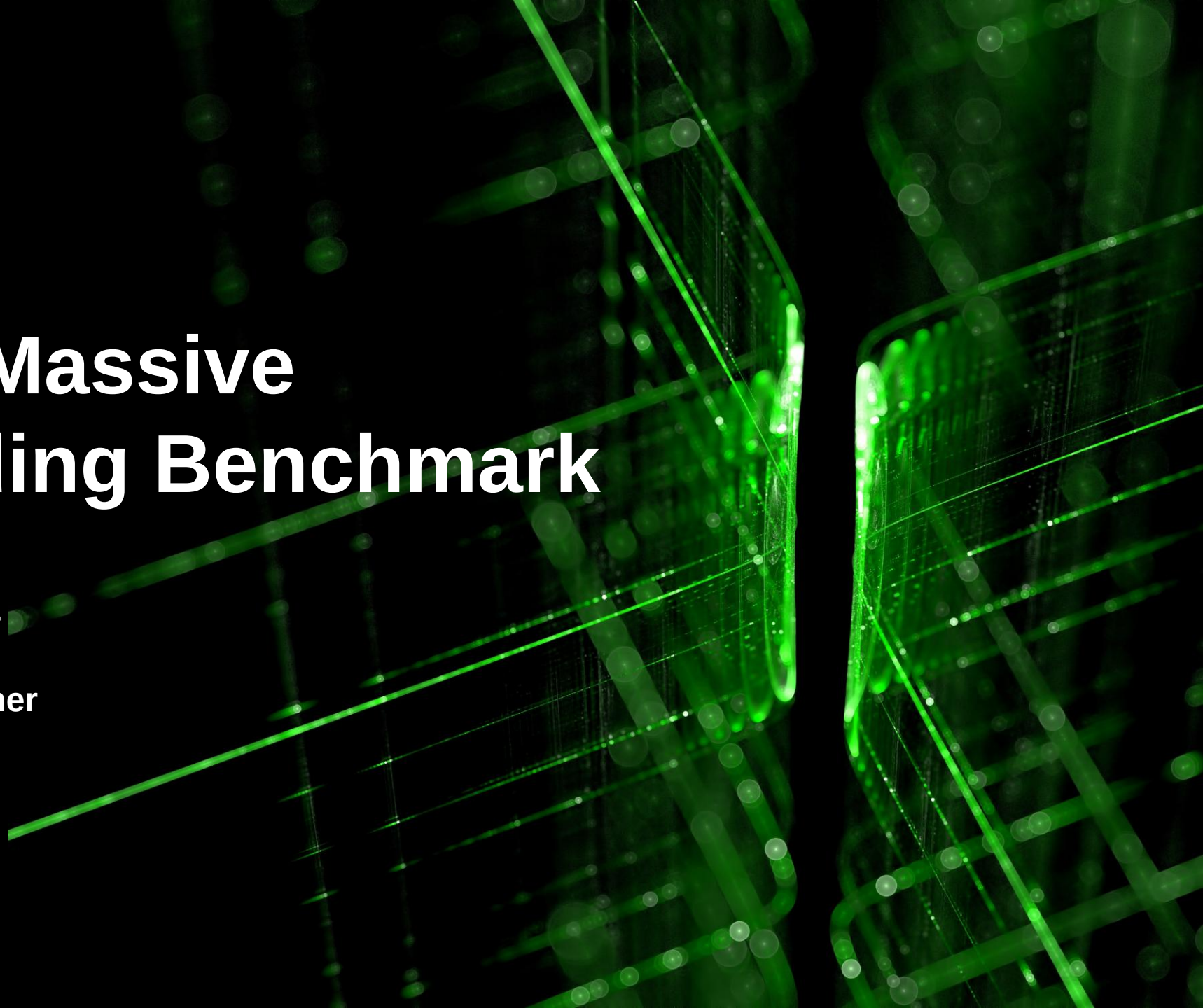


VN-MTEB ↘
Vietnamese Massive Text Embedding Benchmark

Presented by

Pham Bao Loc – AI Researcher
locpb@greennode.ai
Hoang Quoc Viet – AI Researcher
viethq5@greennode.ai
Luu Quy Tung– AI Researcher
tunglq@greennode.ai

