SEA-BED: How Do Embedding Models Represent Southeast Asian Languages?

Anonymous ACL submission

Abstract

Multilingual text embeddings are often assumed to encode meaning in a perspective-independent semantic space, yielding stable similarity judgments across tasks and languages. Our results show that this assumption does not hold in practice. We introduce SEA-BED, a large-scale benchmark covering 10 Southeast Asian (SEA) languages and diverse embedding tasks, designed to systematically examine how embedding performance varies across tasks, languages, and language-task combinations. Across extensive evaluations, we observe that no single model performs uniformly well across SEA languages; task difficulty differs markedly within languages, and success on one task does not reliably generalize to others. Language-task analyses further reveal highly non-uniform performance landscapes, where performance varies across different language-task combinations. These findings call for closer attention to performance measurements that provide an expansive view across languages and tasks to uncover inconsistencies in semantic representation. Based on these observations, we provide insights for future model development, including data, algorithmic, and architectural considerations.

1 Introduction

Text embedding plays a crucial role in NLP by transforming complex linguistic structures into fixed-size vectors that capture semantic locality. These embeddings are fundamental for various downstream tasks, including semantic textual similarity, retrieval, and re-ranking. New embedding models have also shifted toward global multilinguality. For instance, Wang et al. (2024a) proposed a multilingual model on Mistral-7B that supports 93 languages, while Jina-embeddings-v3 (Sturua et al., 2024a) includes 89 languages.

A common **ideal** motivating such multilingual development is that a single embedding model should approximate a robust semantic space handling multiple languages simultaneously. This ideal posits that semantic equivalence should primarily determine geometric similarity, independent of linguistic surface forms, an assumption that underlies the use of multilingual embeddings in cross-lingual retrieval (Conneau and Kiela, 2018), semantic search, and transfer learning.

In **practice**, however, *nothing in modern embedding architectures guarantees such ideal behavior.* Multilingual text embeddings are learned through co-occurrence statistics in training data. It reflects linguistic distributions, translation conventions, cultural norms, and domain-specific usage patterns present in the data. Hence, real embedding spaces often diverge across tasks (the "task consistency problem") and languages (the "language consistency problem"), revealing gaps between the ideal of perspective independence and the structure that models actually capture.

Southeast Asia (SEA) presents a uniquely rich environment for exploring the **real-world limitations** of multilingual text embedding models. The region spans multiple language families (Austronesian, Tai-Kadai, Sino-Tibetan, Austroasiatic, among others), writing systems (Latin alphabet, abugida, logographic), and typological properties (isolating, analytic, and agglutinative structures). Despite the rich epistemic opportunities, **SEA languages remain underrepresented** in existing embedding benchmarks.

A practical approach to multilingual benchmarking is translation. For example, XNLI (Conneau and Kiela, 2018), Tatoeba (community, 2021), and SIB-200 (Adelani et al., 2024) rely on samples translated from English to extend language and task coverage efficiently. While translation can approximate native data for certain tasks, it may also alter semantic or discourse properties in lower-resource settings. An alternative is to use native-authored datasets, which provide stronger linguistic and cul-

Benchmark	# Languages	# SEA Languages	# Datasets	# Task	# New datasets	# SEA datasets	# Human-Crafted datasets (only SEA languages)
MTEB-French (Ciancone et al., 2024a)	1	N/A	18	8	3	N/A	N/A
C-Pack (Xiao et al., 2024a)	1	N/A	35	6	35	N/A	N/A
SEB (Enevoldsen et al., 2024)	4	N/A	24	4	24	N/A	N/A
MMTEB (Enevoldsen et al., 2025)	1,090	9	270	10	5	22	21
SEA-BED (ours)	10	10	169	9	11	169	120 (71.01 %)

Table 1: The statistics of our benchmark compared to existing text embedding benchmarks.

tural grounding, motivating MMTEB's recent focus on provenance nativity (Enevoldsen et al., 2025). However, MMTEB includes only 22 SEA datasets across 10 languages, resulting in limited coverage and task diversity. As a result, current benchmarks provide a narrow view of how embedding models behave in such an epistemically rich testing ground.

Proposed Benchmark. When constructing a multilingual benchmark, the first design decision concerns the evaluation scope, the aspects of model behavior we aim to observe. In this regard, SEA-BED spans 9 task types across 10 SEA languages, enabling broad cross-task and cross-lingual comparative analysis of multilingual embedding models. The second decision concerns data sourcing, with the goal of maximizing breadth and depth of coverage while ensuring data quality within the defined scope. SEA-BED includes 169 datasets, most of which are not covered by existing multilingual sentence embedding benchmarks. We combine native-authored datasets with carefully curated translated datasets, enabling systematic comparison between native and transferred provenance. We additionally contribute 11 new datasets for Thai and Burmese, enabling deeper examination of semantic similarity, relation understanding, and cross-lingual transfer. A comparison between SEA-BED and existing benchmarks is given in Table 1.

Proposed Study. With SEA-BED in place, we can systematically examine aspects of multilingual text embeddings that were previously unmeasurable in the SEA region, particularly how model behavior varies across tasks and languages. *We focus on a single research question: how embedding model behavior varies across tasks and Southeast Asian languages.* Specifically, we investigate performance patterns under various evaluation conditions. To this end, our analysis adopts three complementary comparison views: (i) a language-model view, which analyzes how model behavior varies across languages; (ii) a task-model view, which examines how different embedding models perform across task categories; and (iii) a language-task view, which aggregates model performance over language-task combinations to reveal conditional patterns that emerge at their intersection. These three comparative views provide a structured characterization of embedding behavior across languages, tasks, and model dimensions.

Key Results. Model performance across tasks and SEA languages shows clear drift across languages and task types, with reconfigurations of relative task difficulty rather than uniform scaling, especially in Burmese and Lao. Furthermore, there is no single model that performs uniformly well across SEA languages and task types. These results indicate that embedding performance is inherently task- and language-dependent. **Insights** from our analysis also point to three avenues for performance improvement: enhancing data quality and coverage, refining algorithmic handling of semantic and cross-lingual variation, and adapting architectural designs to better accommodate SEA linguistic diversity.

Our contributions are as follows.

- **Resource.** We introduce SEA-BED, a benchmark of 169 datasets across 9 task types and 10 SEA languages, including 11 new Thai and Burmese datasets that expand coverage of previously missing task-language combinations.
- **Experimental Studies.** We evaluate 17 embedding models through a set of studies to analyze performance variation across tasks, languages, and model.
- **Insights for Future Model Development.** Our findings reveal substantial instability in model performance and highlight concrete directions for improving multilingual embedding robustness in SEA settings.

2 Proposed Benchmark: SEA-BED

2.1 Language-Task Coverage

We constructed SEA-BED with broad language-task coverage to extend existing evaluation resources for Southeast Asian languages. SEA-BED spans 10 languages and 9 task types, covering linguistic contexts that are substantially underrepresented in benchmarks such as MMTEB. This ex-

panded structure enables the systematic investigation of embedding behavior across tasks and languages.

We adopt the MMTEB task taxonomy (Enevoldsen et al., 2025) to ensure comparability with prior work, and extend their coverage of SEA languages. The benchmark spans the following task types. **(i) Classification**. Learn a classifier over sentence embeddings to assign labels to individual sentences. **(ii) Multi-label Classification**. Predict multiple labels for each input text using a classifier trained on embeddings. **(iii) Pair Classification**. Predict a binary relationship between two sentences based on their embedding similarity. **(iv) Semantic Textual Similarity (STS)**. Measure similarity between sentence pairs using continuous scores derived from distance metrics computed over their embeddings. **(v) Clustering**. The task groups embedded texts into clusters based on semantic similarity, using k-means with the number of unique labels of k. **(vi) Bitext Mining**. Identify translation pairs across two languages by retrieving the closest match for each sentence in a source set. **(vii) Retrieval**. Retrieve relevant documents for a given query by computing embedding similarity between the query and candidate texts. **(viii) Instruction Retrieval**. Extend traditional retrieval by incorporating detailed instructions into queries, pairing each query with a corresponding detailed instruction that outlines the criteria for determining document relevance. **(ix) Reranking**. Reorder a set of candidate documents based on embedding similarity to a query to improve relevance ranking. Note that summarization is omitted due to data availability constraints.

Table 2 summarizes the task subtypes and their language coverage. SEA-BED contains 19 subtypes across 9 task types and substantially increases the number of language-subtype pairs relative to MMTEB (from 10 to 19). Notably, several subtypes *Language Identification, Toxic Language Detection, Dialect Pairing, Instruction QA, and Article Reranking* are introduced for the first time in SEA language evaluations.

2.2 Evaluation Data Sourcing

Data Sourcing Overview. We now turn our attention to how datasets were sourced and organized to enable systematic evaluation within the scope described in the previous subsection. SEA-BED consists of 169 datasets spanning 9 task types and 10 languages, as previously discussed in Table 1. While MMTEB contains 270 datasets, only 22 involve SEA languages, which reflects a limited regional representation. In contrast, 147 of our datasets (86%) are not included in MMTEB, highlighting substantial gaps in existing benchmark coverage, making SEA-BED substantially more representative of SEA linguistic diversity. Of these, 120 were authored by native speakers in their respective languages, while the rest are sourced from translated datasets. This combination enables the study of provenance effects across task types and languages. We also introduce 11 new datasets for Thai and Burmese to expand coverage of previously underrepresented task-language combinations. Benchmark efficiency considerations (e.g., caching and downsampling) are described in Appendix B.

Task Type and Subtype	ind	tha	vie	mya	fil	khm	zsm	lao	tam	tet
Classification										
Language Identification	▲	▲	▲	▲	▲	▲	▲	▲		▲
Sentiment	↻	↻	↻	▲	↻	▲			△	
Topic Classification	↻	↻	↻	▲	↻	↻	↻	△	↻	
Toxic Language Detection	▲	▲	▲		▲					
Multi-label Classification										
Sentiment	▲		▲							
Topic Classification		▲		▲	▲	▲				
Toxic Language Detection	▲									
Pair Classification										
Textual Entailment	↻	↻	↻	▲	▲	▲	▲	▲	▲	
STS										
Multilingual STS	↻	▲	▲	▲	▲	▲	▲	▲	▲	
Cross-lingual STS		▲		▲					↻	
Clustering										
Topic Clustering	↻	↻	↻	↻	↻	↻	↻	↻	△	
Bitext Mining										
Cross-lingual pairing	↻	↻	↻	↻	↻	↻	△	↻	↻	▲
Dialect pairing	▲	▲	▲	▲	▲	▲	▲	▲		
Written-forms pairing	▲	▲	▲	▲						
Retrieval										
Article Retrieval	↻	↻	▲	▲						
Long Document Retrieval		▲								
Question Answering	▲	▲	↻				▲		▲	
Instruction Retrieval										
Instruction QA	▲	▲	▲							
Reranking										
Article Reranking	▲	▲								

Table 2: Task coverage of SEA-BED compared to MMTEB. Cells use ↻ (present in MMTEB and directly reused), △ (present in MMTEB and extended in SEA-BED), and ▲ (entirely new). Similar task types are grouped by highlight.

Domain Coverage. Domain diversity is essential for realistic and discriminative evaluation, as semantic representations vary substantially across domains and influence model robustness in real-world applications. MMTEB provides coverage of widely studied domains in SEA languages (e.g., news, non-fiction, and encyclopedias). SEA-BED complements and extends this foundation by spanning 17

domains across 10 SEA languages, including a broader range of formal, informal, and application-oriented texts. Much of this coverage is supported by new or newly sourced datasets, with some domains (academic, blogs, medical, and subtitles) newly introduced. Table 3 presents the domain coverage of the SEA-BED benchmark compared to MMTEB across languages, indicating reused, extended, and newly introduced domains. The full domain descriptions are provided in Appendix C.

Domain	ind	tha	vie	mya	fil	khm	zsm	lao	tam	tet
Academic			▲							
Blog	▲	▲	▲	▲	▲					
Constructed	▲	▲	▲	▲	▲		▲			
Encyclopedia	△	△	△	△	△	△	▲	△	⟳	
Fiction	▲	⟳	△	▲	⟳	▲	▲	▲	⟳	
Government	▲	△	△	▲	▲	▲	▲	▲	⟳	▲
Legal	▲	▲	▲	▲		▲	▲	▲	⟳	
Medical		▲	▲							
News	△	△	△	△	△	△	△	△	△	
Non-fiction	△	△	△	△	△	△	▲	△	⟳	
Religious	△	⟳	⟳	⟳	⟳	▲	▲	▲	⟳	▲
Reviews	△	△	△		▲	▲	▲			
Social	▲	▲	▲	▲	△				⟳	
Spoken	△	△	△	▲	△	△	△	△	△	▲
Subtitles	▲									
Web	△	▲	▲	▲			▲		⟳	▲
Written	△	△	△	△	△	△	▲	△	△	▲

Table 3: Domain coverage of SEA-BED benchmark compared to MMTEB. Cells use ⟳ to denote tasks present in MMTEB and directly reused, △ for tasks present in MMTEB but extended in SEA-BED, and ▲ for entirely new domains.

