | 0.186 |
| AmazonPolarityVNClassification | Classification | 536,435,795 | 141167.31 | 109.797 |
| AmazonReviewsVNClassification | Classification | 82,306,198 | 21659.53 | 16.846 |
| Banking77VNClassification | Classification | 241,685 | 63.60 | 0.049 |
| EmotionVNClassification | Classification | 595,593 | 156.74 | 0.122 |
| ImdbVNClassification | Classification | 18,074,863 | 4756.54 | 3.700 |
| MassiveIntentVNClassification | Classification | 13,809,421 | 3634.06 | 2.826 |
| MassiveScenarioVNClassification | Classification | 13,802,417 | 3632.22 | 2.825 |
| MTOPDomainVNClassification | Classification | 1,439,620 | 378.85 | 0.295 |
| MTOPIntentVNClassification | Classification | 1,439,620 | 378.85 | 0.295 |
| ToxicConversationsVNClassification | Classification | 9,332,763 | 2455.99 | 1.910 |
| TweetSentimentExtractionVNClassification | Classification | 1,011,699 | 266.24 | 0.207 |
| RedditClustering-VN | Clustering | 12,694,431 | 3340.64 | 2.598 |
| RedditClusteringP2P-VN | Clustering | 108,712,751 | 28608.62 | 22.251 |
| StackExchangeClustering-VN | Clustering | 17,157,163 | 4515.04 | 3.512 |
| StackExchangeClusteringP2P-VN | Clustering | 25,618,672 | 6741.76 | 5.244 |
| TwentyNewsgroupsClustering-VN | Clustering | 1,655,500 | 435.66 | 0.339 |
| SprintDuplicateQuestions-VN | PairClassification | 4,711,640 | 1239.91 | 0.964 |
| TwitterSemEval2015-VN | PairClassification | 665,973 | 175.26 | 0.136 |
| TwitterURLCorpus-VN | PairClassification | 3,004,908 | 790.77 | 0.615 |
| AskUbuntuDupQuestions-VN | Reranking | 136,142 | 35.83 | 0.028 |
| SciDocsRR-VN | Reranking | 7,620,209 | 2005.32 | 1.560 |
| StackOverflowDupQuestions-VN | Reranking | 12,324,554 | 3243.30 | 2.523 |
| ArguAna-VN | Retrieval | 2,842,260 | 747.96 | 0.582 |
| ClimateFEVER-VN | Retrieval | 681,973,189 | 179466.63 | 139.585 |
| CQADupstackAndroidRetrieval-VN | Retrieval | 3,902,043 | 1026.85 | 0.799 |
| CQADupstackGisRetrieval-VN | Retrieval | 10,313,933 | 2714.19 | 2.111 |
| CQADupstackMathematicaRetrieval-VN | Retrieval | 6,109,244 | 1607.70 | 1.250 |
| CQADupstackPhysicsRetrieval-VN | Retrieval | 6,224,273 | 1637.97 | 1.274 |
| CQADupstackProgrammersRetrieval-VN | Retrieval | 8,800,245 | 2315.85 | 1.801 |
| CQADupstackStatsRetrieval-VN | Retrieval | 13,178,147 | 3467.93 | 2.697 |
| CQADupstackTexRetrieval-VN | Retrieval | 25,201,127 | 6631.88 | 5.158 |
| CQADupstackUnixRetrieval-VN | Retrieval | 13,401,968 | 3526.83 | 2.743 |
| CQADupstackWebmastersRetrieval-VN | Retrieval | 3,483,317 | 916.66 | 0.713 |
| CQADupstackWordpressRetrieval-VN | Retrieval | 14,241,887 | 3747.86 | 2.915 |
| DBPedia-VN | Retrieval | 414,726,629 | 109138.59 | 84.886 |
| FEVER-VN | Retrieval | 683,783,334 | 179942.98 | 139.956 |
| FiQA2018-VN | Retrieval | 12,536,252 | 3299.01 | 2.566 |
| HotpotQA-VN | Retrieval | 442,305,098 | 116396.08 | 90.530 |
| MSMARCO-VN | Retrieval | 778,538,066 | 204878.44 | 159.350 |