Table of Content

Motivation

Proposed Method

VN-MTEB

Benchmark Result & Conclusion

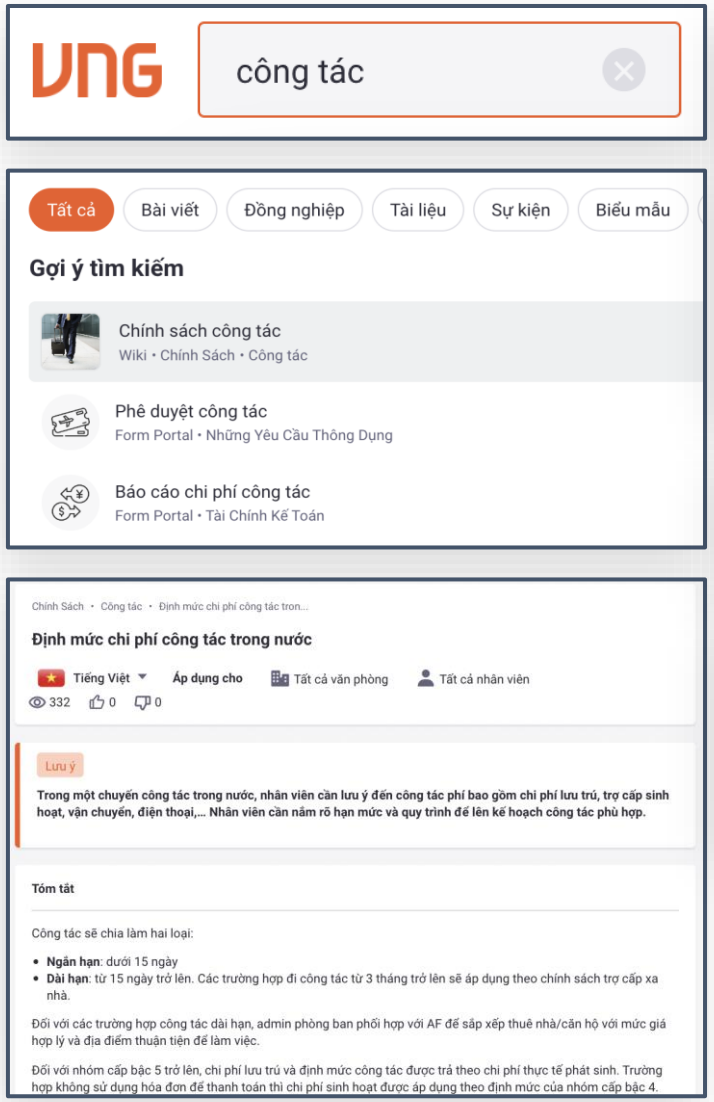
Motivation

AI Chatbot using RAG

Phòng họp TSE ở đâu?

Phòng họp TSE nằm ở VNG Campus, Building 1, Floor 1, Green Zone. Để đến phòng họp TSE, bạn có thể đi từ The Coffee House, đi thẳng đến khu vực IT Helpdesk, rẽ phải rồi đi thẳng, phòng họp TSE nằm bên phía tay trái. Tuy nhiên, phòng họp TSE không hỗ trợ Video Conference qua ZOOM.

Source: Hướng Dẫn - Phòng họp.docx



MTEB

BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, Iryna Gurevych
Ubiquitous Knowledge Processing Lab (UKP Lab)
Department of Computer Science, Technical University of Darmstadt
http://www.ukp.tu-darmstadt.de

Abstract

Existing neural information retrieval (IR) models have often been studied in homogeneous and narrow settings, which has considerably limited insights into their out-of-distribution (OOD) generalization capabilities. To address this, and to facilitate researchers to broadly evaluate the effectiveness of their models, we introduce **Be**nchmarking-**IR** (**BEIR**), a robust and heterogeneous evaluation benchmark for information retrieval. We leverage a careful selection of 18 publicly available datasets from diverse text retrieval tasks and domains and evaluate 10 state-of-the-art retrieval systems including lexical, sparse, dense, late-interaction and re-ranking architectures on the BEIR benchmark. Our results show BM25 is a robust baseline and re-ranking and late-interaction based models on average achieve the best zero-shot performances, however, at high computational costs. In contrast, dense and sparse-retrieval models are computationally more efficient but often underperform other approaches, highlighting the considerable room for improvement in their generalization capabilities. We hope this framework allows us to better evaluate and understand existing retrieval systems, and contributes to accelerating progress towards more robust and generalizable systems in the future. BEIR is publicly available at https://github.com/UKPLab/beir.

1 Introduction

Major natural language processing (NLP) problems rely on a practical and efficient retrieval component as a first step to find relevant information. Challenging problems include open-domain question-answering [8], claim-verification [58], duplicate question detection [77], and many more. Traditionally, retrieval has been dominated by lexical approaches like TF-IDF or BM25 [53]. However, these approaches suffer from lexical gap [5] and are able to only retrieve documents containing keywords present within the query. Further, lexical approaches treat queries and documents as bag-of-words by not taking word ordering into consideration.

