GREENNODE

VN-MTEB

Vietnamese Massive Text ↘ Embedding Benchmark

Presented by

Pham Bao Loc
Senior AI Engineer

locpb@greennode.ai

MTEB
MMTEB & VN-MTEB ↘

MMTEB: Massive Multilingual Text Embedding Benchmark

Kenneth Enevoldsen[1,†,‡], Isaac Chung[2,‡], Imene Kerboua[3,4,‡], Márton Kardos[1,‡],
Ashwin Mathur[2], David Stap[5], Jay Gala[6], Wissam Siblini[2], Dominik Krzemiński[2],
Genta Indra Winata[2], Saba Sturua[7], Saiteja Utpala[8], Mathieu Ciancone[9], Marion Schaeffer[9],
Gabriel Sequeira[2], Diganta Misra[56,57], Shreeya Dhakal[2], Jonathan Rystrøm[11], Roman Solomatin[12,‡],
Ömer Çağatan[13], Akash Kundu[14,15], Martin Bernstorff[1], Shitao Xiao[16], Akshita Sukhlecha[2],
Bhavish Pahwa[8], Rafał Poświata[17], Kranthi Kira
Björn Plüster[19], Jan Philipp Harries[19], Loïc Mag
Dawei Zhu[20], Hippolyte Gisserot-Boukhlef[21,22], T
Taemin Lee[25], Marek Šuppa[27,28], Crystina Zhang
Andrianos Michail[31], John Yang[32], Manuel Fayss
Manan Dey[34], Dipam Vasani[2], Pranjal Chitale[35],
Artem Snegirev[38], Michael Günther[7], Mengzhou
Gayatri Krishnakumar[42], Anna Maksimova[38], Si
Henil Panchal[45], Aleksandr Abramov[38], Malte O
Lester James Miranda[47], Alena Fenogenova[38], G
Alessia Borghini[36], Federico Cassano[51], Hongjin
Sara Hooker[30], Chenghao Xiao[53,‡], Vaibhav Adla
Niklas Muennighoff[32,47,58,‡]

[1]Aarhus University, [2]Individual Contributor, [3]Esker,
[5]University of Amsterdam, [6]MBZUAI, [7]Jina AI, [8]M
[9]Wikit, [10]McGill University, [11]University of Oxford
[13]Koç University, [14]Heritage Institute of Technology
[17]National Information Processing Institute, [18]New
[20]Peking University, [21]CentraleSupélec, [22]Artefact F
[24]Wrocław University [25]Korea University, [26]Illuin T
[27]Comenius University Bratislava, [28]Cisco Systems,
[30]Cohere For AI, [31]University of Zurich, [32]Stanford
[34]Salesforce, [35]IIT Madras, [36]Sapienza University o
[37]University of Pennsylvania, [38]SaluteDevices, [39]Pri
[40]University of Washington, [41]Imperial College Lon
[43]Robert Koch Institute, [44]HSE University, [45]Nirma
[47]Allen Institute for AI, [48]Tano Labs, [49]The London
[50]Cornell University, [51]Northeastern University, [52]H
[53]Durham University, [54]ServiceNow Research, [55]Jol

VN-MTEB: Vietnamese Massive Text Embedding Benchmark

Loc Pham[♠], Tung Luu[♠], Thu Vo[♠], Minh Nguyen[♣], Viet Hoang[♠],
[♠] GreenNode AI, Singapore
[♣]School of Electrical Engineering, International University, VNU-HCMC, Vietnam
{locpb, tunglq, thu, viethq5}@greennode.ai, {nntminh}@hcmiu.edu.vn

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. 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 Em-

ken 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. However, for low-resource languages like Vietnamese, there is still a lack of di-