| NFCorpus-VN | Retrieval | 1,642,900 | 432.34 | 0.336 |
| NQ-VN | Retrieval | 370,480,772 | 97494.94 | 75.829 |
| QuoraRetrieval-VN | Retrieval | 19,285,282 | 5075.07 | 3.947 |
| SCIDOCS-VN | Retrieval | 7,936,076 | 2088.44 | 1.624 |
| SciFact-VN | Retrieval | 2,200,704 | 579.13 | 0.450 |
| Touche2020-VN | Retrieval | 170,315,421 | 44819.85 | 34.860 |
| TRECCOVID-VN | Retrieval | 52,994,734 | 13945.98 | 10.847 |
| BIOSSES-VN | STS | 9,357 | 2.46 | 0.002 |
| SICK-R-VN | STS | 269,368 | 70.89 | 0.055 |
| STSBenchmark-VN | STS | 332,610 | 87.53 | 0.068 |
| Total | Total | 4,620,730,217 | 1215981.64 | 946.066 |

Table 17: GPU Usage to Translate datasets in VN-MTEB

| Model | Public Checkpoint |
|---|---|
| gte-Qwen2-7B-instruct | https://huggingface.co/Alibaba-NLP/gte-Qwen2-7B-instruct |
| e5-Mistral-7B-instruct | https://huggingface.co/intfloat/e5-mistral-7b-instruct |
| bge-multilingual-Gemma2 | https://huggingface.co/BAAI/bge-multilingual-gemma2 |
| gte-Qwen2-1.5B-instruct | https://huggingface.co/Alibaba-NLP/gte-Qwen2-1.5B-instruct |
| m-e5-large-instruct | https://huggingface.co/intfloat/multilingual-e5-large-instruct |
| m-e5-large | https://huggingface.co/intfloat/multilingual-e5-large |
| bge-me | https://huggingface.co/BAAI/bge-m3 |
| Vietnamese-Embedding | https://huggingface.co/AITeamVN/Vietnamese_Embedding |
| KaLM-embedding-m-mini-v1 | https://huggingface.co/HIT-TMG/KaLM-embedding-multilingual-mini-v1 |
| LaBSE | https://huggingface.co/sentence-transformers/LaBSE |
| gte-multilingual-base | https://huggingface.co/Alibaba-NLP/gte-multilingual-base |
| m-e5-base | https://huggingface.co/intfloat/multilingual-e5-base |
| halong-embedding | https://huggingface.co/hiieu/halong_embedding |
| m-e5-small | https://huggingface.co/intfloat/multilingual-e5-small |
| vietnamese-bi-encoder | https://huggingface.co/bkai-foundation-models/vietnamese-bi-encoder |
| sup-SimCSE-VN-phobert-base | https://huggingface.co/VoVanPhuc/sup-SimCSE-VietNamese-phobert-base |
| MiniLM-L12 | https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 |
| MiniLM-L6 | https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L6-v2 |

Table 18: Publicly available model links used for evaluation

| Dataset | gte-Qwen2-7B-instruct | e5-Mistral-7B-instruct | bge-multilingual-Gemma2 | gte-Qwen2-1.5B-instruct | KaLM-mini |
|---|---|---|---|---|---|
| AmazonCounterfactualVNClassification | 66.7 | 68.8 | 68.78 | 64.48 | 62.36 |
| AmazonPolarityVNClassification | 90.89 | 93.8 | 84.14 | 82.0 | 75.84 |
| AmazonReviewsVNClassification | 43.23 | 49.94 | 42.03 | 38.71 | 40.05 |
| Banking77VNClassification | 83.04 | 83.86 | 83.88 | 81.88 | 71.63 |
| EmotionVNClassification | 46.19 | 44.8 | 50.23 | 45.16 | 43.13 |
| ImdbVNClassification | 86.63 | 88.09 | 81.51 | 70.43 | 73.12 |
| MassiveIntentVNClassification | 74.34 | 75.8 | 72.59 | 72.37 | 63.55 |