Recently, deep learning and in particular pre-trained Transformer models like BERT [12] have become popular in information retrieval [75]. These neural retrieval systems can be used in many

MTEB: Massive Text Embedding Benchmark

Niklas Muennighoff[1], Nouamane Tazi[1], Loïc Magne[1], Nils Reimers[2]*
[1]Hugging Face [2]cohere.ai
[1]firstname@huggingface.co [2]info@nils-reimers.de

Abstract

Text embeddings are commonly evaluated on a small set of datasets from a single task not covering their possible applications to other tasks. It is unclear whether state-of-the-art embeddings on semantic textual similarity (STS) can be equally well applied to other tasks like clustering or reranking. This makes progress in the field difficult to track, as various models are constantly being proposed without proper evaluation. To solve this problem, we introduce the Massive Text Embedding Benchmark (MTEB). MTEB spans 8 embedding tasks covering a total of 58 datasets and 112 languages. Through the benchmarking of 33 models on MTEB, we establish the most comprehensive benchmark of text embeddings to date. We find that no particular text embedding method dominates across all tasks. This suggests that the field has yet to converge on a universal text embedding method and scale it up sufficiently to provide state-of-the-art results on all embedding tasks. MTEB comes with open-source code and a public leaderboard at https://github.com/embeddings-benchmark/mteb.

Gurevych, 2019) solely evaluate on STS and classification tasks, leaving open questions about the transferability of the embedding models to search or clustering tasks. STS is known to poorly correlate with other real-world use cases (Neelakantan et al., 2022; Wang et al., 2021). Further, evaluating embedding methods on many tasks requires implementing multiple evaluation pipelines. Implementation details like pre-processing or hyperparameters may influence the results making it unclear whether performance improvements simply come from a favorable evaluation pipeline. This leads to the "blind" application of these models to new use cases in industry or requires incremental work to reevaluate them on different tasks.

The Massive Text Embedding Benchmark (MTEB) aims to provide clarity on how models perform on a variety of embedding tasks and thus serves as the gateway to finding universal text embeddings applicable to a variety of tasks. MTEB consists of 58 datasets covering 112 languages from 8 embedding tasks: Bitext mining, classification, clustering, pair classification, reranking,

Motivation

Spaces | mteb/leaderboard | like 5.72k | Running on CPU UPGRADE | App | Files | Community 167

Select Benchmark

- Multilingual
- English
- Image
- Regional
  - European
  - Indic
  - Scandinavian
- Domain-Specific
- Language-specific
  - Chinese
  - German
  - French
  - Japanese
  - Korean
  - Polish
  - Russian
  - Farsi
- Miscellaneous
- Legacy

Embedding Leaderboard

This leaderboard compares 100+ text and image embedding models across 1000+ languages. We refer to the publication of each selectable benchmark for details on metrics, languages, tasks, and task types. Anyone is welcome to add a model, add benchmarks, help us improve zero-shot annotations or propose other changes to the leaderboard.

MTEB(Multilingual, v2)

A large-scale multilingual expansion of MTEB, driven mainly by highly-curated community contributions covering 250+ languages.

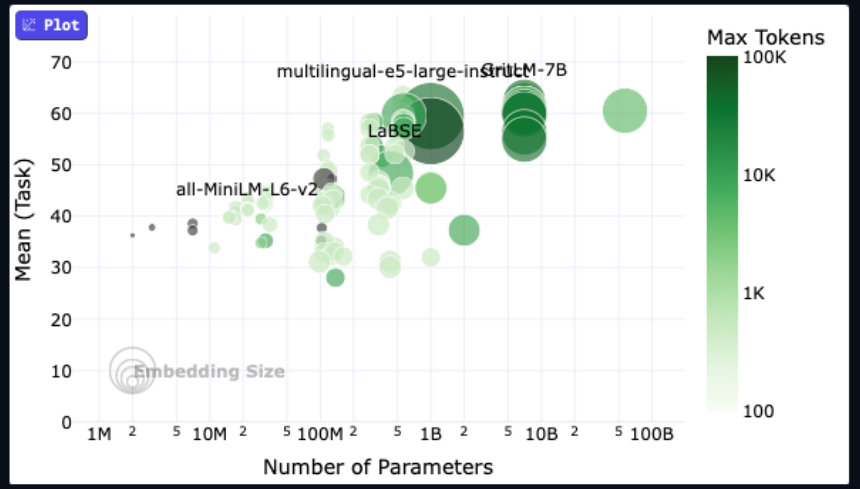
- Number of languages: 1038
- Number of tasks: 131
- Number of task types: 9
- Number of domains: 20

Click for More Info

Cite this benchmark:

Share this benchmark:

Performance per Model Size | Performance per Task Type (Radar Chart)



We only display models that have been run on all tasks in the benchmark

Customize this Benchmark

Advanced Model Filters

Summary | Performance per task | Task information

Filter...