Data Quality Assurance. Data quality assurance is achieved through systematic review processes conducted by native speakers of SEA languages, all of whom are also proficient in English. These annotators verified and validated the data for grammatical correctness, native written style, appropriate language usage (excluding code-switching), and the accuracy of the gold standard annotations. This human verification process results in approximately 8% of the datasets being removed, reducing the number from 182 to 169 datasets. See Appendix A for the annotator guidelines and details.

New Datasets. In addition to curated datasets, we construct new Thai and Burmese datasets for semantic textual similarity, natural language inference, and multi-label classification. These new resources address task-level gaps in SEA languages and expand evaluation coverage in previously underrepresented settings. Given the importance of these tasks for downstream retrieval and re-ranking performance (Gao et al., 2021; Chuang et al., 2022), we release 4 Thai datasets comprising 3,147 samples and 7 Burmese datasets with 13,177 samples to support systematic evaluation.

As shown in Table 4, we construct these datasets by human verification of Google NMT translated data from established English benchmarks for semantic textual similarity and natural language inference, including STSBenchmark (Cer et al., 2017), STS-2017 (Cer et al., 2017), STS-2022 (Chen et al., 2022), STS-2024 (Ousidhoum et al., 2024b), BIOSSES (Soğancıoğlu et al., 2017), and XNLI (Conneau et al., 2018), which serve as the original texts for dataset construction. We also translate the Thai multi-label dataset Prachathai67k (cstorm125, 2019) into Burmese,[1] providing a stronger starting point for creating Burmese resources. Translations were performed by native Thai and Burmese speakers (see Appendix A for annotator demographics and guidelines) following two instructions: (i) produce natural, conversational language, and (ii) make sentence subjects gender-neutral, given that both languages encode gender morphologically.

Dataset	mya	tha
Biosses	100	100
STS17	250	250
STS22	197	197
STS24	2,600	2,600
STSBenchmark	2,880	-
XNLI	5,000	-
Prachathai67k	2,150	-
Total number of samples	13,177	3,147

Table 4: Statistics of the new evaluation datasets included in SEA-BED.

3 Experimental Settings

Models. We evaluate 13 open-source and 4 proprietary multilingual text embedding models on SEA-BED, spanning both encoder-based and decoder-based architectures. Our model selection aims to encompass representative design choices in contemporary multilingual embedding models, rather than providing an exhaustive comparison or establishing a single best-performing approach. All models are treated as black-box embedding functions, and our analysis focuses on observed performance variation across tasks and languages. De-

[1] Thai is a more culturally and politically compatible source for Burmese translation than English.

Model	ind	tha	vie	mya	fil	khm	zsm	lao	tam	tet	Avg.
Number of datasets (→)	(70)	(55)	(41)	(35)	(31)	(22)	(19)	(19)	(18)	(4)	(314)
multilingual-e5-large-instruct (560M)	79.50	81.11	78.00	**78.37**	79.19	**78.13**	**84.60**	**83.94**	77.09	69.40	**78.93**$_{\pm 3.98}$
Qwen3-Embedding-8B (8B)	79.73	**81.49**	**78.99**	74.91	78.05	75.46	82.39	78.20	75.95	67.44	77.26$_{\pm 4.02}$
bge-multilingual-gemma2 (9B)	79.93	80.58	78.76	70.01	**79.61**	74.39	83.38	65.82	**80.96**	65.05	75.85$_{\pm 6.31}$
multilingual-e5-large (560M)	78.59	79.89	78.93	70.28	77.98	72.11	80.10	79.91	77.83	63.55	75.92$_{\pm 5.22}$
bge-m3 (568M)	78.09	77.59	75.91	73.12	75.78	76.23	82.54	82.26	77.51	65.53	76.46$_{\pm 4.55}$
GritLM-7B (7B)	**80.47**	72.84	77.37	45.05	77.49	52.58	78.41	30.07	60.42	**69.67**	64.44$_{\pm 16.13}$
e5-mistral-7b-instruct (7B)	79.23	74.77	75.37	48.85	78.10	56.49	78.82	27.99	66.73	66.73	65.32$_{\pm 15.74}$
Qwen3-Embedding-0.6B (595M)	75.60	75.85	75.13	49.08	63.11	44.10	69.51	29.78	61.12	63.38	60.67$_{\pm 14.55}$
multilingual-mpnet-base (278M)	74.60	73.91	72.66	61.19	52.02	64.44	75.48	65.63	63.31	50.78	65.40$_{\pm 8.53}$
LaBSE (471M)	73.98	70.20	72.60	73.63	76.99	74.06	82.87	79.84	76.59	69.11	74.99$_{\pm 3.99}$
multilingual-MiniLM-L12 (118M)	71.48	70.42	69.90	54.48	47.28	39.92	69.58	45.34	27.88	47.69	54.40$_{\pm 14.53}$
Gemma-SEA-LION-v3-9B-IT (9B)	49.86	41.67	51.90	30.80	54.14	39.53	49.24	22.01	29.20	25.06	39.34$_{\pm 11.27}$
Sailor2-8B-Chat (8B)	49.54	35.98	42.94	30.14	46.16	28.57	30.75	18.31	28.57	25.76	33.67$_{\pm 9.33}$
Proprietary models											
embed-multilingual-v3.0	**79.72**	**80.99**	**78.93**	**76.13**	**78.99**	**77.01**	**82.42**	**83.34**	**78.87**	**66.76**	**78.32**$_{\pm 4.39}$
jina-embeddings-v3	77.35	78.64	76.10	75.10	74.25	74.73	77.91	77.91	76.14	65.11	75.32$_{\pm 3.68}$
voyage-3	75.56	69.78	73.68	48.19	71.43	35.02	69.13	24.27	67.28	61.48	59.58$_{\pm 16.83}$
text-embedding-3-small	78.34	55.24	70.06	32.79	68.08	30.15	69.78	23.97	35.38	65.09	52.89$_{\pm 19.18}$

Table 5: Language-model performance view, where each cell reports scores averaged over all evaluated task types.

tailed model descriptions, training backgrounds, and per-model results are provided in Appendix I.

Evaluation Setup. We employ F1 for Bitext Mining, Classification, and Multi-label Classification. For Pair Classification, we use average precision (AP) as the main metric. For Clustering, we use the V-measure metric. For Retrieval, we use various metrics (i.e., nDCG@k, MRR@k, MAP@k, precision@k, and recall@k), with nDCG@10 as the primary metric. For Reranking, we use Mean Average Precision (MAP). In addition, we use nDCG@5 as the main metric for Instruction Retrieval following Weller et al. (2024). We employ the averaging strategy similar to previous works (Muennighoff et al., 2023; Enevoldsen et al., 2025), where all tasks are averaged equally with the standard deviation (SD) score. We acknowledge that the metrics for each task are different (e.g., F1 for classification and nDCG@10 for retrieval). Thus, we analyze both individual and average results, rather than focusing solely on the average score. All experiments were run on eight H100 (80 GB).

4 Experimental Results

4.1 Language-Model Comparisons

This study presents a language-wise analysis for behavior across language and model to pinpoint which SEA languages and scripts see the largest performance drops. Note that the number of datasets can be duplicated in multilingual scenarios, resulting in a higher number of datasets than the task-model comparison (Table 6).

Results. As shown in Table 5, we observe performance variation across languages that, when compared to the task-level results in Table 6, suggests that strong overall performance does not necessarily imply consistent multilingual coverage. For example, while multilingual-e5-large-instruct achieves the highest average score overall (78.93), its performance varies across languages, ranging from 69.40 points in Tetum to 84.60 points in Malay. We also observe the same trend in proprietary models. Moreover, we found that *some models do not fully support SEA languages*, e.g., GritLM-7B does not support Burmese, Khmer, and Lao, while bge-multilingual-gemma2 does not support Lao. This is because the training datasets for these languages are smaller and of lower quality. Although previous works demonstrate the use of LLMs to generate more datasets (Zhang et al., 2025; Muennighoff et al., 2024), applying this methodology to low-resource languages is underexplored and might not be effective. Thus, although these models performed well overall in this experimental study, their lack of support for some SEA languages renders them less suitable for real-world applications involving SEA languages.

Discussion. Experimental results demonstrate the "language consistency" problem, where models ex-

Model	Dim.	Clf	M. Clf	Pr. Clf	STS	Clust	Btxt	Rtrvl	In. Rtrvl	Rrnk	Avg.
Number of datasets (→)		(73)	(11)	(13)	(11)	(10)	(26)	(20)	(4)	(1)	(169)
multilingual-e5-large-instruct (560M)	1024	77.70	87.84	66.58	**75.59**	**58.09**	**87.86**	77.16	69.10	77.24	**75.24**$_{\pm 9.06}$
Qwen3-Embedding-8B (8B)	4096	**78.60**	90.57	63.10	75.31	52.93	84.78	**81.99**	70.81	78.51	75.18$_{\pm 10.84}$
bge-multilingual-gemma2 (9B)	3584	78.13	**90.89**	**73.87**	72.53	49.14	82.02	80.55	**71.52**	69.04	74.19$_{\pm 10.85}$
multilingual-e5-large (560M)	1024	78.24	88.94	65.79	69.61	47.83	84.51	78.25	66.06	**79.00**	73.14$_{\pm 11.66}$
bge-m3 (568M)	4096	75.98	89.89	68.73	73.27	42.23	86.18	73.56	58.51	75.98	71.59$_{\pm 13.48}$
GritLM-7B (7B)	4096	77.47	88.76	63.86	64.69	46.29	63.63	65.97	67.60	73.37	67.96$_{\pm 10.92}$
e5-mistral-7b-instruct (7B)	4096	76.65	88.32	63.81	63.50	49.48	65.30	72.93	54.46	75.33	67.75$_{\pm 11.24}$
Qwen3-Embedding-0.6B (595M)	1024	74.47	88.19	60.36	65.74	43.94	56.53	76.24	65.80	75.03	67.37$_{\pm 11.58}$
multilingual-mpnet-base (278M)	768	73.79	87.28	70.79	70.15	41.12	68.12	58.28	52.44	64.01	65.11$_{\pm 12.55}$
LaBSE (471M)	768	75.19	86.65	62.32	68.32	41.39	86.84	53.72	39.73	61.23	63.93$_{\pm 16.31}$
multilingual-MiniLM-L12 (118M)	768	70.50	84.88	65.70	64.59	31.50	53.23	52.47	48.66	62.27	59.31$_{\pm 14.25}$
Gemma-SEA-LION-v3-9B-IT (9B)	3584	75.87	89.94	57.77	38.85	39.94	15.31	22.03	11.02	65.49	46.25$_{\pm 26.18}$
Sailor2-8B-Chat (8B)	3584	76.43	90.21	56.71	37.25	38.51	4.31	10.09	3.29	47.05	40.43$_{\pm 29.25}$
Proprietary models											
embed-multilingual-v3.0	1024	**78.52**	**89.98**	**66.11**	73.11	48.99	**88.32**	**78.17**	65.59	**77.77**	**74.06**$_{\pm 11.89}$
jina-embeddings-v3	1024	77.40	88.97	63.61	**73.17**	**50.90**	81.86	76.28	**69.11**	72.49	72.64$_{\pm 10.30}$
voyage-3	1024	75.72	88.70	60.23	61.97	45.15	55.62	62.91	61.77	74.62	65.19$_{\pm 12.01}$
text-embedding-3-small	1536	72.88	88.19	60.16	52.31	39.34	43.12	65.18	52.87	71.25	60.59$_{\pm 14.65}$

Table 6: Task-model performance view, where each cell reports scores averaged over all evaluated languages.

hibit inconsistent performance across different languages. Although multilingual-e5-large-instruct might perform best overall, we found that no model can perform best for all languages. We observe that while multilingual-e5-large-instruct performs well on Burmese, Khmer, Malay, and Lao, Qwen3-Embedding-8B performs well on Thai and Vietnamese, bge-multilingual-gemma2 performs well on Filipino and Tamil, and GritLM-7B performs well on Indonesian and Tetum. This emphasizes the inconsistency of model performance across languages, thus rendering the overall evaluation results inconclusive for all models. Making the embedding model consistent to support all languages equally is a challenge and remains an open question in the field of text understanding. Note that we also experimented with tokenizer and language similarity to understand the underrepresented languages in Appendix E and Appendix F, respectively.

4.2 Task-Model Comparisons

To understand the performance of each task, we ask which tasks remain particularly challenging for state-of-the-art models across SEA languages.

Results. As shown in Table 6, the experiment results demonstrate that multilingual-e5-large-instruct performs the best on our benchmark, achieving 75.24 points on the average score. The performance of the second-best model (Qwen3-Embedding-8B) is lower than that of multilingual-e5-large-instruct by only 0.06 points on average, with a 70-fold difference in model parameters (560M vs. 8B parameters). Moreover, we found that, although Gemma-SEA-LION-v3 and Sailor2 were specifically trained for SEA languages, the models did not perform well on our text embedding benchmark due to their design for generation, rather than embedding purposes. For the proprietary models, in contrast to previous works (Muennighoff et al., 2023), which found that proprietary models outperformed open-source models, we found that all proprietary models perform lower than multilingual-e5-large-instruct and Qwen3-Embedding-8B. This suggests that all proprietary models may be primarily trained in English and not optimized for SEA languages.