| MassiveScenarioVNClassification | 78.28 | 78.74 | 76.48 | 75.88 | 67.37 |
| MTOPDomainVNClassification | 89.62 | 88.43 | 91.66 | 86.99 | 81.04 |
| MTOPIntentVNClassification | 70.43 | 68.7 | 75.72 | 66.48 | 53.63 |
| ToxicConversationsVNClassification | 61.22 | 62.35 | 73.19 | 60.74 | 62.49 |
| TweetSentimentExtractionVNClassification | 58.52 | 63.27 | 61.13 | 60.56 | 59.85 |
| RedditClustering-VN | 49.7 | 45.78 | 29.91 | 46.76 | 45.37 |
| RedditClusteringP2P-VN | 64.06 | 59.34 | 56.5 | 56.65 | 60.68 |
| StackExchangeClustering-VN | 65.05 | 62.72 | 48.83 | 58.9 | 55.67 |
| StackExchangeClusteringP2P-VN | 40.67 | 43.8 | 32.99 | 33.42 | 33.37 |
| TwentyNewsgroupsClustering-VN | 46.27 | 46.9 | 32.42 | 42.46 | 39.16 |
| SprintDuplicateQuestions-VN | 75.07 | 91.78 | 66.68 | 85.03 | 90.6 |
| TwitterSemEval2015-VN | 58.68 | 73.32 | 53.76 | 52.44 | 63.65 |
| TwitterURLCorpus-VN | 82.52 | 86.92 | 80.49 | 80.64 | 85.58 |
| AskUbuntuDupQuestions-VN | 77.03 | 78.17 | 68.05 | 73.01 | 70.93 |
| SciDocsRR-VN | 93.62 | 93.32 | 83.93 | 92.18 | 90.12 |
| StackOverflowDupQuestions-VN | 52.2 | 53.96 | 40.63 | 48.91 | 45.48 |
| ArguAna-VN | 52.77 | 50.36 | 50.61 | 51.99 | 52.66 |
| ClimateFEVER-VN | 21.49 | 24.77 | 16.52 | 23.47 | 7.81 |
| CQADupstackAndroid-VN | 48.36 | 46.82 | 34.54 | 42.33 | 43.3 |
| CQADupstackGis-VN | 36.06 | 35.18 | 15.15 | 28.13 | 29.8 |
| CQADupstackMathematica-VN | 29.41 | 25.26 | 12.22 | 24.46 | 20.73 |
| CQADupstackPhysics-VN | 48.15 | 38.17 | 24.0 | 37.18 | 36.64 |
| CQADupstackProgrammers-VN | 38.86 | 40.42 | 19.15 | 35.66 | 33.66 |
| CQADupstackStats-VN | 34.59 | 29.55 | 10.96 | 26.77 | 26.69 |
| CQADupstackTex-VN | 26.74 | 28.1 | 8.66 | 23.75 | 23.29 |
| CQADupstackUnix-VN | 39.26 | 39.94 | 20.01 | 33.88 | 32.97 |
| CQADupstackWebmasters-VN | 38.71 | 38.59 | 20.35 | 32.3 | 32.5 |
| CQADupstackWordpress-VN | 31.14 | 31.62 | 11.45 | 25.34 | 23.55 |
| DBPedia-VN | 41.89 | 42.78 | 6.96 | 39.51 | 28.61 |
| FEVER-VN | 82.81 | 84.82 | 45.23 | 83.53 | 60.61 |
| FiQA2018-VN | 46.92 | 30.39 | 11.76 | 34.27 | 29.45 |
| HotpotQA-VN | 67.99 | 64.54 | 29.72 | 61.86 | 60.81 |
| MSMARCO-VN | 68.99 | 35.24 | 10.3 | 66.49 | 28.31 |
| NFCorpus-VN | 38.27 | 31.98 | 10.25 | 33.21 | 29.76 |
| NQ-VN | 59.91 | 57.8 | 9.71 | 54.89 | 34.42 |
| Quora-VN | 52.23 | 42.87 | 21.3 | 52.11 | 52.14 |
| SCIDOCS-VN | 20.95 | 15.23 | 8.12 | 18.04 | 13.83 |
| SciFact-VN | 73.8 | 63.77 | 45.29 | 69.67 | 58.74 |
| Touche2020-VN | 28.64 | 25.92 | 11.05 | 30.99 | 22.17 |
| TRECCOVID-VN | 77.3 | 77.42 | 39.2 | 78.46 | 59.33 |
| BIOSSES-VN | 82.09 | 83.72 | 66.85 | 80.8 | 83.52 |
| SICK-R-VN | 76.32 | 77.91 | 66.5 | 78.07 | 74.49 |
| STSBenchmark-VN | 77.79 | 81.98 | 64.97 | 81.03 | 77.6 |

Table 19: All Vietnamese results on RoPE based model. The main score for each task is reported as described in Original MTEB Paper (Muennighoff et al., 2023).