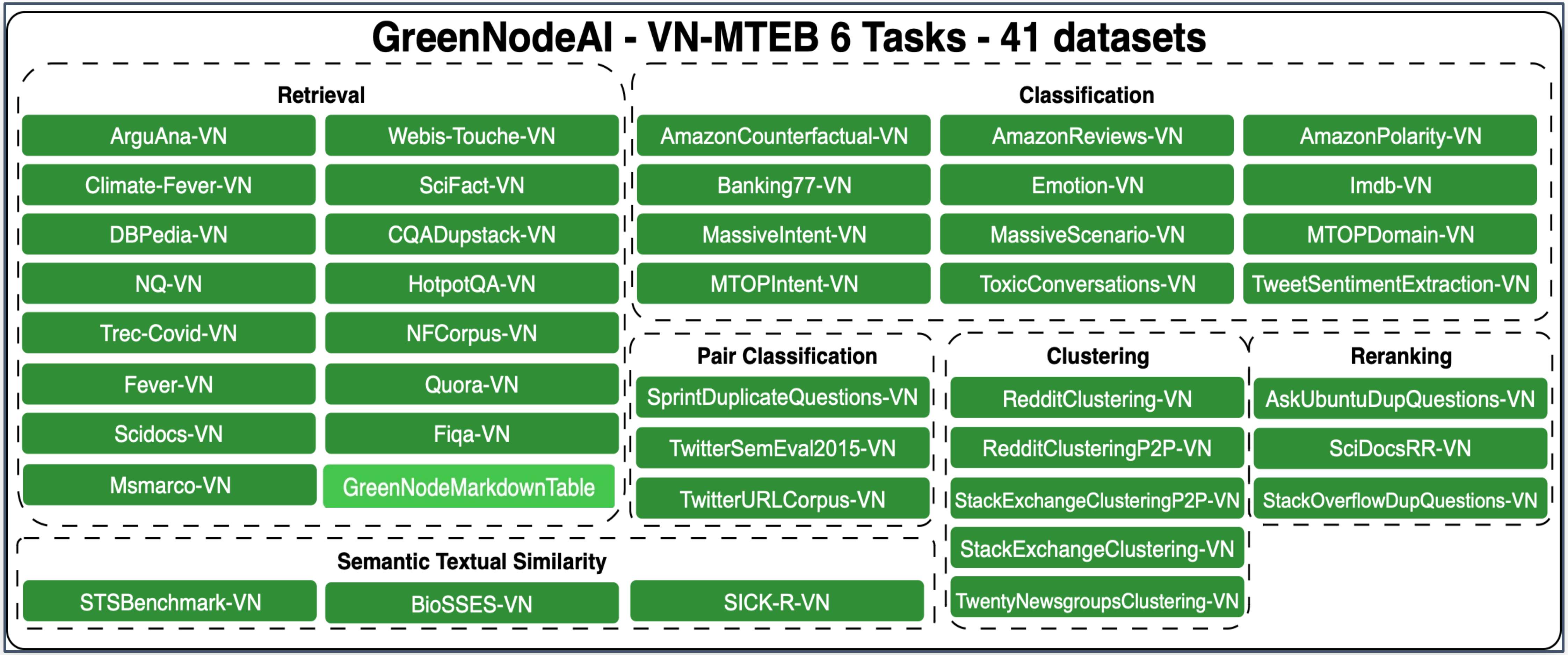
MTEB: Massive Text Embedding Benchmark

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

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,

VN-MTEB



GreenNodeAI - 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
- GreenNodeMarkdownTable

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

Clustering

- RedditClustering-VN
- RedditClusteringP2P-VN
- StackExchangeClusteringP2P-VN
- StackExchangeClustering-VN
- TwentyNewsgroupsClustering-VN