Discussion. We found that task performance consistency is the main challenge for current text embedding models. In particular, a robust model should perform well on all tasks. As shown in Table 6, we found that there is no dominant model that achieves the highest score on all tasks. Notably, model performance varies considerably depending on the task. *This emphasizes that the task consistency problem in our benchmark is still challenging for embedding models*. In short, when using multilingual text embedding in SEA languages, it is essential to select the model based on the specific task at hand, as there is no all-purpose model that suits every solution.

4.3 Language-Task Comparisons

Let us now turn our attention to language-task comparisons, averaging performance across all models.

Results. Table 7 presents our third comparative view: language-task. Substantial variation is observed across both dimensions, with no task exhibiting uniformly strong performance across all languages. For relatively well-studied tasks such as classification and bitext mining, performance is generally strong for higher-resource languages (ind and tha), while noticeable degradation persists for more resource-constrained languages, e.g., lao, mya, and tet. In contrast, clustering remains challenging across most languages, exhibiting lower and more variable performance, indicating that some task categories pose intrinsic difficulties that are not alleviated by language coverage alone. Bitext mining exhibits mixed behavior, with strong performance in some languages but substantial variation across others, highlighting how task difficulty interacts with language-specific factors.

Discussion. The results reveal a landscape of conditional behaviors across tasks and languages, underscoring the importance of disentangling evaluation dimensions for meaningful assessment.

Lang.	Clf	M. Clf	Pr. Clf	STS	Clust	Btxt	Rtrvl	In. Rtrvl	Rrnk
ind	77.77	92.49	64.51	61.04	43.55	75.65	72.71	34.82	69.28
tha	73.73	76.60	70.31	68.73	38.49	70.99	67.71	70.43	71.86
vie	78.93	96.49	63.67	77.15	38.56	76.36	56.80	42.75	-
mya	78.78	69.71	67.10	61.95	28.09	48.83	55.44	-	-
fil	72.78	90.06	50.63	68.41	52.69	69.56	-	-	-
khm	68.56	100.00	66.41	63.74	23.83	54.21	-	-	-
zsm	72.94	-	71.50	75.39	-	75.98	46.27	-	-
lao	70.55	-	64.36	59.11	18.23	51.79	-	-	-
tam	84.35	-	68.14	52.49	38.75	61.23	46.27	-	-
tet	99.81	-	-	-	-	45.75	-	-	-

Table 7: Language-task performance view, where each cell reports scores averaged over all evaluated models.

5 Insights for Future Model Development

From the main results and ablation studies, embedding models for SEA languages can be enhanced in three aspects: (i) datasets, (ii) training algorithm, and (iii) architecture.

Dataset. A common technique to improve downstream tasks is to introduce data aligned with the domain of the target task. However, in resource-constrained settings, model developers often have to resort to machine-assisted dataset generation. As shown in Appendix G, we examine the possibility of using MT on low-resource languages and found that we can use machine translation to translate from English to SEA languages with a marginal difference between human and machine translations. There are various English datasets that are not yet available in SEA languages, i.e., MSMACO (Nguyen et al., 2016b) and NQ (Kwiatkowski et al., 2019), and some datasets are only available in a subset of SEA languages, e.g., only Thai or Indonesian, like Mr.TyDi (Clark et al., 2020) and MIRACL (Zhang et al., 2023). As demonstrated by previous works, having these datasets in SEA languages will increase robust representation in embeddings.

Training Algorithm. From our ablation study in Appendix F, we found that there are a lot of false positive and negative occurrences during the testing of the models in Table 5. As shown in Figure 2a, the experimental results from multilingual-e5-large-instruct demonstrate that the similarity of positive pairs is high (more than 0.87 in all cases). However, the similarity of negative pairs is also high (ranging from 0.74 to 0.81), resulting in the overlap between positive and negative pairs (Figure 2c). Moreover, when the model performs worst in SEA-BED (multilingual-MiniLM-L12-v2), Figures 2b and 2d show that the contrast between positive and negative samples is better than robust models, where this model did not employ contrastive learning, unlike the SOTA model. This highlights the inconsistency of robust models, which necessitates immediate correction. To mitigate this, recent works (Limkonchotiwat et al., 2022; Li and Li, 2023; Wang et al., 2024c) demonstrate the possibility of contrasting positive and negative samples more effectively than vanilla contrastive learning (Gao et al., 2021), which is employed in current models. However, these techniques have not been well explored on SEA languages; the effects and failure cases need further study. Applying these techniques can mitigate the issue of overlap.

Architecture. Similar to the findings from previous works in Appendix E, the tokenizer plays a crucial role in embeddings. In our findings, we discovered that the current models' tokenizer does not include a Telugu token; adding Telugu would enhance the representation of these models. However, Appendix H also shows that adding tokens is not effective for all languages; in most cases, non-Latin scripts will have the most effect. We can also employ adding token techniques (Cui et al., 2024; Nguyen et al., 2024) by adding new tokens with minimal effort during continual pre-training to maintain previous knowledge and incorporate new knowledge into the model.

In summary, future work can benefit from the insights gained from our discussions. The fastest and

most cost-effective way is to obtain more datasets using machine translation. In particular, we can utilize models that demonstrate robust performance for English-to-SEA language translation on SEA translation benchmarks (Susanto et al., 2025), such as ChatGPT or Google Gemini. Then, we can focus on applying and adapting novel training objectives to SEA embeddings. We also need to study how to adapt from an English-centric to a SEA-centric approach, leaving a gap for future work to propose a new training objective for the multilingual scenario. Lastly, we can focus on the changes in architecture since this will require a new round of pre-training. However, not all languages will benefit from this change; we need to carefully add new tokens to the model, as previous studies have shown.

6 Related Work

6.1 Text Embedding Benchmarks

Existing text embedding benchmarks primarily focus on high-resource languages. Notable examples include SentEval (Conneau and Kiela, 2018), which provides a preliminary benchmark for understanding text embeddings in STS and transfer learning. USEB (Wang et al., 2021) is an unsupervised embedding benchmark focusing on pair-text classification. BEIR (Thakur et al., 2021) is a heterogeneous benchmark focusing only on 18 information retrieval datasets. MTEB (Muennighoff et al., 2023) is a large-scale version of BEIR that not only focuses on retrieval tasks but also on diverse tasks, i.e., bitext mining, classification, and semantic textual similarity. However, these benchmarks primarily focus on English, while many works extend MTEB from English to Chinese (Xiao et al., 2024b), German (Wehrli et al., 2023), and French (Ciancone et al., 2024b). Recently, an attempt has been made to create a multilingual version of MTEB, called MMTEB (Enevoldsen et al., 2025). MMTEB multilingual benchmark evaluates 10 tasks and 270 datasets; notably, only 22 of these datasets are from SEA languages. Thus, results from MMTEB might not be representative of performance in SEA languages, given the reliance on machine-translated datasets.

6.2 SEA Benchmarks

There have been many efforts to formulate SEA benchmarks. NusaCrowd (Cahyawijaya et al., 2023) proposed a large-scale Indonesian benchmark focusing on natural language understanding and generation, especially for decoder models. SEACrowd (Lovenia et al., 2024) and SEA-VL (Cahyawijaya et al., 2025) are data collection projects that gather SEA benchmarks in their own repositories. The experiment from SEA projects focuses primarily on large language models, particularly the Llama (Dubey et al., 2024) and T5 (Raffel et al., 2020) families. Moreover, these benchmarks do not accurately measure the effectiveness of embedding in SEA texts. In particular, previous works studied large language models and generative outputs, while embeddings have not been experimented with in SEA languages.

7 Conclusion

Our analyses show that multilingual embedding performance in SEA languages is highly conditional, varying across models, tasks, and language-task combinations. Language-model comparisons reveal that no single model consistently performs well across all languages, with substantial disparities persisting even among the strongest models. Task-model analyses reveal clear differences in task difficulty: while classification and bitext mining approach saturation for well-resourced languages, clustering and semantic similarity remain challenging and unstable. Language-task comparisons demonstrate uneven performance within individual languages, showing that success on one task does not reliably generalize to others.

Our results suggest that observed performance gaps arise from interrelated limitations in data coverage, training objectives, and architectural design. **Dataset availability** plays a central role: models trained predominantly on English-centric or weakly aligned multilingual data struggle to generalize across languages and tasks. **Training algorithms** need to address the high positive-negative similarity overlap in low-resource languages, suggesting the application of cross-lingual transfer to ensure that task-relevant semantic structures learned in one language generalize to others. **Architectural factors**, including tokenizer design and language coverage, introduce additional structural constraints that disproportionately affect non-Latin and underrepresented languages. Consequently, improving multilingual embeddings for SEA languages requires coordinated advances across datasets, training algorithms, and architectures.

Limitations

While SEA-BED substantially expands the landscape of multilingual sentence-embedding evaluation, several limitations remain.

First, the coverage is uneven across the 10 SEA languages. Although the benchmark encompasses the region's major language families and scripts, some languages are represented in fewer task categories due to the limited availability of publicly accessible datasets. This asymmetry limits the breadth of task-language combinations that can be examined uniformly.

Second, the scope of evaluation data is regionally bounded. SEA-BED focuses on locally grounded and native-verified data from Southeast Asian linguistic communities. While this enables strong internal validity for SEA-specific probing, it does not capture cultural or pragmatic phenomena unique to other low-resource regions, nor does it fully represent global diversity.

Finally, the benchmark incorporates machine-generated and machine-derived data. While this supports broader task coverage and scalability, such data may differ systematically from human-authored text. Given persistent coverage constraints, the inclusion of machine-derived data is often unavoidable in multilingual evaluation. Accordingly, we distinguish evaluation results by data source where applicable to examine differences between human-authored and machine-derived conditions in Appendix G. However, we do not conduct controlled studies isolating the causal effects of machine generation, nor do we claim that machine-derived data uniformly approximates human-authored data across tasks.

Ethical Statement

For the annotator details, we hired annotators (graduated students) who speak SEA languages natively (see Appendix A for more details). We first ran the annotation experiment and selected only the annotators who passed the annotation test, i.e., the English test and NLP understanding, to test whether annotators understand and can perform work in a high-quality manner. In addition, the payment rate for each annotator is 18 USD/Hr, which is considered higher than the average payment.

References

M. L. Khodra A. N. Azhar and A. P. Sutiono. 2019. Multi-label aspect categorization with convolutional neural networks and extreme gradient boosting. In *Proceedings of the 2019 International Conference on Electrical Engineering and Informatics (ICEEI)*, pages 35–40.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2023. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2024. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects.

Divyanshu Aggarwal, Vivek Gupta, and Anoop Kunchukuttan. 2022. Indicxnli: Evaluating multilingual inference for indian languages.

Pawitsapak Akarajaradwong, Pirat Pothavorn, Chompakorn Chaksangchaichot, Panuthep Tasawong, Thitiwat Nopparatbundit, and Sarana Nutanong. 2025. Nitibench: A comprehensive studies of llm frameworks capabilities for thai legal question answering.

Vesa Akerman, David Baines, Damien Daspit, Ulf Hermjakob, Taeho Jang, Colin Leong, Michael Martin, Joel Mathew, Jonathan Robie, and Marcus Schwarting. 2023. The ebible corpus: Data and model benchmarks for bible translation for low-resource languages. *arXiv preprint arXiv:2304.09919*.

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the indonesian language: A dataset and preliminary study.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Schulze Buschhoff, Charvi Jain, Alexander Arno Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, Stefan Kesselheim, and Nicolas Flores-Herr. 2024. Tokenizer choice for LLM training: Negligible or crucial? In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 3907–3924. Association for Computational Linguistics.

Samuel Cahyawijaya Arfinda Ilmania, Abdurrahman and Ayu Purwarianti. 2018. Aspect detection and sentiment classification using deep neural network for indonesian aspect-based sentiment analysis. In *Proceedings of the 2018 International Conference on Asian Language Processing(IALP)*, pages 62–67. IEEE.

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 6607–6623. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Laksmita Widya Astuti, Yunita Sari, and Suprapto. 2023. Code-mixed sentiment analysis using transformer for twitter social media data. *International Journal of Advanced Computer Science and Applications*, 14(10).

Nofa Aulia and Indra Budi. 2019. Hate speech detection on indonesian long text documents using machine learning approach. In *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence*, ICCAI '19, page 164–169, New York, NY, USA. Association for Computing Machinery.

Thura Aung, Eaint Kay Khaing Kyaw, and Ye Kyaw Thu. 2024. An empirical study of simple text augmentation on news classification for myanmar language. Preprint, Language Understanding Laboratory, Myanmar.

Thura Aung and Pyi Hein San. 2025. Askcoviddrbot: Retrieval based tf-idf english and burmese bilingual chatbot for covid-19 domain. GitHub repository.