| Dataset | m-e5-large-instruct | m-e5-large | bge-m3 | Vietnamese-Emb | LaBSE | gte-multilingual-base | m-e5-base | halong | m-e5-small | vietnamese-bi | sup-SimCSE-VN | MiniLM-L12 | MiniLM-L6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AmazonCounterfactualVNClassification | 67.7 | 70.39 | 69.4 | 71.44 | 71.61 | 66.24 | 66.09 | 65.6 | 63.07 | 61.37 | 67.96 | 64.7 | 64.59 |
| AmazonPolarityVNClassification | 95.05 | 76.42 | 87.54 | 88.78 | 70.39 | 80.06 | 75.91 | 69.99 | 74.86 | 66.52 | 79.05 | 55.4 | 56.19 |
| AmazonReviewsVNClassification | 49.8 | 39.68 | 44.33 | 44.48 | 36.37 | 42.36 | 40.31 | 36.36 | 38.55 | 32.79 | 37.69 | 27.22 | 26.99 |
| Banking77VNClassification | 83.84 | 73.74 | 78.1 | 79.21 | 67.11 | 74.71 | 70.96 | 75.02 | 67.14 | 75.51 | 69.05 | 50.8 | 48.94 |
| EmotionVNClassification | 49.64 | 46.81 | 49.93 | 48.95 | 40.91 | 44.31 | 45.24 | 46.85 | 40.5 | 34.57 | 36.69 | 20.14 | 20.96 |
| ImdbVNClassification | 91.92 | 72.68 | 82.71 | 83.06 | 62.59 | 75.31 | 68.51 | 65.2 | 66.77 | 59.01 | 69.93 | 51.53 | 54.24 |
| MassiveIntentVNClassification | 74.38 | 65.73 | 68.18 | 67.74 | 60.59 | 64.96 | 63.02 | 65.4 | 60.06 | 62.29 | 57.94 | 42.39 | 41.47 |
| MassiveScenarioVNClassification | 77.62 | 68.32 | 72.75 | 72.85 | 64.1 | 69.37 | 67.24 | 70.88 | 64.38 | 65.36 | 60.76 | 47.84 | 45.9 |
| MTOPDomainVNClassification | 87.74 | 84.75 | 86.56 | 85.54 | 79.72 | 82.82 | 83.98 | 84.29 | 79.35 | 79.35 | 70.58 | 58.9 | 56.41 |
| MTOPIntentVNClassification | 71.92 | 57.33 | 57.01 | 58.01 | 53.23 | 50.94 | 52.01 | 53.99 | 45.5 | 55.84 | 48.21 | 30.43 | 29.93 |
| ToxicConversationsVNClassification | 67.03 | 64.22 | 69.08 | 66.27 | 65.05 | 68.67 | 65.67 | 66.59 | 63.34 | 62.94 | 61.46 | 55.56 | 54.75 |
| TweetSentimentExtractionVNClassification | 64.02 | 60.25 | 63.49 | 62.34 | 59.53 | 60.11 | 60.56 | 59.77 | 59.69 | 51.45 | 56.92 | 41.89 | 41.89 |
| RedditClustering-VN | 49.06 | 42.12 | 43.25 | 43.89 | 28.29 | 49.91 | 42.6 | 38.31 | 37.74 | 28.6 | 29.08 | 17.97 | 13.23 |
| RedditClusteringP2P-VN | 61.47 | 60.64 | 57.38 | 56.16 | 50.09 | 59.75 | 58.34 | 55.67 | 56.39 | 43.82 | 43.66 | 31.45 | 27.61 |
| StackExchangeClustering-VN | 63.63 | 56.39 | 58.42 | 57.24 | 38.58 | 60.8 | 57.15 | 55.26 | 54.88 | 44.31 | 38.63 | 21.91 | 16.62 |