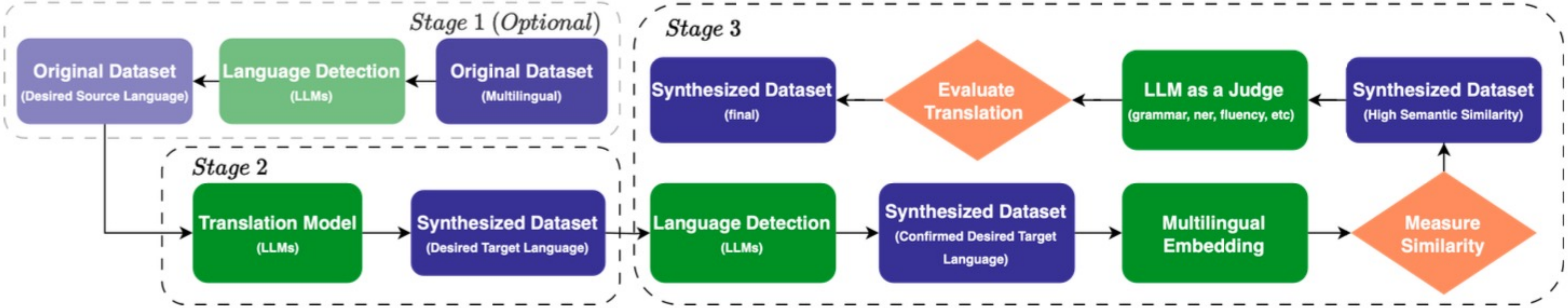
Reranking

- AskUbuntuDupQuestions-VN
- SciDocsRR-VN
- StackOverflowDupQuestions-VN

Semantic Textual Similarity

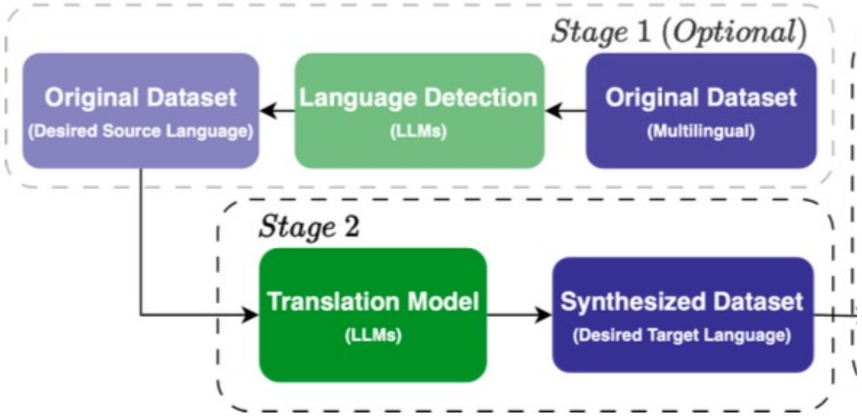
- STSBenchmark-VN
- BioSSES-VN
- SICK-R-VN

VN-MTEB ↘



Translation pipeline overview

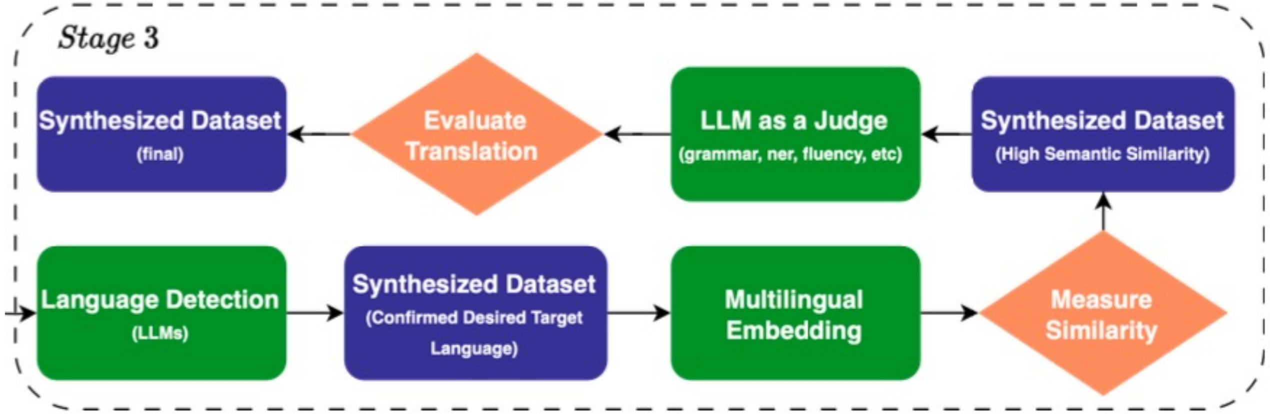
VN-MTEB



we chose the best model according to
SouthEast Asian Holistic Evaluation of Language Models (SEA Healms)
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).

Stage 2: Translation

VN-MTEB ↘



Stage 3
Synthesized Dataset (final)
Evaluate Translation
LLM as a Judge (grammar, ner, fluency, etc)
Synthesized Dataset (High Semantic Similarity)
Language Detection (LLMs)
Synthesized Dataset (Confirmed Desired Target Language)
Multilingual Embedding
Measure Similarity