Bianka Buschbeck and Miriam Exel. 2020. A parallel evaluation data set of software documentation with document structure annotation.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Indra Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Halim Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Akbar Septiandri, James Jaya, Kaustubh D. Dhole, Arie Ardiyanti Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Farid Adilazuarda, Ryan Hadiwijaya, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapuspita, Haryo Akbarianto Wibowo, Cuk Tho, Ichwanul Muslim Karo Karo, Tirana Fatyanosa, Ziwei Ji, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Pascale Fung, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2023. Nusacrowd: Open source initiative for indonesian NLP resources. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13745–13818. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Joel Ruben Antony Moniz, Tack Hwa Wong, Mohammad Rifqi Farhansyah, Thant Thiri Maung, Frederikus Hudi, David Anugraha, Muhammad Ravi Shulthan Habibi, Muhammad Reza Qorib, Amit Agarwal, Joseph Marvin Imperial, Hitesh Laxmichand Patel, Vicky Feliren, Bahrul Ilmi Nasution, Manuel Antonio Rufino, Genta Indra Winata, Rian Adam Rajagede, Carlos Rafael Catalan, Mohamed Fazli Mohamed Imam, Priyaranjan Pattnayak, Salsabila Zahirah Pranida, Kevin Pratama, Yeshil Bangera, Adisai Na-Thalang, Patricia Nicole Monderin, Yueqi Song, Christian Simon, Lynnette Hui Xian Ng, Richardy Lobo Sapan, Taki Hasan Rafi, Bin Wang, Supryadi, Kanyakorn Veerakanjana, Piyalitt Ittichaiwong, Matthew Theodore Roque, Karissa Vincentio, Takdanai Kreangphet, Phakphum Artkaew, Kadek Hendrawan Palgunadi, Yanzhi Yu, Rochana Prih Hastuti, William Nixon, Mithil Bangera, Adrian Xuan Wei Lim, Aye Hninn Khine, Hanif Muhammad Zhafran, Teddy Ferdinan, Audra Aurora Izzani, Ayushman Singh, Evan Evan, Jauza Akbar Krito, Michael Anugraha, Fenal Ashokbhai Ilasariya, Haochen Li, John Amadeo Daniswara, Filbert Aurelian Tjiaranata, Eryawan Presma Yulianrifat, Can Udomcharoenchaikit, Fadil Risdian Ansori, Mahardika Krisna Ihsani, Giang Nguyen, Anab Maulana Barik, Dan John Velasco, Rifo Ahmad Genadi, Saptarshi Saha, Chengwei Wei, Isaiah Edri W. Flores, Kenneth Chen Ko Han, Anjela Gail D. Santos, Wan Shen Lim, Kaung Si Phyo, Tim Santos, Meisyarah Dwiastuti, Jiayun Luo, Jan Christian Blaise Cruz, Ming Shan Hee, Ikhlasul Akmal Hanif, M.Alif Al Hakim, Muhammad Rizky Sya'ban, Kun Kerdthaisong, Lester James Validad Miranda, Fajri Koto, Tirana Noor Fatyanosa, Alham Fikri Aji, Jostin Jerico Rosal, Jun Kevin, Robert Wijaya, Onno P. Kampman, Ruochen Zhang, Börje F. Karlsson, and Peerat Limkonchotiwat. 2025. Crowdsource, crawl, or generate? creating SEA-VL, a multicultural vision-language dataset for Southeast Asia. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18685–18717, Vienna, Austria. Association for Computational Linguistics.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 2021. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jasper Kyle Catapang and Moses Visperas. 2023. Emotion-based morality in Tagalog and English scenarios (EMoTES-3K): A parallel corpus for explaining (im)morality of actions. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pages 1–6, Tokyo, Japan. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Andreas Chandra. 2020. Indonesian news dataset. Online. Accessed: 2024-02-13.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, and Mattia Samory. 2022. SemEval-2022 task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106, Seattle, United States. Association for Computational Linguistics.

Jirat Chiaranaipanich, Naiyarat Hanmatheekuna, Jitkapat Sawatphol, Krittamate Tiankanon, Jiramet Kinchagawat, Amrest Chinkamol, Parinthapat Pengpun, Piyalitt Ittichaiwong, and Peerat Limkonchotiwat. 2024. Can general-purpose large language models generalize to english-thai machine translation ?

Antonius Rachmat Chrismanto, Anny Kartika Sari, and Yohanes Suyanto. 2022. Spamid-pair: A novel indonesian post–comment pairs dataset containing emoji. *International Journal of Advanced Computer Science and Applications*, 13(11).

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218, Seattle, United States. Association for Computational Linguistics.

Mathieu Ciancone, Imene Kerboua, Marion Schaeffer, and Wissam Siblini. 2024a. Extending the massive text embedding benchmark to french. *CoRR*, abs/2405.20468.

Mathieu Ciancone, Imene Kerboua, Marion Schaeffer, and Wissam Siblini. 2024b. Mteb-french: Resources for french sentence embedding evaluation and analysis.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*.

Tatoeba community. 2021. Tatoeba: Collection of sentences and translations.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

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, May 7-12, 2018*. European Language Resources Association (ELRA).

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jan Christian Blaise Cruz, Jose Kristian Resabal, James Lin, Dan John Velasco, and Charibeth Cheng. 2020a. Investigating the true performance of transformers in low-resource languages: A case study in automatic corpus creation. *arXiv preprint arXiv:2010.11574*.

Jan Christian Blaise Cruz, Julianne Agatha Tan, and Charibeth Cheng. 2020b. Localization of fake news detection via multitask transfer learning. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2596–2604.

lukkiddd cstorm125. 2019. prachathai67k. `https://github.com/PyThaiNLP/prachathai-67k`.

Yiming Cui, Ziqing Yang, and Xin Yao. 2024. Efficient and effective text encoding for chinese llama and alpaca.

Hoang-Quan Dang, Duc-Duy-Anh Nguyen, and Trong-Hop Do. 2022. Multi-task solution for aspect category sentiment analysis on vietnamese datasets. In *2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, pages 404–409.

Mai Hoang Dao, Thinh Hung Truong, and Dat Quoc Nguyen. 2021. Intent Detection and Slot Filling for Vietnamese. In *Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreyansh Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2022. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. *Annual Meeting of the Association for Computational Linguistics*.

Longxu Dou, Qian Liu, Fan Zhou, Changyu Chen, Zili Wang, Ziqi Jin, Zichen Liu, Tongyao Zhu, Cunxiao Du, Penghui Yang, Haonan Wang, Jiaheng Liu, Yongchi Zhao, Xiachong Feng, Xin Mao, Man Tsung Yeung, Kunat Pipatanakul, Fajri Koto, Min Si Thu, Hynek Kydlíček, Zeyi Liu, Qunshu Lin, Sittipong Sripaisarnmongkol, Kridtaphad Sae-Khow, Nirattisai Thongchim, Taechawat Konkaew, Narong Borijindargoon, Anh Dao, Matichon Maneegard, Phakphum Artkaew, Zheng-Xin Yong, Quan Nguyen, Wannaphong Phatthiyaphaibun, Hoang H. Tran, Mike Zhang, Shiqi Chen, Tianyu Pang, Chao Du, Xinyi Wan, Wei Lu, and Min Lin. 2025. Sailor2: Sailing in south-east asia with inclusive multilingual llm. *arXiv preprint arXiv:2502.12982*.

Kerenza Doxolodeo and Adila Alfa Krisnadhi. 2024. Ac-iquad: Automatically constructed indonesian question answering dataset by leveraging wikidata. *Language Resources and Evaluation*. Publisher Copyright: extcopyright 2024, The Author(s).

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Kenneth C. Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzeminski, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çagatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafal Poswiata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Suppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. 2025. MMTEB: massive multilingual text embedding benchmark. *CoRR*, abs/2502.13595.

Kenneth C. Enevoldsen, Márton Kardos, Niklas Muennighoff, and Kristoffer L. Nielbo. 2024. The scandinavian embedding benchmarks: Comprehensive assessment of multilingual and monolingual text embedding. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Ridi Fe. 2019. Indonesia sentiment analysis dataset. https://github.com/ridife/dataset-idsa.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 878–891. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan,

Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

H Fujita and H Perez-Meana. 2021. An empirical investigation of online news classification on an open-domain, large-scale and high-quality dataset in vietnamese. In *New Trends in Intelligent Software Methodologies, Tools and Techniques: Proceedings of the 20th International Conference on New Trends in Intelligent Software Methodologies, Tools and Techniques (SoMeT_21)*, volume 337, page 367. SAGE Publications Limited.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Valfrid Galinato, Lawrence Amores, Gino Ben Magsino, and David Rafael Sumawang. 2023. Context-based profanity detection and censorship using bidirectional encoder representations from transformers.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, and Francisco Guzmán. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 19–35.

Tri Wahyu Guntara, Alham Fikri Aji, and Radityo Eko Prasojo. 2020. Benchmarking multidomain English-Indonesian machine translation. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 35–43, Marseille, France. European Language Resources Association.

Mika Hämäläinen, Pattama Patpong, Khalid Alnajjar, Niko Partanen, and Jack Rueter. 2021. Detecting depression in Thai blog posts: a dataset and a baseline. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 20–25, Online. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703.

Rommel Hernandez Urbano Jr, Jeffrey Uy Ajero, Angelic Legaspi Angeles, Maria Nikki Hacar Quintos, Joseph Marvin Regalado Imperial, and Ramon Llabanes Rodriguez. 2021. A bert-based hate speech classifier from transcribed online short-form videos. In *2021 5th International Conference on E-Society, E-Education and E-Technology*.

Ahmad Fathan Hidayatullah, Siwi Cahyaningtyas, and Rheza Daffa Pamungkas. 2020. Attention-based cnn-bilstm for dialect identification on javanese text. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, pages 317–324.

Vong Anh Ho, Duong Huynh-Cong Nguyen, Danh Hoang Nguyen, Linh Thi-Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2020. Emotion recognition for vietnamese social media text. In *Computational Linguistics: 16th International Conference of the Pacific Association for Computational Linguistics, PACLING 2019, Hanoi, Vietnam, October 11–13, 2019, Revised Selected Papers 16*, pages 319–333. Springer.

Aung Kyaw Htet and Mark Dras. 2024. Myanmar xnli: Building a dataset and exploring low-resource approaches to natural language inference with myanmar. PREPRINT (Version 1) available at Research Square.

Muhammad Okky Ibrohim and Indra Budi. 2018. A dataset and preliminaries study for abusive language detection in indonesian social media. *Procedia Computer Science*, 135:222–229. The 3rd International Conference on Computer Science and Computational Intelligence (ICCSCI 2018) : Empowering Smart Technology in Digital Era for a Better Life.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57, Florence, Italy. Association for Computational Linguistics.

Ahmad Izzan, Christian Wibisono, and Ilham Firdausi Putra. 2025. Netifier: Negativity classifier. GitHub repository.

Jakarta Artificial Intelligence Research. 2023. Indoqa: Building indonesian qa dataset.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock,

Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Shengyi Jiang, Sihui Fu, Nankai Lin, and Yingwen Fu. 2021a. Pre-trained models and evaluation data for the khmer language. *Tsinghua Science and Technology*.

Shengyi Jiang, Sihui Fu, Nankai Lin, and Yingwen Fu. 2022. Pretrained models and evaluation data for the khmer language. *Tsinghua Science and Technology*, 27(4):709–718.

Shengyi Jiang, Xiuwen Huang, Xiaonan Cai, and Nankai Lin. 2021b. Pre-trained models and evaluation data for the myanmar language. In *The 28th International Conference on Neural Information Processing*, Cham. Springer International Publishing.

Sarah Samson Juan, Suhaila Saee, and Fitri Suraya Mohamad. 2022. Social versus physical distancing: Analysis of public health messages at the start of covid-19 outbreak in malaysia using natural language processing. In *Proceedings of the 8th International Conference on Computational Science and Technology*, volume 835 of *Lecture Notes in Electrical Engineering*, pages 577–589. Springer Singapore.

A. H. Khine, K. T. Nwet, and K. M. Soe. 2017. Automatic myanmar news classification. In *15th Proceedings of International Conference on Computer Applications*, pages 401–408.

Dhamir Raniah Kiasati Desrul and Ade Romadhony. 2019. Abusive language detection on indonesian online news comments. In *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pages 320–325.

Fajri Koto and Ikhwan Koto. 2020. Towards computational linguistics in minangkabau language: Studies on sentiment analysis and machine translation. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, Vietnam.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2020a. Liputan6: A large-scale indonesian dataset for text summarization. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 598–608.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020b. Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian NLP. *CoRR*, abs/2011.00677.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2020. Qed: A framework and dataset for explanations in question answering.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *Preprint*. Publisher: Open Science Framework.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. XLM-V: overcoming the vocabulary bottleneck in multilingual masked language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 13142–13152. Association for Computational Linguistics.