Rank (Bor…	Model	Zero-shot	Memory U…	Number of P…	Embedding D…	Max Tokens	Mean (T…	Mean (TaskT…	Bitext …
1	gemini-embedding-001	99%	Unknown	Unknown	3072	2048	**68.37**	**59.59**	79.28
2	Linq-Embed-Mistral	99%	13563	7B	4096	32768	61.47	54.14	70.34
3	gte-Qwen2-7B-instruct	⚠ NA	29040	7B	3584	32768	62.51	55.93	73.92
4	multilingual-e5-large-instruct	99%	1068	560M	1024	514	63.22	55.08	**80.13**
5	SFR-Embedding-Mistral	96%	13563	7B	4096	32768	60.90	53.92	70.00
6	GritLM-7B	99%	13813	7B	4096	4096	60.92	53.74	70.53
	text-multilingual-								

GREENNODE

Motivation

Language-specific

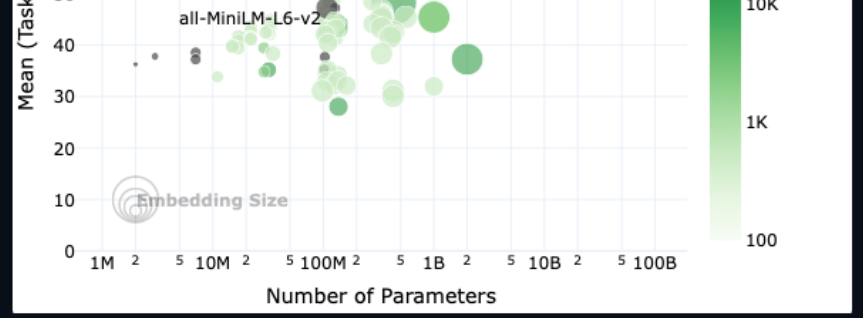
- Chinese
- German
- French
- Japanese
- Korean
- Polish
- Russian
- Farsi

Miscellaneous

Legacy

Cite this benchmark:

Share this benchmark:



We only display models that have been run on all tasks in the benchmark

Customize this Benchmark

Advanced Model Filters

Summary | Performance per task | Task information

Filter...

Rank (Bor…	Model	Zero-shot	Memory U…	Number of P…	Embedding D…	Max Tokens	Mean (T…	Mean (TaskT…	Bitext …
1	gemini-embedding-001	99%	Unknown	Unknown	3072	2048	**68.37**	**59.59**	79.28
2	Linq-Embed-Mistral	99%	13563	7B	4096	32768	61.47	54.14	70.34
3	gte-Qwen2-7B-instruct	⚠ NA	29040	7B	3584	32768	62.51	55.93	73.92
4	multilingual-e5-large-instruct	99%	1068	560M	1024	514	63.22	55.08	**80.13**
5	SFR-Embedding-Mistral	96%	13563	7B	4096	32768	60.90	53.92	70.00
6	GritLM-7B	99%	13813	7B	4096	4096	60.92	53.74	70.53
7	text-multilingual-embedding-002	99%	Unknown	Unknown	768	2048	62.16	54.25	70.73
8	GritLM-8x7B	99%	89079	57B	4096	4096	60.49	53.31	68.17
9	e5-mistral-7b-instruct	99%	13563	7B	4096	32768	60.25	53.08	70.58
10	Cohere-embed-multilingual-v3.0	⚠ NA	Unknown	Unknown	1024	Unknown	61.12	53.23	70.50

Download Table

Frequently Asked Questions

Acknowledgment: We thank Google, ServiceNow, Contextual AI and Hugging Face for their generous sponsorship. If you'd like to sponsor us, please get in touch.

We also thank the following companies which provide API credits to evaluate their models: OpenAI, Voyage AI

GREENNODE

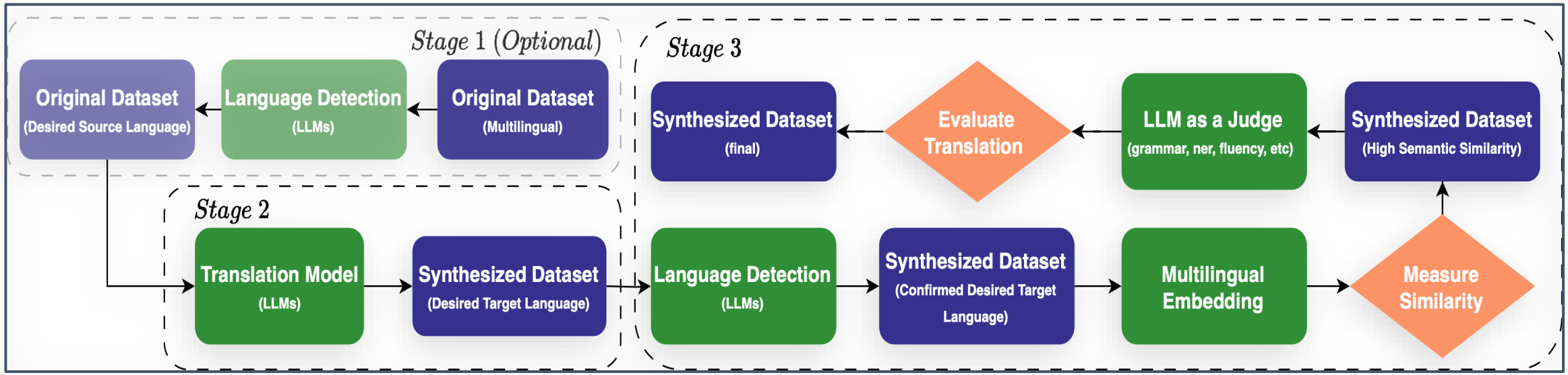
Why is this research important?