Stage 3: Evaluate and filtering bad samples

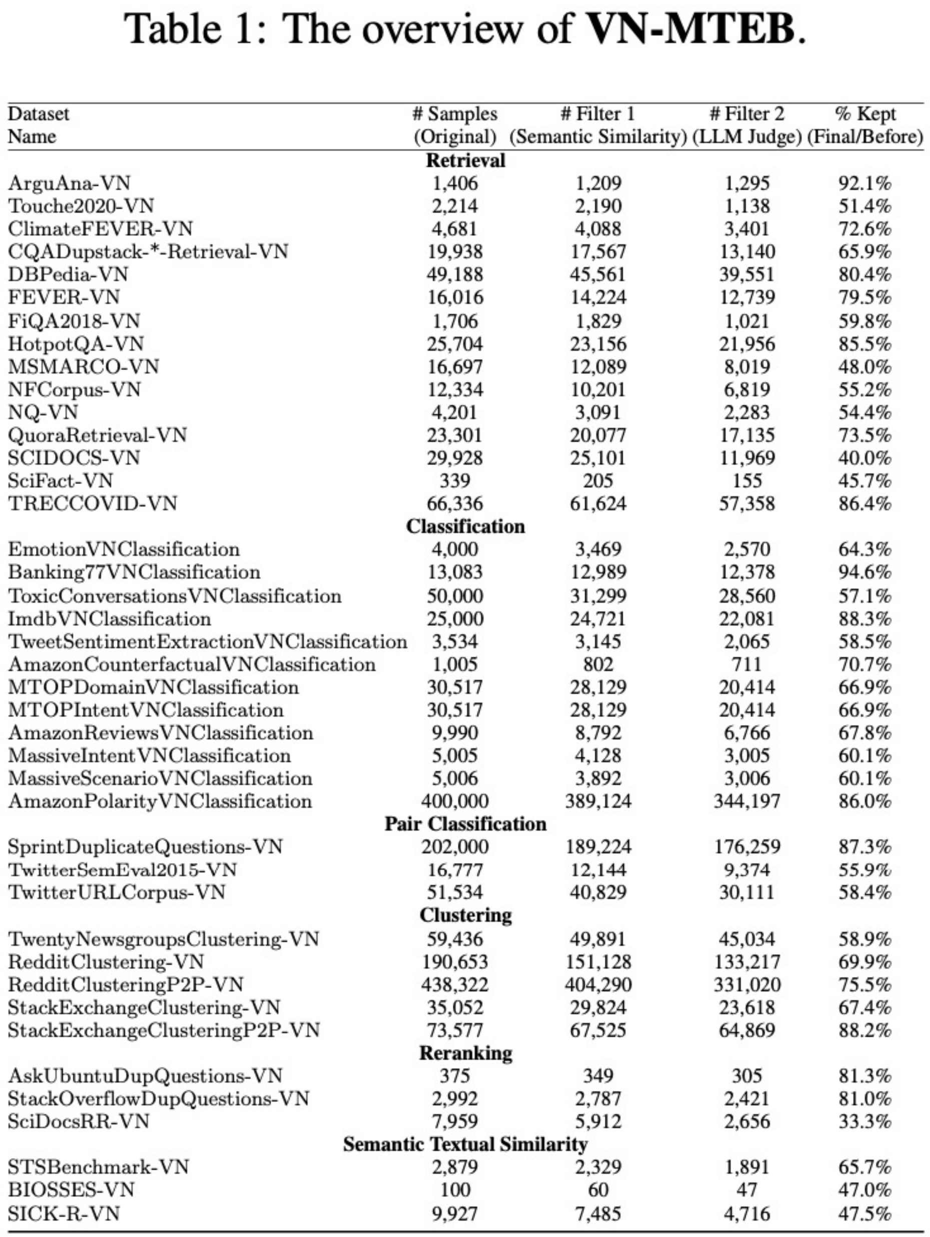
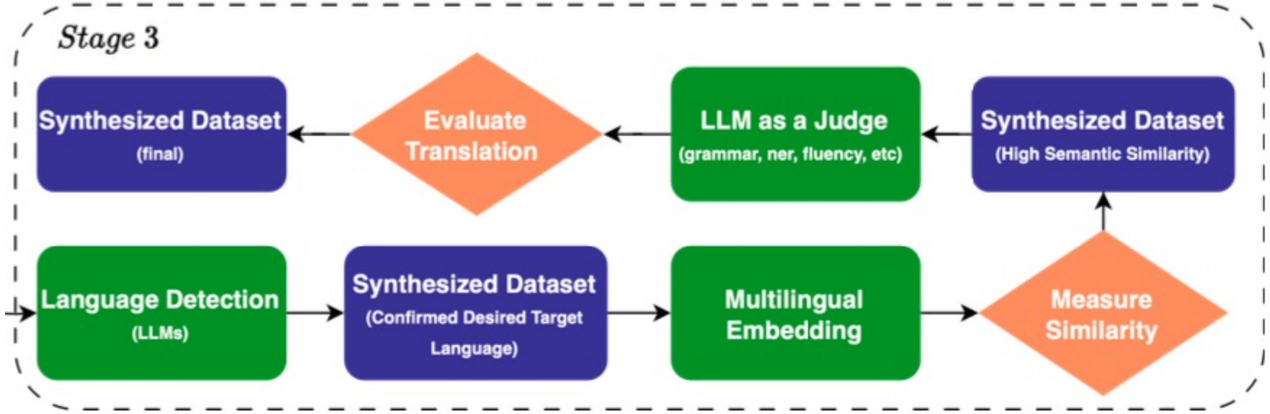


Table 1: The overview of **VN-MTEB**.