Peerat Limkonchotiwat, Wuttikorn Ponwitayarat, Lalita Lowphansirikul, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2022. ConGen: Unsupervised control and generalization distillation for sentence representation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6467–6480, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

E. D. Livelo and C. Cheng. 2018. Intelligent dengue infoveillance using gated recurrent neural learning and cross-label frequencies. In *2018 IEEE International Conference on Agents (ICA)*, pages 2–7.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Jann Montalan, Ryan Hadiwijaya, Joanito Agili Lopo, William Nixon, Börje Karlsson, James Jaya, Ryandito Diandaru, Yuze Gao, Patrick Amadeus Irawan, Bin Wang, Jan Christian Blaise Cruz, Chenxi Whitehouse, Ivan Halim Parmonangan, Maria Khelli, Wenyu Zhang, Lucky

Susanto, Reynard Adha Ryanda, Sonny Lazuardi Hermawan, Dan John Velasco, Muhammad Dehan Al Kautsar, Willy Fitra Hendria, Yasmin Moslem, Noah Flynn, Muhammad Farid Adilazuarda, Haochen Li, Johanes Lee, R. Damanhuri, Shuo Sun, Muhammad Reza Qorib, Amirbek Djanibekov, Wei Qi Leong, Quyet V. Do, Niklas Muennighoff, Tanrada Pansuwan, Ilham Firdausi Putra, Yan Xu, Ngee Tai Chia, Ayu Purwarianti, Sebastian Ruder, William-Chandra Tjhi, Peerat Limkonchotiwat, Alham Fikri Aji, Sedrick Keh, Genta Indra Winata, Ruochen Zhang, Fajri Koto, Zheng Xin Yong, and Samuel Cahyawijaya. 2024. Seacrowd: A multilingual multimodal data hub and benchmark suite for southeast asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5155–5203. Association for Computational Linguistics.

Lalita Lowphansirikul, Charin Polpanumas, Attapol T. Rutherford, and Sarana Nutanong. 2022. A large english-thai parallel corpus from the web and machine-generated text. *Lang. Resour. Evaluation*, 56(2):477–499.

Luong Luc Phan, Phuc Huynh Pham, Kim Thi-Thanh Nguyen, Sieu Khai Huynh, Tham Thi Nguyen, Luan Thanh Nguyen, Tin Van Huynh, and Kiet Van Nguyen. 2021. Sa2sl: From aspect-based sentiment analysis to social listening system for business intelligence. In *Knowledge Science, Engineering and Management*, pages 647–658, Cham. Springer International Publishing.

Son T. Luu, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. A large-scale dataset for hate speech detection on vietnamese social media texts. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices*, pages 415–426, Cham. Springer International Publishing.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Rahmad Mahendra, Alham Fikri Aji, Samuel Louvan, Fahrurrozi Rahman, and Clara Vania. 2021. IndoNLI: A natural language inference dataset for Indonesian. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10511–10527, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rahmad Mahendra Mei Silviana Saputri and Mirna Adriani. 2018. Emotion classification on indonesian twitter dataset. In *Proceedings of the 2018 International Conference on Asian Language Processing(IALP)*, pages 90–95. IEEE.

Min Si Thu,Khin Myat Noe. Myanmar-agriculture-1k.

Sepideh Mollanorozy, Marc Tanti, and Malvina Nissim. 2023. Cross-lingual transfer learning with Persian. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 89–95, Dubrovnik, Croatia. Association for Computational Linguistics.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning.

Niklas Muennighoff, Nouamane Tazi, Loïc 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, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2006–2029. Association for Computational Linguistics.

Huyen TM Nguyen, Hung V Nguyen, Quyen T Ngo, Luong X Vu, Vu Mai Tran, Bach X Ngo, and Cuong A Le. 2018a. Vlsp shared task: sentiment analysis. *Journal of Computer Science and Cybernetics*, 34(4):295–310.

Kiet Nguyen, Son Quoc Tran, Luan Thanh Nguyen, Tin Van Huynh, Son Thanh Luu, and Ngan Luu-Thuy Nguyen. 2022. Vlsp 2021-vimrc challenge: Vietnamese machine reading comprehension. *VNU Journal of Science: Computer Science and Communication Engineering*, 38(2).

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

Luan Thanh Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021a. Constructive and toxic speech detection for open-domain social media comments in vietnamese. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices*, pages 572–583, Cham. Springer International Publishing.

Minh-Tien Nguyen, Dac Viet Lai, Phong-Khac Do, Duc-Vu Tran, and Minh-Le Nguyen. 2016a. VSoLSC-Sum: Building a Vietnamese sentence-comment dataset for social context summarization. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 38–48, Osaka, Japan. The COLING 2016 Organizing Committee.

Nhung Thi-Hong Nguyen, Phuong Phan-Dieu Ha, Luan Thanh Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021b. Vietnamese complaint detection on e-commerce websites.

Phu-Vinh Nguyen, Minh-Nam Tran, Long Nguyen, and Dien Dinh. 2025. Advancing vietnamese information retrieval with learning objective and benchmark.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016b. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs - large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.

Tran Nhiem. 2023. Vietnamese instruction data corpus for large-scale finetuning of language models.

Hiroki Nomoto, Kenji Okano, David Moeljadi, and Hideo Sawada. 2018. Tufs asian language parallel corpus (talpco). pages 436–439.

Hiroki Nomoto, Kenji Okano, Sunisa Wittayapanyanon, and Junta Nomura. 2019. Interpersonal meaning annotation for asian language corpora: The case of tufs asian language parallel corpus (talpco). pages 846–849.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Abinew Ali Ayele, Pavan Baswani, Meriem Beloucif, Chris Biemann, Sofia Bourhim, Christine De Kock, Genet Shanko Dekebo, Oumaima Hourrane, Gopichand Kanumolu, Lokesh Madasu, Samuel Rutunda, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Hailegnaw Getaneh Tilaye, Krishnapriya Vishnubhotla, Genta Winata, Seid Muhie Yimam, and Saif M. Mohammad. 2024a. Semrel2024: A collection of semantic textual relatedness datasets for 13 languages. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Meriem Beloucif, Christine De Kock, Oumaima Hourrane, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Krishnapriya Vishnubhotla, Seid Muhie Yimam, and Saif M. Mohammad. 2024b. SemEval-2024 task 1: Semantic textual relatedness for african and asian languages. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization.

Kitsuchart Pasupa, Ponruedee Netisopakul, and Ratthawut Lertsuksakda. 2016. Sentiment analysis on thai children stories. *Artificial Life and Robotics*, 21(3):357–364.

Patomporn Payoungkhamdee, Peerachet Porkaew, Atthasith Sinthunyathum, Phattharaphon Songphum, Witsarut Kawidam, Wichayut Loha-Udom, Prachya Boonkwan, and Vipas Sutantayawalee. 2021. Limesoda: Dataset for fake news detection in healthcare domain. In *2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–6.

Parinthapat Pengpun, Can Udomcharoenchaikit, Weerayut Buaphet, and Peerat Limkonchotiwat. 2024. Seed-free synthetic data generation framework for instruction-tuning LLMs: A case study in Thai. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 438–457, Bangkok, Thailand. Association for Computational Linguistics.

Wannaphong Phatthiyaphaibun. 2020. Pythainlp/thai-lao-parallel-corpus: Thai lao parallel corpus v0.5.

Wannaphong Phatthiyaphaibun. 2025. Lao news classification.

Wannaphong Phatthiyaphaibun, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita Lowphansirikul, Pattarawat Chormai, Peerat Limkonchotiwat, Thanathip Suntorntip, and Can Udomcharoenchaikit. 2023. PyThaiNLP: Thai natural language processing in Python. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 25–36, Singapore, Singapore. Empirical Methods in Natural Language Processing.

Inggrid Yanuar Risca Pratiwi, Rosa Andrie Asmara, and Faisal Rahutomo. 2017. Study of hoax news detection using naïve bayes classifier in indonesian language. In *2017 11th International Conference on Information, Communication Technology and System (ICTS)*, pages 73–78.

Ayu Purwarianti and Ida Ayu Putu Ari Crisdayanti. 2019. Improving bi-lstm performance for indonesian sentiment analysis using paragraph vector. In *Proceedings of the 2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–5. IEEE.

Rifki Afina Putri and Alice Oh. 2022. IDK-MRC: Unanswerable questions for Indonesian machine reading comprehension. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6918–6933, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Deepak Kumar, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Trans. Assoc. Comput. Linguistics*, 10:145–162.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Riccosan and Karen Etania Saputra. 2023. Multilabel multiclass sentiment and emotion dataset from indonesian mobile application review. *Data in Brief*, 50.

Riccosan, Karen Etania Saputra, Galih Dea Pratama, and Andry Chowanda. 2022. Emotion dataset from indonesian public opinion. *Data in Brief*, 43:108465.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thái, Rapid Sun, Vichet Chea, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding. 2019. Asian language treebank. In *Proceedings of O-COCOSDA*. National Institute of Information and Communication Technology (NICT), Japan.

Muhammad Razif Rizqullah, Ayu Purwarianti, and Alham Fikri Aji. 2023. Qasina: Religious domain question answering using sirah nabawiyah. In *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, pages 1–6.

Ken Nabila Setya and Rahmad Mahendra. 2018. Semi-supervised textual entailment on indonesian wikipedia data. In *Proceedings of the 2018 International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*.

M. Si Thu. 2024. Burmese microbiology 1k dataset (1.1).

AI Singapore. 2024. Sea-lion (southeast asian languages in one network): A family of large language models for southeast asia. `https://github.com/aisingapore/sealion`.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Sebastian Ruder, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation.

Rohayani Sitepu et al. 2024. Sentiment analysis in karonese tweet using machine learning algorithms. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, 12(4):2482–2489.

Gizem Soğancıoğlu, Hakime "Ozt"urk, and Arzucan "Ozg"ur. 2017. Biosses: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024a. jina-embeddings-v3: Multilingual embeddings with task lora. *CoRR*, abs/2409.10173.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas,

Nan Wang, and Han Xiao. 2024b. jina-embeddings-v3: Multilingual embeddings with task lora.

Arthit Suriyawongkul, Ekapol Chuangsuwanich, Pattarawat Chormai, and Charin Polpanumas. 2019. Pythainlp/wisesight-sentiment: First release.

Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Montalan, Jian Gang Ngui, Xian Bin Yong, Wei Qi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Yifan Mai, and William-Chandra Tjhi. 2025. SEA-HELM: southeast asian holistic evaluation of language models. *CoRR*, abs/2502.14301.

Gemma Team. 2024. Gemma.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

C Tho, Y Heryadi, L Lukas, and A Wibowo. 2021. Code-mixed sentiment analysis of indonesian language and javanese language using lexicon based approach. *Journal of Physics: Conference Series*, 1869(1):012084.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Kanokorn Trakultaweekoon, Santipong Thaiprayoon, Pornpimon Palingoon, and Anocha Rugchatjaroen. 2019. The first wikipedia questions and factoid answers corpus in the thai language. In *2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–4. IEEE.

Co Van Dinh, Son T. Luu, and Anh Gia-Tuan Nguyen. 2022. Detecting spam reviews on vietnamese e-commerce websites. In *Intelligent Information and Database Systems*, pages 595–607, Cham. Springer International Publishing.

Kobkrit Viriyayudhakorn and Charin Polpanumas. 2021. iapp_wiki_qa_squad.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. TSDAE: using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. *CoRR*, abs/2104.06979.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Xinghao Wang, Junliang He, Pengyu Wang, Yunhua Zhou, Tianxiang Sun, and Xipeng Qiu. 2024c. Denosent: A denoising objective for self-supervised sentence representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19180–19188.

Silvan Wehrli, Bert Arnrich, and Christopher Irrgang. 2023. German text embedding clustering benchmark. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 187–201, Ingolstadt, Germany. Association for Computational Linguistics.

Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. 2024. Followir: Evaluating and teaching information retrieval models to follow instructions.

Andika William and Yunita Sari. 2020. Click-id: A novel dataset for indonesian clickbait headlines.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024a. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 641–649. ACM.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024b. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 641–649, New York, NY, USA. Association for Computing Machinery.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang,

Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Evi Yulianti, Ajmal Kurnia, Mirna Adriani, and Yoppy Setyo Duto. 2021. Normalisation of indonesian-english code-mixed text and its effect on emotion classification. *International Journal of Advanced Computer Science and Applications*.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Appendix

A Annotator Demographics and Guidelines

Data Assurance Annotators We hired two native speakers of each SEA language to check the quality of the datasets before adding them to SEA-BED. These annotators are graduate students who have passed the English test and the NLP test (i.e., understanding the concepts of each task in SEA-BED). We give them the guidelines as follows. *Please re-check the datasets that (i) the correctness of text and written style is natural and understandable for a native speaker; (ii) the correctness of the gold label, e.g., a correct class of text classification is assigned.* We have asked them to check 182 datasets. There are some datasets that utilize machine translation with low-quality outputs or poor-quality labels; we remove them from our SEA-BED.

New Dataset Annotators In this work, our collaborators helped us translate the data from English to Thai and Burmese for STS and NLI tasks. These people are Thai and Burmese undergraduate and graduate students studying in Thailand, aged from 20 to 25 years old, who can speak English and their native language (Thai or Burmese). We use three Thai annotators and one Burmese annotator to create new datasets. We also removed some examples that contain special characters that cannot be shown in Google Sheets. We give them the guidelines as follows. *Translate the selected datasets to make them a human-like or everyday conversation in your native languages* and *change the subject of a sentence to be gender-neutral* since both the Thai and Burmese languages have words or morphemes that can express the gender of the speaker. Therefore, the quality of our new human-crafted dataset is higher than that of using machine translations or LLMs to generate data, as such methods have been observed to be less native-like or unrepresentative of natural language use (Lovenia et al., 2024; Singh et al., 2025).