- Saving Effort
- Advance research progress
- Advance product development

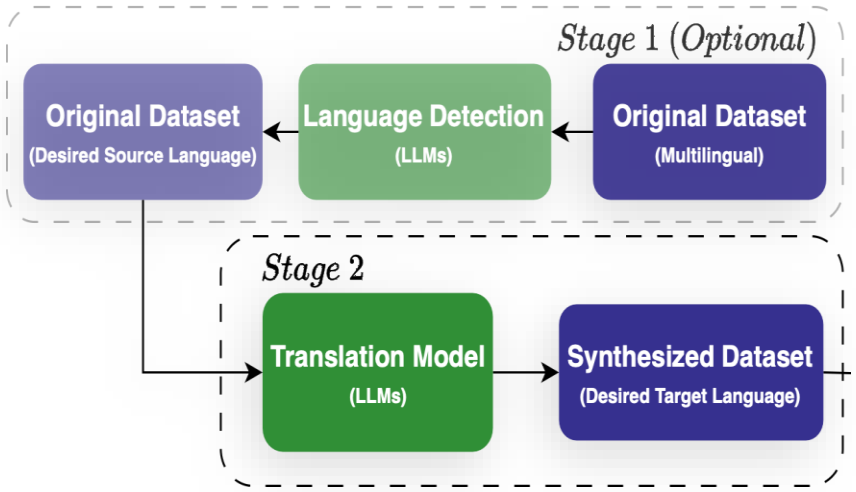
```python
import mteb
from sentence_transformers import SentenceTransformer

# Define the sentence-transformers model name
model_name = "average_word_embeddings_komninos"

model = mteb.get_model(model_name) # if the model is not implemented in MTEB it will be eq. to SentenceTransformer(model_name)
tasks = mteb.get_tasks(tasks=["VN-MTEB"])
evaluation = mteb.MTEB(tasks=tasks)
results = evaluation.run(model, output_folder=f"results/{model_name}")
```

```
mteb available_tasks # list _all_ available tasks

mteb run -m sentence-transformers/all-MiniLM-L6-v2 \
    -t VN-MTEB  \
    --verbosity 3

# if nothing is specified default to saving the results in the results/{model_name} folder
```

Our propose method



Our propose method



In translation model:

- We tried many Translation models, LLMs and picked **Aya-23-35B** as the main Translation model.

SEA-HELM Leaderboard

SEA-HELM (SouthEast Asian Holistic Evaluation of Language Models) is an assessment of large language models across various tasks, emphasizing Southeast Asian languages. The leaderboard evaluates four key multilingual capabilities for instruction-tuned models: performance on the Open LLM Leaderboard v2 in English, proficiency in Southeast Asian chat, instruction-following in Southeast Asian languages, and Southeast Asian linguistic tasks. For base models, it assesses performance on the Open LLM Leaderboard v2 in English and proficiency in Southeast Asian linguistic tasks.

Hosted by the AI Products Team @ AI Singapore

Instruction-Tuned/Reasoning Models (<10B) | Instruction-Tuned/Reasoning Models (>10B) | Base Models (<10B) | Base Models (>10B) | Comparison

Overall | Delta | Delta (%) | Performance Plot

Search: Search by model name'

Model Size: - | 70B | 671B | 27B | 12B | 72B | 32B | 21B | 123B | 14B | 83B | 24B | 35B | 20B | 22B

Select Columns to Show: Model | Organisation | Model Variant | Architecture | Type | Supported SEA Languages | MMLU-PRO (EN) | MUSR (EN) | IFEval (EN) | SEA Average | ID | VI | TH | TA | TL

Model Alias	Size	Organisation	MMLU-PRO (EN)	MUSR (EN)	IFEval (EN)	SEA Average	ID	VI	TH	TA	TL
Mistral Small (2409)	22B	Mistral AI	49.96	45.74	62.66	38.07	57.01	40.95	39.2	7.55	45.61
Sailor2	20B	Sea AI Lab	46.9	28.8	34.75	39.24	55.81	41.06	39.58	14.27	45.48
Mistral Nemo	12B	Mistral AI	41.48	35.24	56.38	35.17	42.23	46.44	36.33	17.35	33.5
Llama 3	70B	Meta	57	51.65	78	54.45	57.43	53.55	50.36	52.16	58.76
Babel	83B	Alibaba DAMO	56.92	44.34	41.59	50.7	62.32	55.03	36.62	48.87	50.65
Mistral Large (2407)	123B	Mistral AI	67.5	63.78	78.74	56.54	65.35	57.31	52.96	43.73	63.36
R1 Distill (Qwen)	14B	DeepSeek	62.82	65.47	73.94	44.34	62.62	57.36	55.74	3.97	41.99
Aya 23	35B	Cohere For AI	28.9	24.3	57.86	42.72	54.59	58.53	28.23	33.93	38.35
Mistral Small (3.1)	24B	Mistral AI	48.48	56.32	76.16	48.88	63.48	59.92	47.64	11.11	62.26
Reka Flash 3	21B	RekaAI	-	-	-	59.89	62.68	60.63	57.27	59.83	59.04
Mistral Large (2411)	123B	Mistral AI	67.92	63.92	82.62	59.87	68.11	60.76	53.97	50.32	66.22