Dataset Name	# Samples (Original)	# Filter 1 (Semantic Similarity)	# Filter 2 (LLM 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%

VN-MTEB


Stage 3
Synthesized Dataset (final)
Evaluate Translation
LLM as a Judge (grammar, ner, fluency, etc)
Synthesized Dataset (High Semantic Similarity)
Language Detection (LLMs)
Synthesized Dataset (Confirmed Desired Target Language)
Multilingual Embedding
Measure Similarity

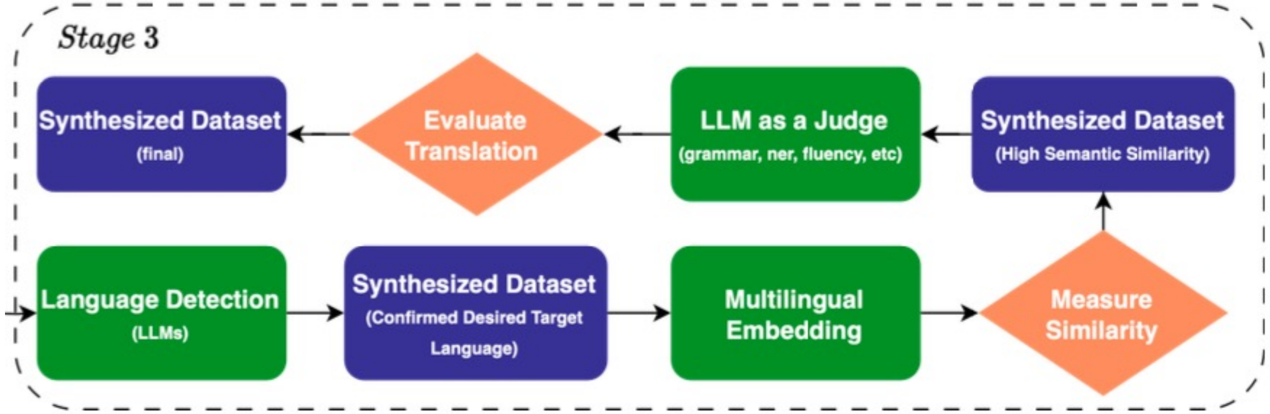
In our pipeline, we refer to the Seahelm leaderboar and select:

- **Qwen/Qwen2.5-3B-Instruct** to perform detecting language

Stage 3: Evaluate and filtering bad samples

- **Alibaba-NLP/gte-Qwen2-7B-Instruct** to create embedding and calculate cosine similarity.

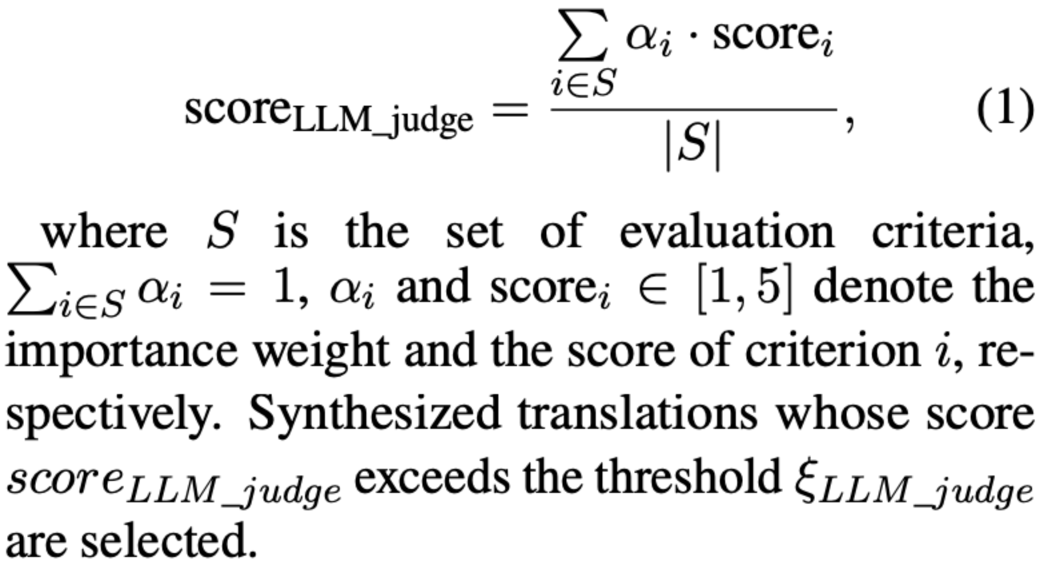
VN-MTEB


Stage 3
Synthesized Dataset (final)
Evaluate Translation
LLM as a Judge (grammar, ner, fluency, etc)
Synthesized Dataset (High Semantic Similarity)
Language Detection (LLMs)
Synthesized Dataset (Confirmed Desired Target Language)
Multilingual Embedding
Measure Similarity