B Benchmark Efficiency

Caching Embeddings. To improve the run-time efficiency, we use embedding caching to store embedded texts in memory and cache files; when seen texts are input to the same model, we will use the cached embedding instead of computing the new one to decrease the run-time of our benchmark.

Downsampling. Enevoldsen et al. (2025) proposed a downsampling technique for the English benchmark, decreasing the number of samples by 98%. However, as shown in Table 8, we applied the same technique to our benchmark (bitext mining datasets) and found that the performance of each model increased in all cases. This is because all challenging samples may have been removed from the dataset, leading to improved performance for most models. Moreover, the ranking of each model changed, in contrast to the findings of Enevoldsen et al. (2025), where the rankings remained largely unchanged. Therefore, we did not apply the downsampling technique to our benchmark.

Model	100% Dataset	30% Dataset	Rank after downsampling
multilingual-e5-large-instruct (560M)	87.86	93.03	0
Qwen3-Embedding-8B (8B)	84.78	90.31	↓1
bge-multilingual-gemma2 (9B)	82.02	90.71	↑3
multilingual-e5-large (560M)	84.51	88.19	↓1
bge-m3 (568M)	86.18	91.89	↑1
GritLM-7B (7B)	63.63	69.68	0
e5-mistral-7b-instruct (7B)	65.30	73.42	0
Qwen3-Embedding-0.6B (595M)	56.53	62.95	0
multilingual-mpnet-base (278M)	68.12	73.97	0
LaBSE (471M)	86.84	90.51	↓2
multilingual-MiniLM-L12 (118M)	53.23	59.06	0
Gemma-SEA-LION-v3-9B-IT (9B)	15.31	3.21	↓1
Sailor2-8B-Chat (8B)	4.31	6.01	↑1

Table 8: We evaluate 13 models on bitext mining using 100% and 30% dataset sizes. We also indicate the rank change of the model before and after downsampling to show the performance discrepancy.

C Domains

For domains in the SEA-BED benchmark, we include the following:

- **Academic:** Formal writing and research publications commonly found in scholarly journals, theses, and dissertations.
- **Blog:** Informal, conversational writings about a variety of topics published on websites or personal blogs.
- **Constructed:** Artificially created text or speech, often in experiments to target particular abilities.
- **Encyclopedic:** Structured, reference-based texts offering thorough and factual information on various topics.
- **Fiction:** Narrative writing that involves creative content, such as novels, short stories, and other storytelling forms.
- **Government:** Documents, reports, and publications officially issued by government agencies.
- **Legal:** Documents and texts concerning laws, legal processes, contracts, and legal theories.
- **Medical:** Scientific and clinical publications focused on healthcare, treatments, patient care, and medical studies.
- **News:** News articles and reports that address current events, political developments, economic trends, and other timely topics.
- **Non-fiction:** Texts grounded in real events and factual information, including biographies, essays, and documentaries.
- **Religious:** Writings concerning religious teachings, doctrines, sacred texts, and discussions on spirituality.
- **Reviews:** Analytical assessments of books, films, music, products, or services.
- **Social:** Messages and conversations shared on social media, online forums, and other digital platforms.
- **Spoken:** Spoken content such as speeches, dialogues, interviews, and recorded discussions.
- **Subtitles:** Written transcriptions or translations of spoken content from films, videos, or multimedia presentations.
- **Web:** Web-based content spanning diverse topics, often featuring hyperlinks and multimedia elements.
- **Written:** A broad category encompassing all forms of text-based communication, both print and digital.

D Performance Changes Analysis

Here, we are examining how SEA-focused performance contrasts with the broader multilingual benchmark (MMTEB). To study the robustness of embeddings in world and SEA languages, we compare the ranking changes between our benchmark and the multilingual text embedding benchmark, MMTEB. We use the task average metric (Table 6), similar to MMTEB.

As shown in Figure 1, based on the experiment from MMTEB, Qwen3-Embedding-8B performed the best on world results, which includes 1,090 languages[2]. However, when we focus only on SEA languages using SEA-BED, the ranking of Qwen3-Embedding-8B dropped from first to second place. In addition, Qwen3-Embedding-0.6B dropped from second rank to eighth rank. This is because some of the linguistic and dialect knowledge will be different compared to other groups of languages, when we evaluate them only for the SEA languages. This highlights that the challenges, gaps, and model capabilities measured in MMTEB and our benchmark differ, particularly in the supported languages for embedding models that do not fully support SEA languages. Although some models are effective in performing well on MMTEB, they are not guaranteed to achieve the same performance for SEA languages.

[2] We obtained the model rankings on Nov 28th, 2025.

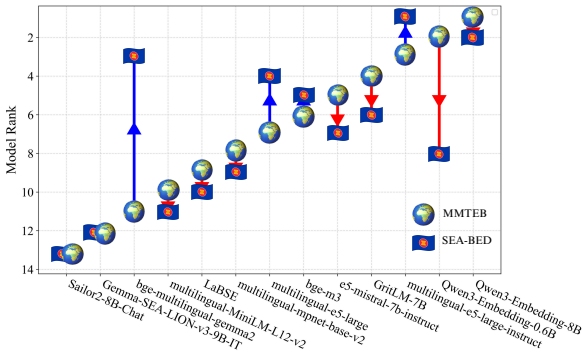


Figure 1: Ranking difference between MMTEB and SEA-BED.

E Tokenizer Analysis

This section examines whether limited coverage of SEA vocabulary in multilingual tokenizers correlates with poor downstream results. We highlight scripts like Lao or Khmer, which are often underrepresented in tokenizers. Previous works (Ali et al., 2024; Arnett and Bergen, 2025; Liang et al., 2023) demonstrated that vocabularies in a tokenizer affect the model performance in downstream tasks. In particular, when the multilingual tokenizer represents more vocabulary in some languages, the performance on those languages has also been observed to improve. In this study, we want to investigate whether the vocabulary in the tokenizer affects SEA-BED's overall performance or not. To answer this question, we count the SEA tokens in each text embedding model and compare their performance from Table 5.

As shown in Table 9, the language with the most tokens represented in a tokenizer is Filipino, with an average of 2.94 percent of vocabulary tokens in 13 models. However, compared to the language performance (Table 5), Filipino performance is lower than Indonesian. Surprisingly, there are no tokens for Tetum at all in the 13 models. We observe that performance on Tetum is also the worst compared to other SEA languages. Moreover, the performance is mixed for languages that do not use Latin characters, i.e., Thai, Burmese, Lao, and Tamil.

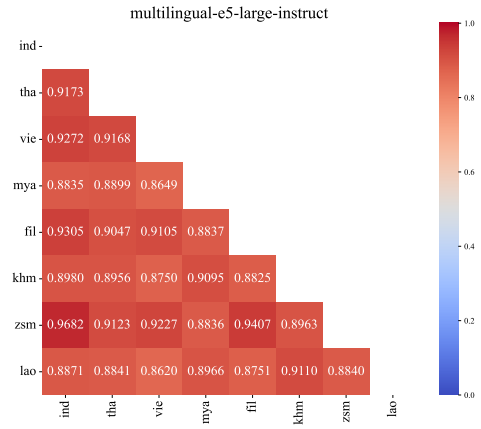
Model	ind	tha	vie	mya	fil	khm	zsm	lao	tam	tet
multilingual-e5-large-instruct (560M)	1.20	1.61	0.73	0.91	3.59	0.66	0.20	0.56	0.98	0.00
Qwen3-Embedding-8B (8B)	0.39	1.70	0.84	0.02	1.13	0.03	0.11	0.02	0.02	0.00
bge-multilingual-gemma2 (9B)	0.59	0.50	0.55	0.45	3.04	0.03	0.11	0.02	0.13	0.00
multilingual-e5-large (560M)	1.20	1.61	0.73	0.91	3.59	0.66	0.20	0.56	0.98	0.00
bge-m3 (568M)	1.20	1.61	0.73	0.91	3.59	0.66	0.20	0.56	0.98	0.00
GritLM-7B (7B)	0.27	0.19	0.55	0.45	3.04	0.03	0.11	0.02	0.13	0.00
multilingual-mpnet-base (278M)	1.20	1.61	0.73	0.91	3.59	0.66	0.20	0.56	0.98	0.00
LaBSE (471M)	1.12	0.45	0.81	0.45	4.65	0.54	0.19	0.29	1.28	0.00
e5-mistral-7b-instruct (7B)	0.27	0.19	0.55	0.45	3.04	0.03	0.11	0.02	0.13	0.00
Qwen3-Embedding-0.6B (595M)	0.39	1.70	0.84	0.02	1.13	0.03	0.11	0.02	0.02	0.00
multilingual-MiniLM-L12 (118M)	1.20	1.61	0.73	0.91	3.59	0.66	0.20	0.56	0.98	0.00
Gemma-SEA-LION-v3-9B-IT (9B)	0.59	0.50	0.55	0.45	3.04	0.03	0.11	0.02	0.13	0.00
Sailor2-8B-Chat (8B)	0.39	1.70	0.84	0.02	1.13	0.03	0.11	0.02	0.02	0.00
Average	0.77	1.15	0.71	0.53	2.94	0.31	0.15	0.25	0.52	0.00

Table 9: The percentage number of vocabulary tokens for each model in each language.

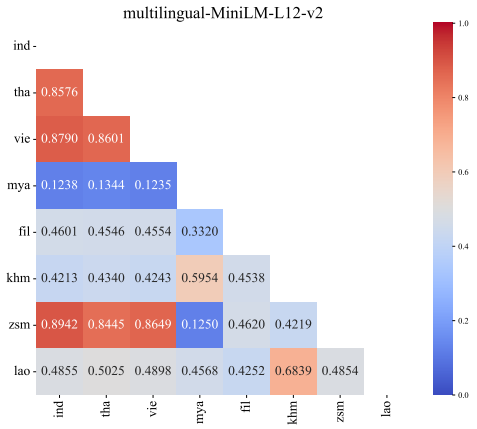
F Language Similarity

To further understand the similarity between language and performance, we analyze the performance of bi-text retrieval datasets in SEA languages. In particular, we study the language similarity of robust and non-robust models, e.g., top-performing and worst-performing embedding models, to see what the desired property is to improve our benchmark. We utilize the dialect pairing subset task in this experiment, where we use a batch size of 128 for the negative pair evaluation. In addition, we use cosine similarity as the main metric, where higher values indicate greater embedding similarity between language pairs.

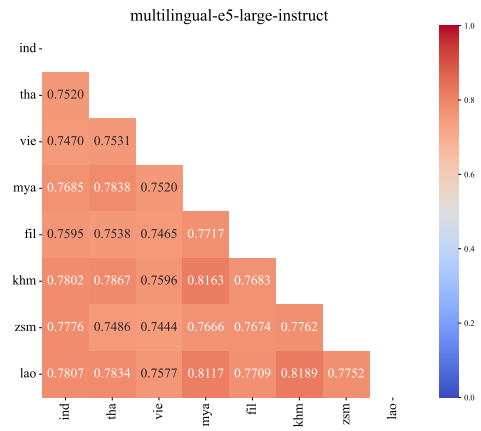
As shown in Figure 2, the top-performing model, multilingual-e5-large-instruct, shows consistently high similarity for positive samples, especially Indonesian-Malay (0.9682 points), Indonesian-Filipino (0.9305 points), and Thai-Vietnamese (0.9168 points), indicating strong cross-lingual embeddings. However, multilingual-e5-large-instruct unexpectedly maintains high similarity for negative samples (0.75-0.81 points), indicating limited distinction between unrelated sentence pairs and highlighting a gap for improvement. In contrast, multilingual-MiniLM-L12-v2 struggles with related positive pairs, showing lower similarity for Indonesian-Filipino (0.4601 points) and notably weak similarity with Burmese (around 0.12-0.59 points). Interestingly, this model achieves low similarity for negative pairs, mostly under 0.08 points, clearly distinguishing unrelated samples. Although it falls short in overall embedding quality, multilingual-MiniLM-L12-v2's distinct negative sample separation provides valuable insights into desirable characteristics for embedding models. These findings suggest that a balanced approach, achieving both strong cross-lingual similarity for positive examples and clear differentiation for negative examples, is essential to improve future embedding benchmarks.



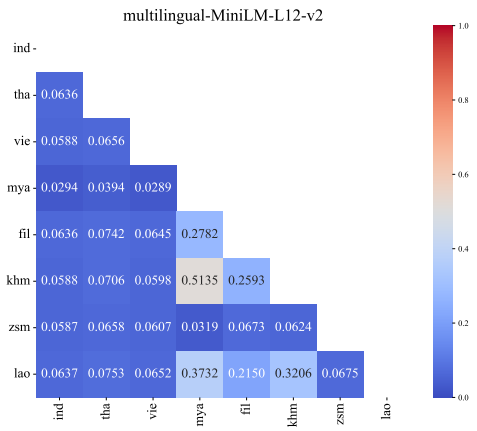
(a) The top-performing model on the positive pairs



(b) The worst-performing model on the positive pairs



(c) The top-performing model on the negative pairs



(d) The worst-performing model on the negative pairs

Figure 2: We perform cross-lingual similarity using the bitext mining task (dialect pairing subset). (Top) Cross-lingual similarity metrics of the top-performing and worst-performing embedding models on the positive parallel samples. (Bottom) Cross-lingual correlation metrics of the top-performing and worst-performing embedding models on the negative parallel samples.