Quick Definitions

SEA Average: Covers **Filipino, Indonesian, Tamil, Thai and Vietnamese** across various tasks such as Sentiment, Toxicity, Translation, QA, to Summarization.

Our propose method

Stage 3: Quality control

We measure between the Original (English) vs the Translated (Vietnamese)

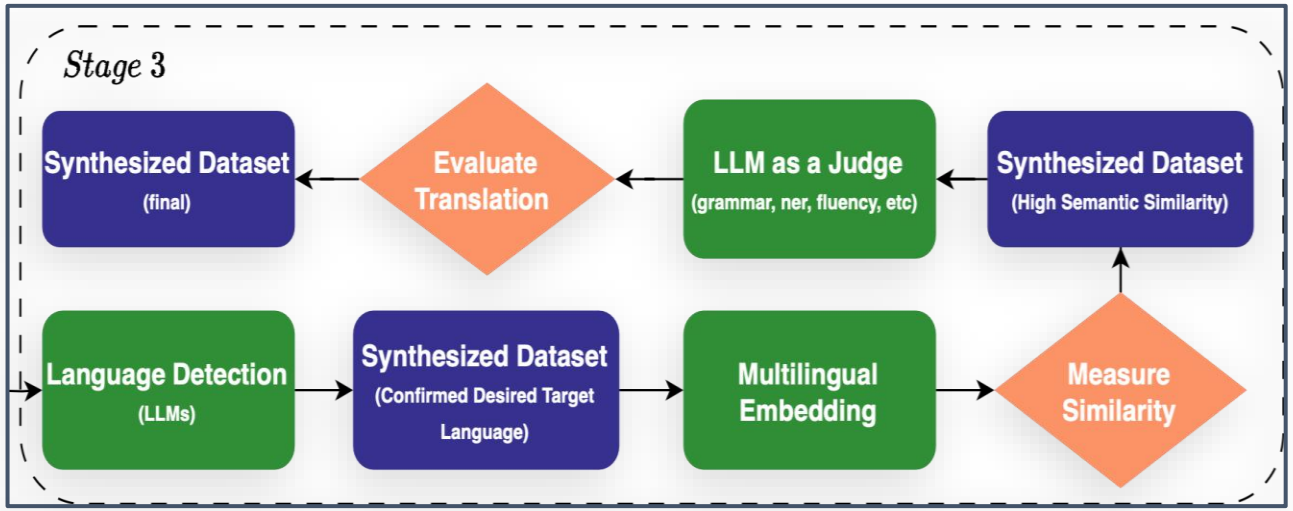
Measure cosine similarity

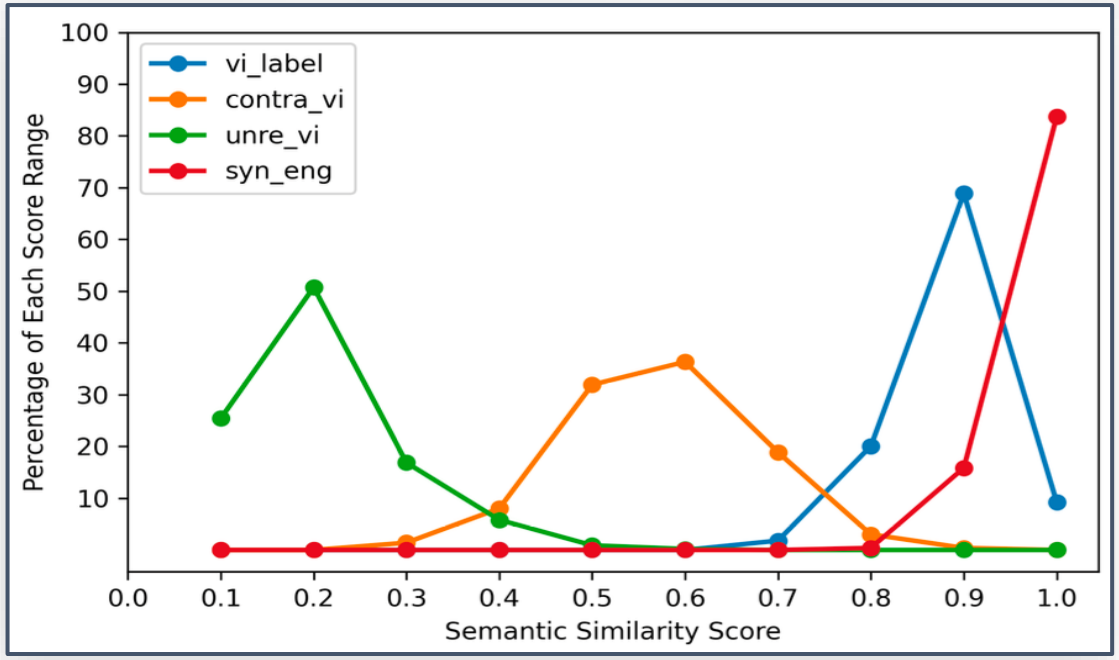
Using **gte-Qwen2-7B-instruct:**