Stage 3: Evaluate and filtering bad samples

Evaluate the following criteria:

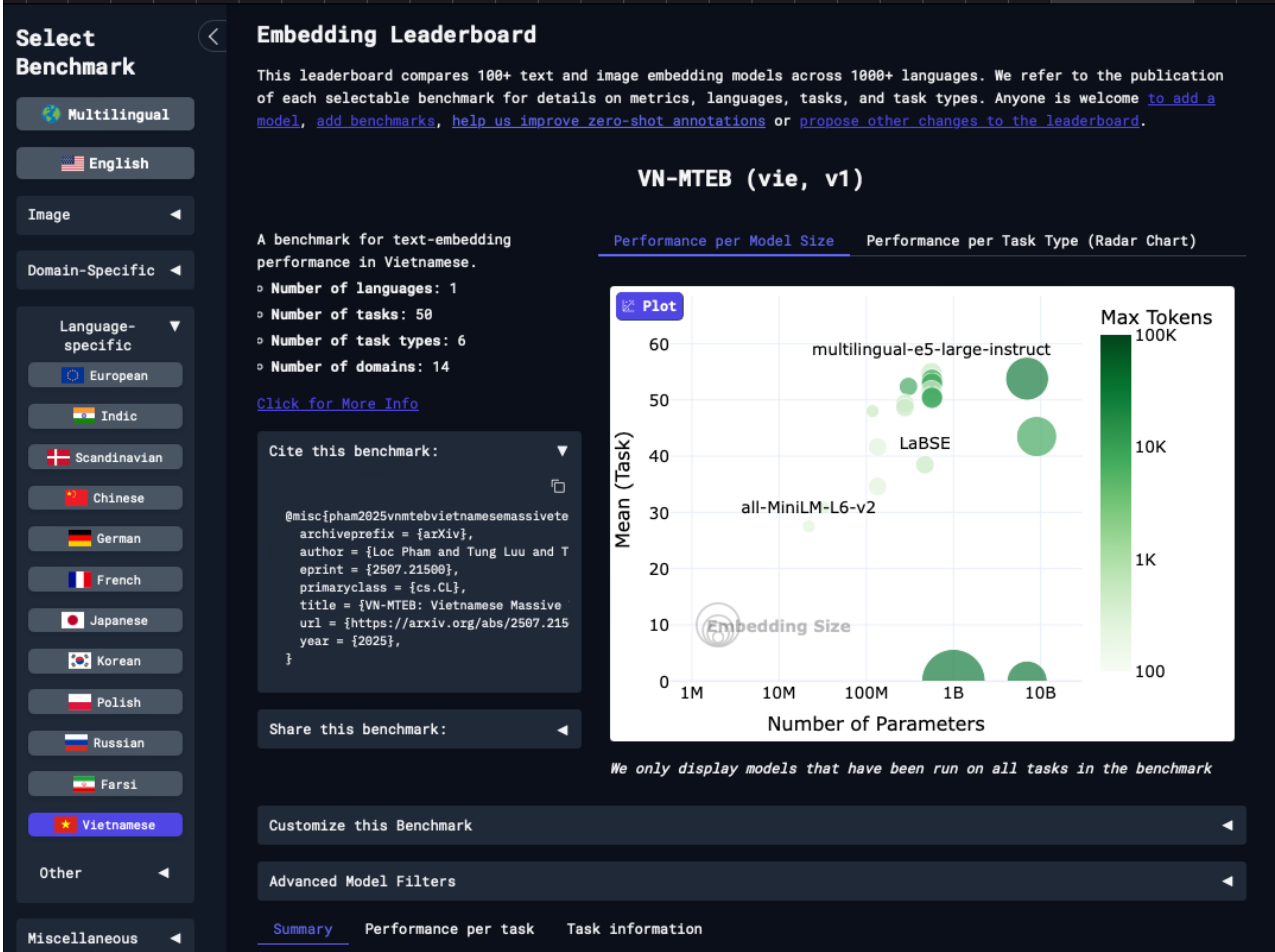
- Grammar
- Named entity recognition (NER)
- Numbers/links/special characters
- Fluency
- Meaning preservation.



$$\text{score}_{\text{LLM\_judge}} = \frac{\sum\limits_{i\in S} \alpha_i \cdot \text{score}_i}{|S|}, \tag{1}$$

where S is the set of evaluation criteria, $\sum_{i\in S} \alpha_i = 1$, α_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 ξ_{LLM_judge} are selected.

MOTIVATION ↘



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.

VN-MTEB (vie, v1)

A benchmark for text-embedding performance in Vietnamese.