| StackExchangeClusteringP2P-VN | 41.26 | 32.07 | 32.63 | 31.49 | 27.87 | 35.23 | 32.22 | 31.94 | 32.5 | 27.8 | 27.66 | 29.49 | 24.89 |
| TwentyNewsgroupsClustering-VN | 49.4 | 37.69 | 37.83 | 39.29 | 28.11 | 45.55 | 38.2 | 35.92 | 34.31 | 26.12 | 26.2 | 20.98 | 19.66 |
| SprintDuplicateQuestions-VN | 90.27 | 93.58 | 96.54 | 95.28 | 82.55 | 97.08 | 93.16 | 95.23 | 91.42 | 89.14 | 76.27 | 80.74 | 69.59 |
| TwitterSemEval2015-VN | 76.1 | 71.78 | 70.99 | 68.24 | 65.26 | 70.21 | 68.76 | 64.02 | 67.47 | 61.05 | 57.67 | 50.2 | 52.33 |
| TwitterURLCorpus-VN | 87.03 | 85.75 | 85.77 | 84.99 | 84.91 | 85.99 | 85.62 | 84.36 | 84.66 | 82.02 | 79.99 | 77.45 | 76.46 |
| AskUbuntuDupQuestions-VN | 75.09 | 71.39 | 72.73 | 72.97 | 67.88 | 73.23 | 69.99 | 70.51 | 68.5 | 68.0 | 62.43 | 66.57 | 65.84 |
| SciDocsRR-VN | 93.34 | 91.1 | 90.01 | 88.77 | 84.72 | 91.83 | 89.52 | 88.91 | 88.2 | 83.01 | 78.88 | 75.74 | 74.62 |
| StackOverflowDupQuestions-VN | 51.41 | 48.72 | 51.09 | 50.94 | 44.33 | 50.29 | 47.69 | 50.08 | 46.36 | 43.83 | 35.28 | 39.01 | 37.91 |
| ArguAna-VN | 48.15 | 47.88 | 50.68 | 51.07 | 36.46 | 52.75 | 45.49 | 52.48 | 42.97 | 38.08 | 26.35 | 9.89 | 9.72 |
| ClimateFEVER-VN | 25.01 | 15.43 | 21.27 | 13.25 | 2.41 | 21.05 | 12.62 | 14.48 | 15.13 | 11.14 | 7.0 | 1.63 | 0.4 |
| CQADupstackAndroid-VN | 43.13 | 42.28 | 44.04 | 41.93 | 28.47 | 39.66 | 42.35 | 42.12 | 41.69 | 26.73 | 16.56 | 20.84 | 17.25 |
| CQADupstackGis-VN | 30.73 | 31.28 | 33.13 | 31.91 | 17.24 | 29.12 | 28.61 | 30.76 | 29.12 | 17.8 | 8.28 | 13.8 | 9.19 |
| CQADupstackMathematica-VN | 22.31 | 24.06 | 23.64 | 21.44 | 12.85 | 20.8 | 21.33 | 21.85 | 19.33 | 13.19 | 4.54 | 9.72 | 6.62 |
| CQADupstackPhysics-VN | 35.7 | 36.53 | 37.99 | 35.52 | 21.19 | 39.08 | 35.15 | 36.89 | 36.96 | 26.19 | 14.16 | 14.54 | 10.19 |
| CQADupstackProgrammers-VN | 36.74 | 34.53 | 34.12 | 32.71 | 18.51 | 34.19 | 31.9 | 32.85 | 31.42 | 20.42 | 10.74 | 14.72 | 7.77 |
| CQADupstackStats-VN | 26.19 | 27.81 | 30.12 | 26.86 | 15.08 | 27.79 | 25.81 | 28.57 | 26.51 | 18.64 | 7.3 | 14.8 | 7.48 |
| CQADupstackTex-VN | 23.4 | 22.68 | 26.11 | 24.78 | 12.73 | 21.37 | 20.78 | 23.97 | 22.08 | 10.99 | 5.59 | 10.53 | 6.07 |