G Machine vs. Human Datasets

This section testing whether human-crafted data yields results different from machine-generated data. We split the experiment into machine generation and translation studies.

Machine Translation vs. Human-translated Datasets. To compare machine-translated and human-translated data, we evaluate our new Thai and Burmese STS datasets from Table 4 against versions translated by Google's MT system. As shown in Table 10, Thai results differ by less than 2 Spearman

points across all settings, aligning with prior work showing that English-Thai NMT is already reliable for practical use (Lowphansirikul et al., 2022; Chiaranaipanich et al., 2024). In contrast to Thai, the performance gap between Burmese human and machine translation datasets is larger than that of Thai in most cases. We found that the Google NMT results for Burmese sometimes show code-switching between Thai and Burmese characters, as shown in Figure 3. This emphasizes that, in underrepresented languages, using humans to create evaluation datasets is still better than relying on machine translations.

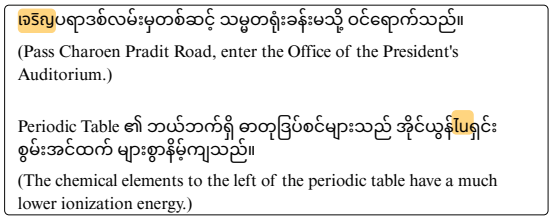


Figure 3: Code-switching between Thai and Burmese words (translated by Google NMT).

Machine Generation vs. Human Datasets. While many recent works introduce datasets generated with machine learning for scalability, we argue that fully machine-generated data remains unstable and should not dominate benchmarks, as it can distort model performance and research conclusions. To illustrate this, we compare human-crafted and machine-generated datasets using the top five models from our previous study. Keeping the same tasks and languages, we evaluate only datasets created by either humans or machines. The results for both settings are reported in Tables 12, 13 and 14.

Our results reveal two main effects: (i) shifts in average performance and (ii) shifts in model ranking. First, machine-generated datasets almost always reduce performance relative to human-crafted ones, with the sole exception of bge-multilingual-gemma2. This finding aligns with our results for machine-translated data (Table 10), indicating that machine-generated inputs can degrade model performance. Second, rankings become unstable, making evaluations unreliable. A robust benchmark should align with human-based outcomes, yet machine-generated datasets fail to preserve ranking consistency. Tetum offers a clear example: in Table 12, bge-multilingual-gemma2 leaps from 30.91 to 99.18 points. This occurs because the Tetum machine-generated set resembles a simple language-detection task from MADLAD-400, where all models score above 90, suggesting data leakage or in-domain overlap. Additional analysis of the machine-generated datasets is provided in Appendix H.

Model	Original (eng)	Mach. (mya)	Hum. (mya)	Diff. (mya)	Mach. (tha)	Hum. (tha)	Diff. (tha)
multilingual-e5-large-instruct (560M)	82.87	74.82	75.06	0.24	79.66	79.80	0.14
Qwen3-Embedding-8B (595M)	81.17	74.02	75.81	1.79	80.75	80.74	0.01
bge-multilingual-gemma2 (9B)	84.64	75.51	72.87	2.64	80.25	78.97	1.28
multilingual-e5-large (560M)	80.00	71.49	71.55	0.06	76.44	76.50	0.06
bge-m3 (568M)	80.86	74.57	71.96	2.61	77.75	76.07	1.68
GritLM-7B (7B)	82.65	65.60	66.03	0.43	74.64	74.81	0.17
e5-mistral-7b-instruct (7B)	81.86	62.36	64.63	2.27	74.57	74.57	0.00
Qwen3-Embedding-0.6B (8B)	80.11	67.10	69.23	2.13	77.88	77.79	0.09
multilingual-mpnet-base (278M)	80.54	72.34	71.16	1.18	72.61	72.60	0.01
LaBSE (471M)	73.50	69.06	70.04	0.98	69.29	68.83	0.46
multilingual-MiniLM-L12 (118M)	78.89	69.27	67.26	2.01	72.25	72.23	0.02
Gemma-SEA-LION-v3-9B-IT (9B)	60.42	46.50	49.29	2.79	55.97	56.01	0.04
Sailor2-8B-Chat (8B)	57.94	48.79	52.89	4.1	55.85	54.36	1.49
Proprietary models							
embed-multilingual-v3.0	82.62	74.56	73.92	0.64	78.40	78.63	0.23
jina-embeddings-v3	78.37	75.92	75.61	0.31	76.40	76.36	0.04
voyage-3	81.77	69.54	68.20	1.34	77.42	73.64	3.78
text-embedding-3-small	82.37	54.86	55.65	0.79	66.47	66.45	0.02

Table 10: Model performance on Machine Translation vs. Human Datasets on our STS datasets.

H Dataset Analysis

Performance Analysis. In addition to studying in Appendix G, we analyze the correlation between the percentage of vocabulary token coverage and performance scores for the two top-performing and two lowest-performing embedding models, as shown in Figure 4. The results indicate that the vocabulary size of each model does not have a direct effect on model performance in the text embedding benchmark

for SEA languages. Although some models have a larger number of tokens in their tokenizers covering SEA vocabularies, their performance in the benchmark is not significantly higher than that of models with lower vocabulary coverage. This indicates that simply increasing vocabulary size does not necessarily lead to better performance in text embedding tasks for SEA languages.

Discussion. In contrast to previous works, we summarize that the number of tokens present in the tokenizer might not strongly correlate with the performance in a language. There are many SEA languages with diverse scripts, and solely having a larger vocabulary for each language might not necessarily yield significant improvement. As shown in the language performances of GritLM-7B and bge-multilingual-gemma2 (Table 5), omitting SEA languages from the training data results in poor performance in those languages. To achieve a promising result, we can add more SEA training datasets in the training step to improve downstream task performance rather than adding more tokens in the tokenizer.

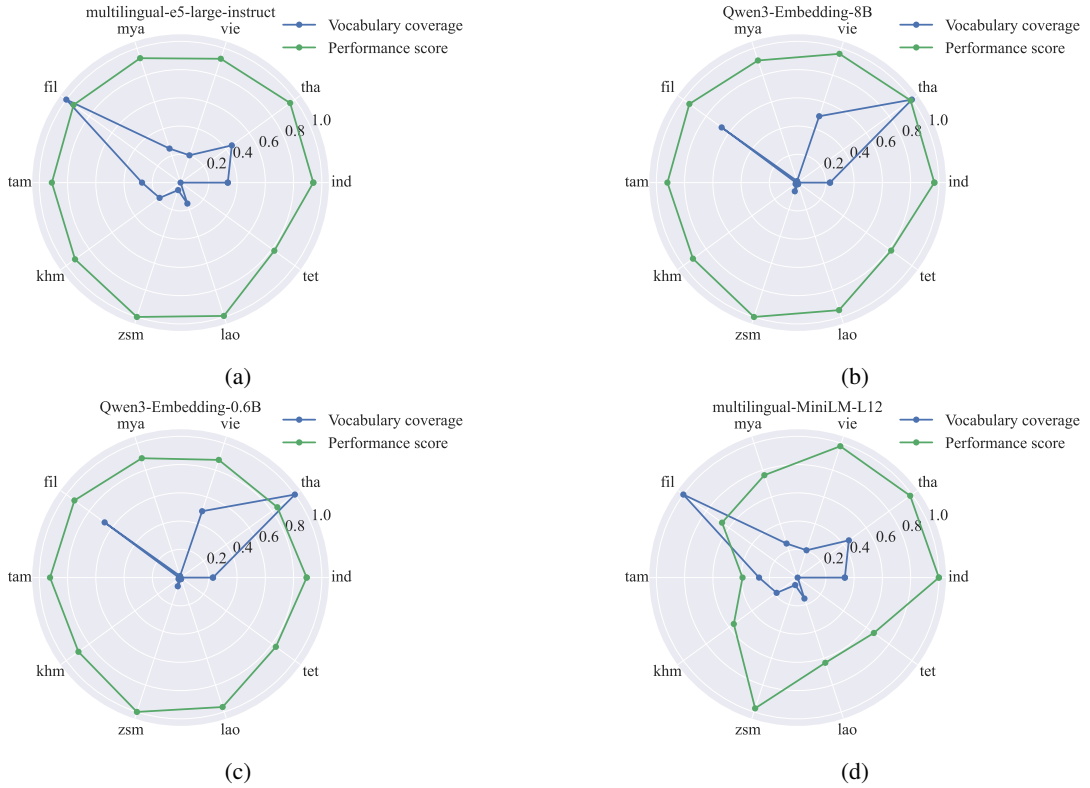


Figure 4: (Top) Correlation between the percentage of vocabulary token coverage and performance score for the two top-performing models, multilingual-e5-large-instruct and Qwen3-Embedding-8B. (Bottom) Correlation between the percentage of vocabulary token coverage and performance score for the two lowest-performing models, Qwen3-Embedding-0.6B and multilingual-MiniLM-L12. Both values are normalized to a [0, 1] scale for comparability across languages and models.

I Models

To evaluate text embedding on SEA texts, we experiment on 13 open-source models across encoder and decoder models as follows:

- **multilingual-e5-large** (Wang et al., 2024b). A multilingual-e5-large model that is trained on over 100 languages using a combination of contrastive pre-training on diverse multilingual text pairs and supervised fine-tuning on high-quality labeled datasets using mined hard negatives and knowledge distillation techniques.
- **multilingual-e5-large-instruct** (Wang et al., 2024b). The multilingual-e5-large-instruct model is similar to multilingual-e5-large, with additional fine-tuning on instructional data.
- **e5-mistral-7b-instruct** (Wang et al., 2024a). The e5-mistral-7b-instruct model is a text embedding model based on Mistral-7B (Jiang et al., 2023), fine-tuned with contrastive learning on synthetic

instruction data across 93 languages. Using a two-step prompting strategy, the model learns from diverse embedding tasks and achieves strong multilingual performance with under 1,000 training steps.

- **multilingual-mpnet-base-v2** (Reimers and Gurevych, 2020). The multilingual-mpnet-base-v2 model is trained on parallel data for over 50 languages via multilingual knowledge distillation using paraphrase-mpnet-base-v2 (Reimers and Gurevych, 2019) as a teacher model, xlm-roberta-base (Conneau et al., 2019) as a student model, and MSE loss to align their embeddings.
- **LaBSE** (Feng et al., 2022). The LaBSE model is trained on over 109 languages using a dual-encoder transformer architecture based on BERT (Devlin et al., 2018), leveraging a translation ranking loss function to produce sentence embeddings that align semantically similar sentences across languages into a shared vector space
- **multilingual-MiniLM-L12-v2** (Reimers and Gurevych, 2020). The multilingual-MiniLM-L12-v2 model is trained using a similar multilingual knowledge distillation approach to multilingual-mpnet-base-v2, with paraphrase-MiniLM-L12-v2 (Reimers and Gurevych, 2019) as a teacher model, Multilingual-MiniLM-L12-H384 (Wang et al., 2020) as a student model, and MSE loss to align their embeddings.
- **bge-m3** (Chen et al., 2024). The BGE-M3 model is trained on over 100 languages using a combination of contrastive pre-training on diverse multilingual corpora and supervised fine-tuning with high-quality labeled and synthetic datasets, leveraging hard negative mining and a self-knowledge distillation framework that integrates dense, sparse, and multi-vector retrieval signals.
- **bge-multilingual-gemma2** (Chen et al., 2024). The bge-multilingual-gemma2 model is built on Gemma-2-9b (Team, 2024) and trained on diverse multilingual data across tasks such as retrieval, classification, and clustering using embedding techniques.
- **GritLM-7B** (Muennighoff et al., 2024). The GritLM-7B model is built on the Mistral-7B (Jiang et al., 2023) architecture and trained using Generative Representational Instruction Tuning (GRIT), a unified framework combining contrastive learning for embeddings and next-token prediction for generation, with task-specific instructions and a joint loss to enable strong performance across both tasks.
- **Qwen3-Embedding-0.6B and Qwen3-Embedding-8B** (Zhang et al., 2025). The Qwen3-Embedding-0.6B and Qwen3-Embedding-8B models were trained on multiple languages using a multi-stage training pipeline that combines large-scale weakly supervised pre-training on synthetic multilingual data with supervised fine-tuning and model merging techniques to enhance robustness and generalization.
- **Sailor2-8B-Chat** (Dou et al., 2025). The Sailor2-8B-Chat model, based on an expanded Qwen2.5-7B (Yang et al., 2024), was trained on 13 SEA languages using two-stage continual pre-training with balanced and high-quality data, followed by two-stage instruction tuning and preference tuning with length-regularized DPO (Park et al., 2024).
- **Gemma-SEA-LION-v3-9B-IT** (Singapore, 2024). The Gemma-SEA-LION-v3-9B-IT model is fine-tuned from the Gemma2 9B (Rivière et al., 2024) base model on English and multiple SEA languages (such as Indonesian, Thai, and Vietnamese), using a combination of full parameter fine-tuning, on-policy alignment, and model merging techniques.