- **Number of languages**: 1
- **Number of tasks**: 50
- **Number of task types**: 6
- **Number of domains**: 14

Click for More Info

Cite this benchmark: ▼

```
@misc{pham2025vnmtebvietnamesemassivete
  archiveprefix = {arXiv},
  author = {Loc Pham and Tung Luu and T
  eprint = {2507.21500},
  primaryclass = {cs.CL},
  title = {VN-MTEB: Vietnamese Massive
  url = {https://arxiv.org/abs/2507.215
  year = {2025},
}
```

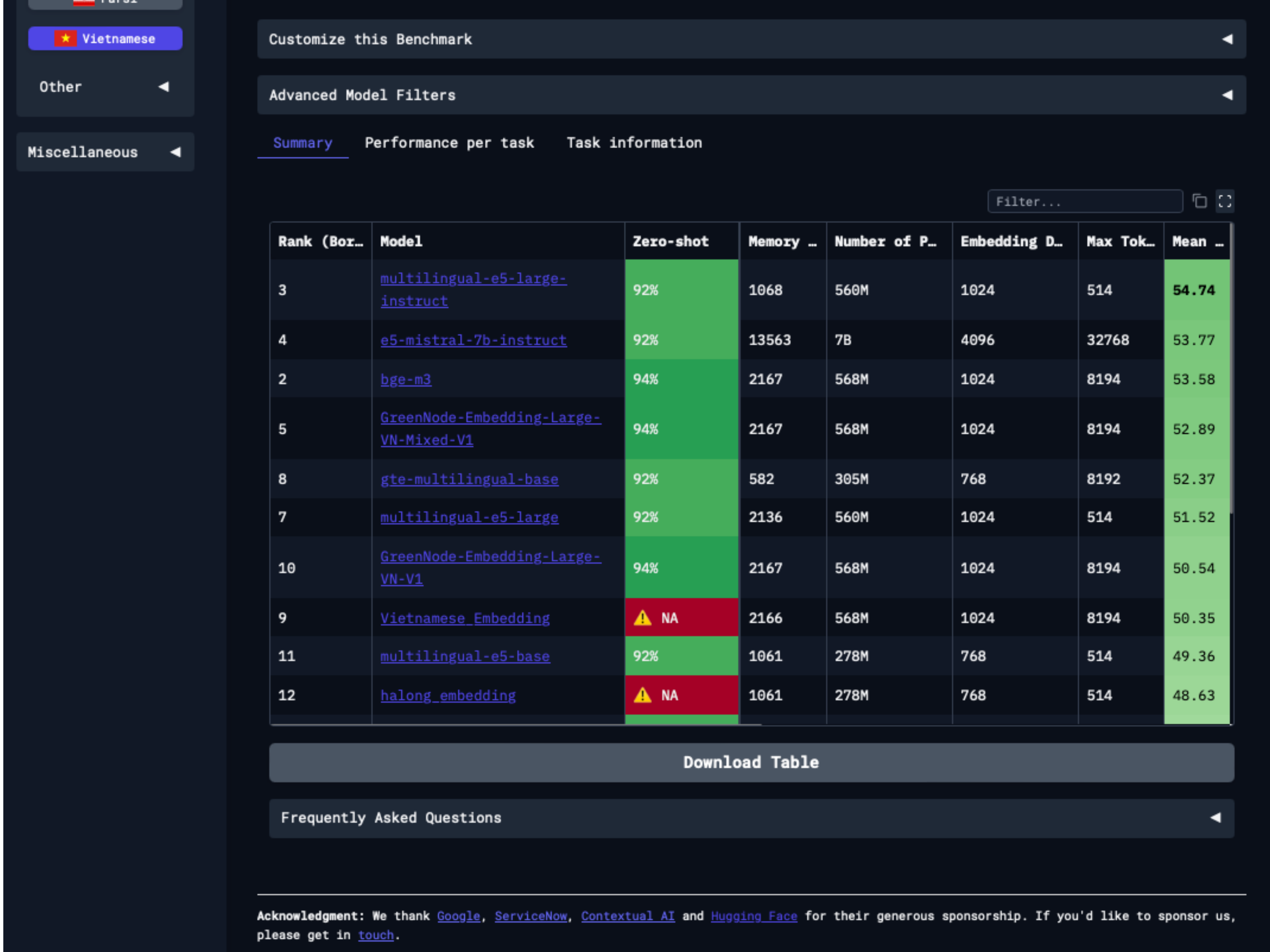
Share this benchmark: ◀

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

Customize this Benchmark ◀

Advanced Model Filters ◀

MOTIVATION ↘



Vietnamese

Other ◀

Miscellaneous ◀

Customize this Benchmark ◀

Advanced Model Filters ◀

Summary | Performance per task | Task information

Filter...