- Top performing model on MTEB multilingual.
- Long context (Up to 32k token)

Calculate the cosine similarity of the Original vs the Translated

Filter out samples where cosine score < **0.8**



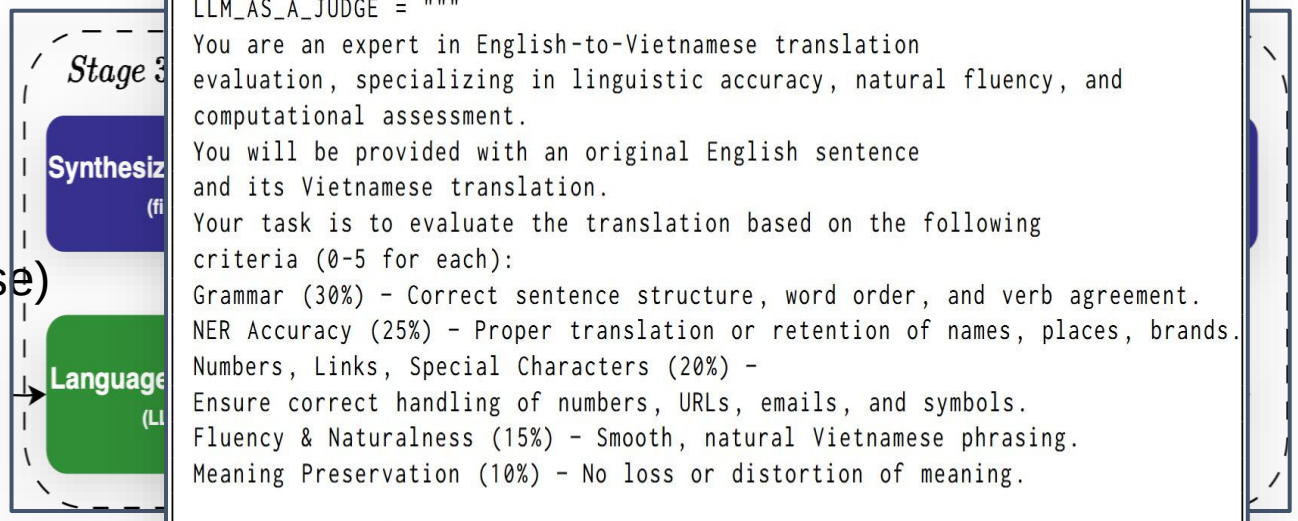


Our propose method

Stage 3: Quality control

We measure between
the Original (English) vs the Translated (Vietnamese)

LLM as a Judge




```
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:
"""
```