| CQADupstackUnix-VN | 32.98 | 33.62 | 35.67 | 34.52 | 22.5 | 30.61 | 32.94 | 32.65 | 31.12 | 19.48 | 8.82 | 16.44 | 11.78 |
| CQADupstackWebmasters-VN | 33.85 | 33.07 | 34.47 | 31.67 | 20.78 | 28.51 | 31.04 | 32.6 | 30.58 | 21.39 | 11.41 | 16.52 | 9.56 |
| CQADupstackWordpress-VN | 25.3 | 25.56 | 28.18 | 24.74 | 14.05 | 23.38 | 23.87 | 24.78 | 23.39 | 16.21 | 6.45 | 13.23 | 8.54 |
| DBPedia-VN | 39.9 | 31.58 | 36.7 | 34.2 | 15.92 | 37.46 | 30.77 | 23.8 | 28.54 | 20.22 | 11.16 | 14.81 | 10.9 |
| FEVER-VN | 83.34 | 58.3 | 70.14 | 48.81 | 12.58 | 86.24 | 49.6 | 52.87 | 54.25 | 53.11 | 11.89 | 29.4 | 12.35 |
| FiQA2018-VN | 36.46 | 31.51 | 34.38 | 29.94 | 7.38 | 32.88 | 25.14 | 26.23 | 22.71 | 17.29 | 6.62 | 4.31 | 1.44 |
| HotpotQA-VN | 63.99 | 65.11 | 63.76 | 70.07 | 17.0 | 58.6 | 60.79 | 53.36 | 54.99 | 34.48 | 13.65 | 17.16 | 13.31 |
| MSMARCO-VN | 37.86 | 39.08 | 36.22 | 30.5 | 10.12 | 35.16 | 36.19 | 29.75 | 33.11 | 30.12 | 4.99 | 9.41 | 8.16 |
| NFCorpus-VN | 33.4 | 31.5 | 30.96 | 25.39 | 20.5 | 31.48 | 26.75 | 27.03 | 27.28 | 23.38 | 15.82 | 17.32 | 14.05 |
| NQ-VN | 56.86 | 52.32 | 54.98 | 42.61 | 11.7 | 50.65 | 45.1 | 36.15 | 38.54 | 30.63 | 7.08 | 10.79 | 7.44 |
| Quora-VN | 57.9 | 66.49 | 64.57 | 61.0 | 38.17 | 56.68 | 63.29 | 58.79 | 60.47 | 37.55 | 32.33 | 26.55 | 20.43 |
| SCIDOCS-VN | 16.81 | 13.74 | 15.01 | 13.03 | 8.36 | 14.49 | 12.9 | 13.35 | 11.71 | 9.18 | 4.93 | 5.39 | 3.61 |
| SciFact-VN | 65.52 | 68.5 | 62.31 | 55.12 | 41.49 | 65.62 | 67.61 | 60.87 | 65.78 | 39.29 | 19.67 | 26.75 | 20.93 |
| Touche2020-VN | 25.03 | 16.01 | 21.53 | 11.98 | 3.93 | 22.69 | 13.13 | 15.88 | 17.74 | 18.8 | 12.15 | 2.68 | 2.66 |
| TRECCOVID-VN | 80.56 | 54.71 | 66.22 | 27.32 | 16.96 | 60.82 | 44.86 | 54.73 | 53.57 | 54.57 | 21.23 | 18.11 | 11.75 |
| BIOSSES-VN | 84.26 | 81.69 | 77.5 | 78.14 | 76.77 | 84.45 | 81.82 | 80.2 | 79.08 | 66.13 | 55.13 | 64.14 | 56.09 |
| SICK-R-VN | 80.17 | 78.22 | 77.88 | 77.11 | 68.77 | 77.5 | 76.77 | 74.0 | 75.49 | 69.65 | 74.46 | 61.92 | 62.05 |
| STSBenchmark | 84.38 | 82.03 | 81.15 | 77.19 | 70.59 | 82.58 | 79.77 | 77.96 | 78.09 | 69.97 | 76.24 | 60.94 | 56.62 |

Table 20: All Vietnamese results on APE based model. The main score for each task is reported as described in Original MTEB paper (Muennighoff et al., 2023).