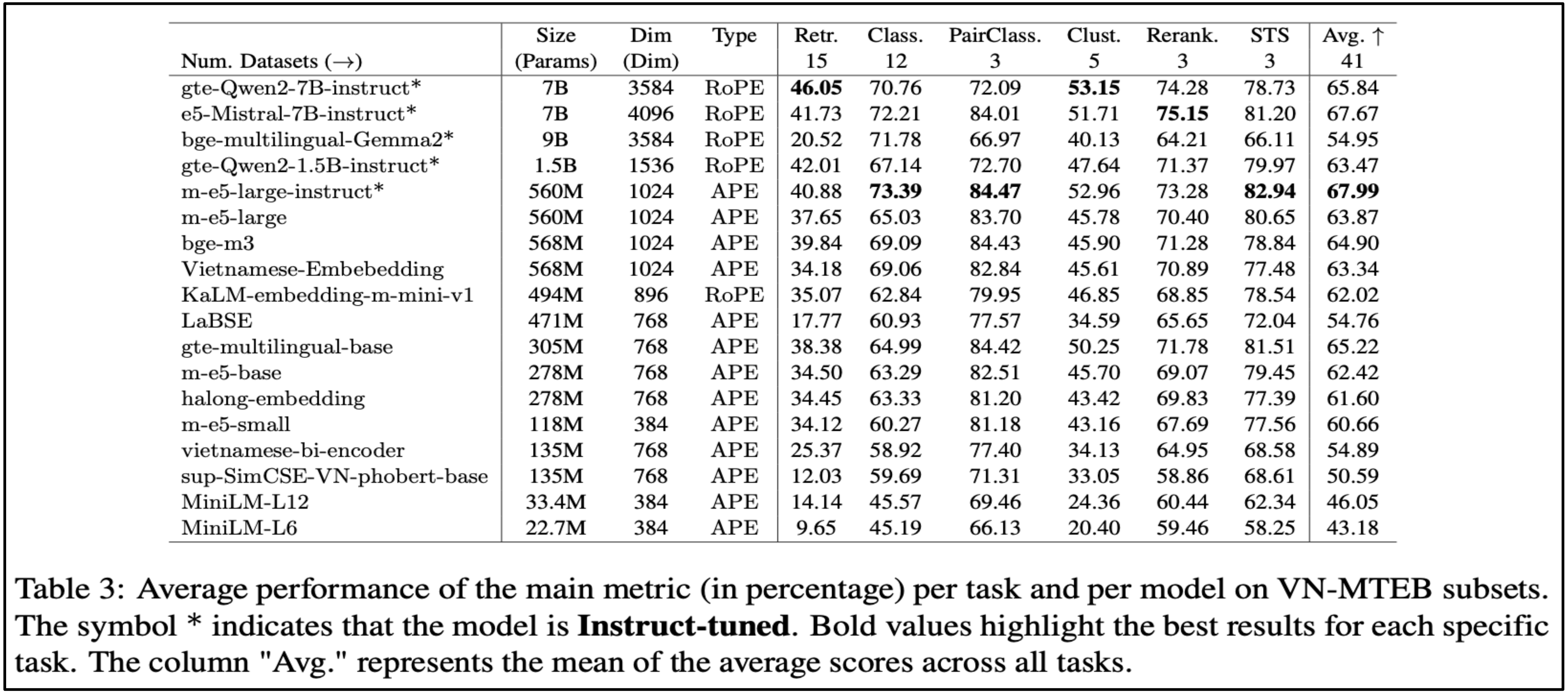
Rank (Bor…	Model	Zero-shot	Memory …	Number of P…	Embedding D…	Max Tok…	Mean …
3	multilingual-e5-large-instruct	92%	1068	560M	1024	514	**54.74**
4	e5-mistral-7b-instruct	92%	13563	7B	4096	32768	53.77
2	bge-m3	94%	2167	568M	1024	8194	53.58
5	GreenNode-Embedding-Large-VN-Mixed-V1	94%	2167	568M	1024	8194	52.89
8	gte-multilingual-base	92%	582	305M	768	8192	52.37
7	multilingual-e5-large	92%	2136	560M	1024	514	51.52
10	GreenNode-Embedding-Large-VN-V1	94%	2167	568M	1024	8194	50.54
9	Vietnamese_Embedding	⚠ NA	2166	568M	1024	8194	50.35
11	multilingual-e5-base	92%	1061	278M	768	514	49.36
12	halong_embedding	⚠ NA	1061	278M	768	514	48.63

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.

BENCHMARK RESULT & CONCLUSION ↘



	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.