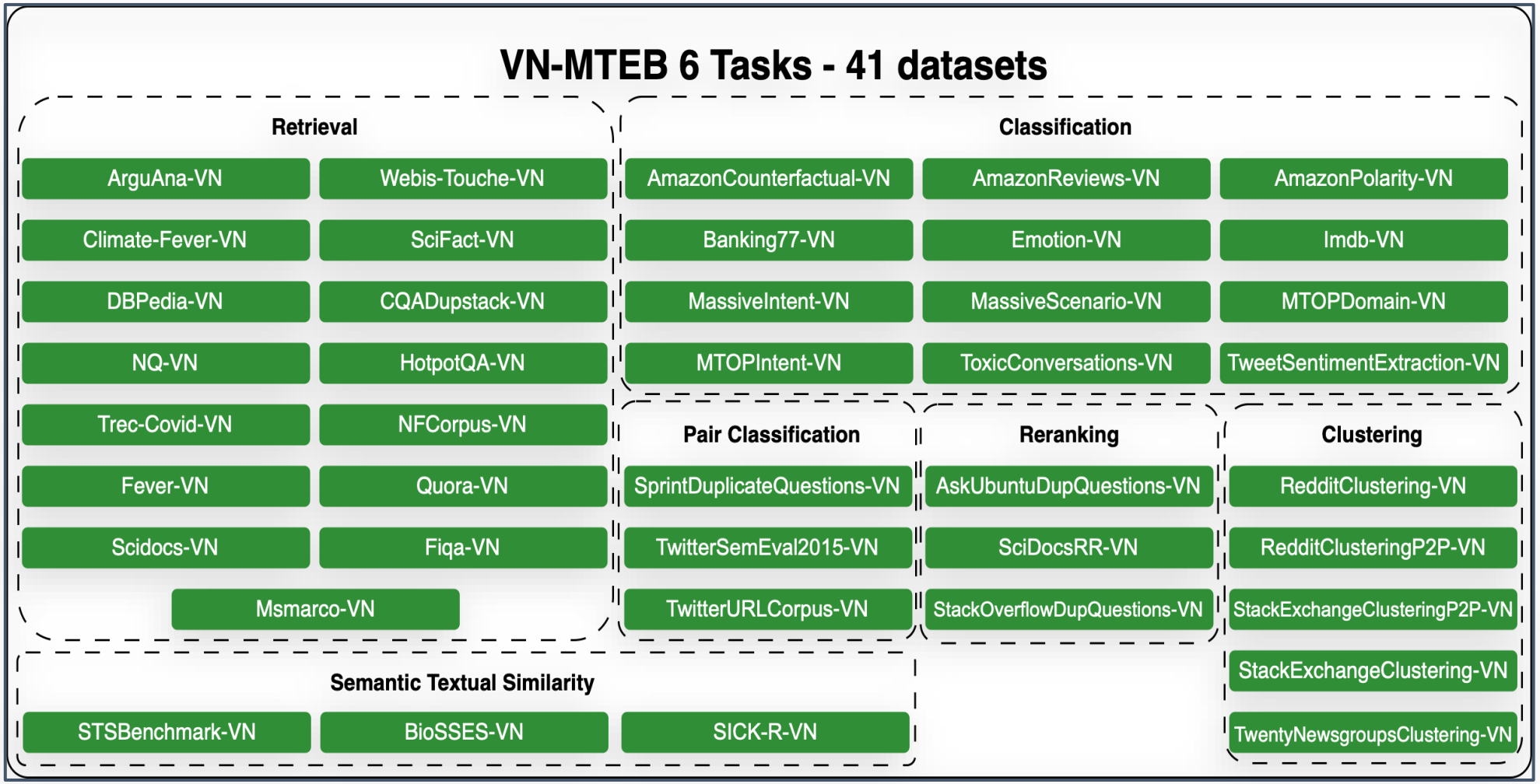
Explain the reason first, score later

- Using **Llama-SEA-LION-v3-70B-IT**.
- There are criteria: grammar; named entity recognition (NER); number, links, special characters; fluency; meaning preservation

$$score_{LLM_{Judge}} = \frac{\sum_{i \in S} \alpha_i \cdot score_i}{|S|}$$

- Where S is the set of evaluation criteria.
- $\sum_{i \in S} \alpha_i = 1$ and $score_i \in [1, 5]$ denote important weight.

VN-MTEB


VN-MTEB 6 Tasks - 41 datasets
Retrieval
ArguAna-VN
Webis-Touche-VN
Climate-Fever-VN
SciFact-VN
DBPedia-VN
CQADupstack-VN
NQ-VN
HotpotQA-VN
Trec-Covid-VN
NFCorpus-VN
Fever-VN
Quora-VN
Scidocs-VN
Fiqa-VN
Msmarco-VN
Classification
AmazonCounterfactual-VN
AmazonReviews-VN
AmazonPolarity-VN
Banking77-VN
Emotion-VN
Imdb-VN
MassiveIntent-VN
MassiveScenario-VN
MTOPDomain-VN
MTOPIntent-VN
ToxicConversations-VN
TweetSentimentExtraction-VN
Pair Classification
SprintDuplicateQuestions-VN
TwitterSemEval2015-VN
TwitterURLCorpus-VN
Reranking
AskUbuntuDupQuestions-VN
SciDocsRR-VN
StackOverflowDupQuestions-VN
Clustering
RedditClustering-VN
RedditClusteringP2P-VN
StackExchangeClusteringP2P-VN
StackExchangeClustering-VN
TwentyNewsgroupsClustering-VN
Semantic Textual Similarity
STSBenchmark-VN
BioSSES-VN
SICK-R-VN

Benchmark Result & Conclusion

Num. Datasets (→)	Size (Params)	Dim (Dim)	Type	Retr. 15	Class. 12	PairClass. 3	Clust. 5	Rerank. 3	STS 3	Avg. ↑ 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.

Scan to receive 150$ FREE CREDIT
and experience powerful GPU AI Cloud Infrastructure

GreenNode at a glance

NVIDIA Cloud Service Partner in APAC

#AI #Cloud #Security

- Serving enterprise clients & AI startups in US, EMEA, APAC
- AI Cloud Infrastructure is currently located in Bangkok, Hanoi and Ho Chi Minh City
- Compliant with ISO 27000, PCI DSS, and TVRA standards

GREENNODE | NVIDIA

GreenNode in South-East Asia

AI Cloud Infrastructure is currently located in Bangkok, Hanoi and Ho Chi Minh City

Thank You
For Listening

Scan to download the GreenNode slide!