Moreover, we also evaluate the performance of proprietary models as follows:

- **text-embedding-3-small**. We evaluate the text-embedding-3-small [3] model, which provides a highly efficient embedding model suitable for various downstream applications.
- **embed-multilingual-v3.0**. We evaluate the embed-multilingual-v3.0 [4] model, designed for multilingual representation learning across over 100 languages.
- **voyage-3**. We evaluate the voyage-3 [5] model, which provides efficient, high-quality embeddings optimized for retrieval across diverse domains.

[3] https://openai.com/index/new-embedding-models-and-api-updates
[4] https://cohere.com/blog/introducing-embed-v3
[5] https://blog.voyageai.com/2024/09/18/voyage-3/

- **jina-embeddings-v3**. We evaluate the jina-embeddings-v3 (Sturua et al., 2024b) model, which is designed for efficient semantic similarity and search applications, supporting various multilingual scenarios.

All proprietary models were accessed and evaluated using their latest publicly available versions during experimentation (April 4th, 2025). The full model links are shown in Table 11.

Model	Hugging Face Link
multilingual-e5-large-instruct	https://huggingface.co/intfloat/multilingual-e5-large-instruct
Qwen3-Embedding-8B	https://huggingface.co/Qwen/Qwen3-Embedding-8B
bge-multilingual-gemma2	https://huggingface.co/BAAI/bge-multilingual-gemma2
multilingual-e5-large	https://huggingface.co/intfloat/multilingual-e5-large
bge-m3	https://huggingface.co/BAAI/bge-m3
GritLM-7B	https://huggingface.co/GritLM/GritLM-7B
e5-mistral-7b-instruct	https://huggingface.co/intfloat/e5-mistral-7b-instruct
Qwen3-Embedding-0.6B	https://huggingface.co/Qwen/Qwen3-Embedding-0.6B
multilingual-mpnet-base	https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2
LaBSE	https://huggingface.co/sentence-transformers/LaBSE
multilingual-MiniLM-L12	https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2
Gemma-SEA-LION-v3-9B-IT	https://huggingface.co/aisingapore/Gemma-SEA-LION-v3-9B-IT
Sailor2-8B-Chat	https://huggingface.co/sail/Sailor2-8B-Chat

Table 11: Models and Hugging Face links used for the evaluation.

J Example of Our Evaluation Tool

Similar to the previous text embedding benchmarks (Muennighoff et al., 2023; Enevoldsen et al., 2025), the evaluation tool of SEA-BED can be simply run using Python as shown in Figure 5. We will release all the evaluation tools, codes, results, and datasets in the final version of our paper.

```python
from seabed import SEABED
from seabed.results_to_dataframe import results_to_dataframe
from sentence_transformers import SentenceTransformer

# Define the sentence-transformers model name
model_name = "sentence-transformers/paraphrase-multilingual-mpnet-base-v2"

model = SentenceTransformer(model_name)
evaluation = SEABED(task_types=["STS", "PairClassification"])
results = evaluation.run(model, output_folder=f"results/{model_name}", batch_size=32)
results_to_dataframe(results, output_path=f"results/{model_name}")
```

Figure 5: Example usage of the SEA-BED evaluation framework for Semantic Textual Similarity (STS) and Pair Classification tasks.

K Task Examples

Figures 6 to 14 provide examples for each task covered in SEA-BED benchmark.

L Data Links

The complete dataset information, such as citations, languages, domains, annotation creators, and licenses, are shown in Tables 15 and 16.

Model	ind	tha	vie	mya	fil	khm	zsm	lao	tam	tet	Avg.
Number of datasets (→)	(51)	(50)	(36)	(32)	(28)	(18)	(15)	(16)	(17)	(2)	(265)
multilingual-e5-large-instruct (560M)	82.06	**83.12**	79.54	**78.27**	80.33	**79.63**	88.68	**87.07**	76.89	41.06	**77.67**$_{\pm 12.71}$
Qwen3-Embedding-8B (8B)	82.98	82.85	**80.67**	75.35	78.78	75.86	86.87	80.70	75.66	36.00	75.57$_{\pm 13.65}$
bge-multilingual-gemma2 (9B)	83.28	82.08	80.29	70.43	**80.67**	74.03	85.18	66.50	**80.99**	30.91	73.44$_{\pm 15.27}$
multilingual-e5-large (560M)	81.90	81.72	**80.67**	70.56	79.42	72.63	82.92	82.81	78.03	32.24	74.29$_{\pm 14.58}$
bge-m3 (568M)	81.28	79.67	77.48	73.70	76.87	76.56	88.07	85.03	77.70	36.91	75.33$_{\pm 13.43}$
GritLM-7B (7B)	**83.32**	74.64	78.80	42.56	78.48	50.20	80.76	24.36	60.09	40.72	61.39$_{\pm 19.77}$
e5-mistral-7b-instruct (7B)	82.40	76.50	77.32	46.95	79.17	54.68	81.45	23.12	66.61	37.21	62.54$_{\pm 20.00}$
Qwen3-Embedding-0.6B (595M)	78.83	77.16	76.70	47.62	62.83	39.19	71.48	24.31	60.67	30.69	56.95$_{\pm 19.22}$
multilingual-mpnet-base (278M)	77.09	76.06	74.01	60.52	51.01	62.34	77.23	65.09	63.15	34.54	64.10$_{\pm 12.83}$
LaBSE (471M)	75.72	72.59	74.82	73.98	78.27	74.66	**89.21**	83.41	77.05	**43.46**	74.42$_{\pm 11.34}$
multilingual-MiniLM-L12 (118M)	73.54	72.49	71.26	53.68	46.20	35.01	70.63	42.72	27.57	30.54	52.36$_{\pm 17.52}$
Gemma-SEA-LION-v3-9B-IT (9B)	51.93	42.10	52.09	28.02	53.69	35.63	52.25	16.07	27.58	0.08	35.94$_{\pm 17.17}$
Sailor2-8B-Chat (8B)	51.30	36.18	42.23	27.43	45.19	22.91	29.08	11.84	27.22	1.45	29.48$_{\pm 14.42}$
Proprietary models											
embed-multilingual-v3.0	**83.03**	**82.94**	**80.67**	**76.94**	**80.16**	**78.02**	**86.66**	**86.64**	**78.86**	**38.08**	**77.20**$_{\pm 13.42}$
jina-embeddings-v3	80.37	80.35	77.25	75.61	74.94	74.59	81.96	79.56	75.94	33.89	73.45$_{\pm 13.41}$
voyage-3	78.28	71.60	75.25	46.42	72.18	29.32	73.03	18.49	67.37	25.62	55.76$_{\pm 22.19}$
text-embedding-3-small	82.04	55.86	71.64	30.68	68.52	24.25	71.76	18.36	34.01	35.25	49.24$_{\pm 22.02}$

(a) Human-crafted datasets

Model	ind	tha	vie	mya	fil	khm	zsm	lao	tam	tet	Avg.
Number of datasets (→)	(19)	(5)	(5)	(3)	(3)	(4)	(4)	(3)	(1)	(2)	(49)
multilingual-e5-large-instruct (560M)	72.04	61.03	66.92	**79.47**	68.54	71.38	69.31	67.23	80.45	97.73	73.41$_{\pm 9.79}$
Qwen3-Embedding-8B (8B)	69.95	**67.83**	66.91	70.19	**71.24**	73.66	65.62	64.85	**80.86**	98.89	73.00$_{\pm 9.68}$
bge-multilingual-gemma2 (9B)	71.17	65.64	**67.80**	65.55	69.77	**76.00**	**76.63**	62.19	80.40	**99.18**	**73.43**$_{\pm 10.14}$
multilingual-e5-large (560M)	68.58	61.60	66.43	67.30	64.61	69.77	69.52	64.45	74.34	94.86	70.15$_{\pm 8.89}$
bge-m3 (568M)	69.14	56.74	64.57	66.90	65.62	74.74	61.77	67.47	74.39	94.15	69.55$_{\pm 9.65}$
GritLM-7B (7B)	**72.24**	54.86	67.10	71.61	68.20	63.31	69.58	60.49	66.04	98.62	69.21$_{\pm 11.01}$
e5-mistral-7b-instruct (7B)	70.13	57.49	61.31	69.09	68.11	64.61	68.93	53.93	68.74	96.26	67.86$_{\pm 10.82}$
Qwen3-Embedding-0.6B (595M)	66.02	62.77	63.78	64.62	65.68	66.20	62.14	58.99	67.51	96.07	67.38$_{\pm 9.84}$
multilingual-mpnet-base (278M)	67.93	52.53	62.97	68.37	61.37	73.89	68.93	**68.51**	66.07	67.01	65.76$_{\pm 5.47}$
LaBSE (471M)	68.42	46.30	56.59	69.90	65.05	71.38	59.14	60.84	68.88	94.85	66.14$_{\pm 12.01}$
multilingual-MiniLM-L12 (118M)	65.86	49.79	60.10	62.95	56.71	61.99	65.64	59.30	33.12	64.84	58.03$_{\pm 9.50}$
Gemma-SEA-LION-v3-9B-IT (9B)	44.34	37.41	50.56	60.52	58.38	57.09	37.94	53.72	56.71	50.04	50.67$_{\pm 7.89}$
Sailor2-8B-Chat (8B)	44.86	33.92	48.10	59.03	55.23	56.18	36.99	52.86	51.61	50.08	48.89$_{\pm 7.77}$
Proprietary models											
embed-multilingual-v3.0	**69.76**	61.41	66.40	67.52	**68.06**	72.48	**66.53**	65.73	79.09	95.43	71.24$_{\pm 9.20}$
jina-embeddings-v3	68.40	**61.53**	**67.83**	**69.73**	67.76	**75.37**	62.74	**69.11**	**79.70**	96.33	**71.85**$_{\pm 9.59}$
voyage-3	67.32	51.57	62.38	67.12	64.41	60.69	54.49	55.08	65.71	**97.34**	64.61$_{\pm 12.12}$
text-embedding-3-small	67.53	49.03	58.65	55.30	63.93	56.71	62.36	53.88	58.56	94.92	62.09$_{\pm 12.03}$

(b) Machine-generated datasets

Table 12: Language-model performance view, where each cell reports scores averaged over all evaluated task types, separated into human-crafted datasets (Top) and machine-generated datasets (Bottom).

Model	Dim.	Clf	M. Clf	Pr. Clf	STS	Clust	Btxt	Rtrvl	In. Rtrvl	Rrnk	Avg.
Number of datasets (→)		(64)	(9)	(7)	(10)	(10)	(20)	(18)	(1)	(1)	(140)
multilingual-e5-large-instruct (560M)	1024	80.33	86.61	69.10	73.08	**53.18**	**87.62**	80.41	96.38	77.24	**78.22**$_{\pm 11.72}$
Qwen3-Embedding-8B (8B)	4096	**81.48**	89.35	65.45	**73.38**	43.91	84.67	**83.89**	96.18	78.51	77.42$_{\pm 14.49}$
bge-multilingual-gemma2 (9B)	3584	81.19	**89.44**	**78.53**	69.57	41.98	81.86	82.81	**96.89**	69.04	76.81$_{\pm 14.80}$
multilingual-e5-large (560M)	1024	80.97	87.87	67.63	68.07	37.03	84.51	81.66	96.01	**79.00**	75.86$_{\pm 16.09}$
bge-m3 (568M)	4096	78.59	88.78	71.88	71.49	33.73	86.34	77.92	87.73	75.98	74.72$_{\pm 15.73}$
GritLM-7B (7B)	4096	79.80	87.16	65.82	62.27	36.67	62.15	68.64	93.50	73.37	69.93$_{\pm 15.66}$
e5-mistral-7b-instruct (7B)	4096	79.03	86.96	65.99	60.77	40.91	64.06	75.51	94.31	75.33	71.43$_{\pm 14.85}$
Qwen3-Embedding-0.6B (595M)	1024	77.02	86.83	62.33	64.14	34.79	55.19	78.23	94.30	75.03	69.76$_{\pm 16.89}$
multilingual-mpnet-base (278M)	768	76.18	85.56	75.02	67.84	33.87	67.67	61.05	86.80	64.01	68.67$_{\pm 14.92}$
LaBSE (471M)	768	77.53	85.13	63.45	67.92	32.53	86.61	56.82	79.92	61.23	67.90$_{\pm 16.06}$
multilingual-MiniLM-L12 (118M)	768	72.85	83.02	69.86	64.23	23.51	52.14	55.10	83.58	62.27	62.95$_{\pm 17.35}$
Gemma-SEA-LION-v3-9B-IT (9B)	3584	78.67	88.34	57.89	32.62	28.74	16.05	23.74	31.16	65.49	46.97$_{\pm 24.64}$
Sailor2-8B-Chat (8B)	3584	79.44	88.80	56.63	32.32	26.28	4.25	10.94	10.57	47.05	39.59$_{\pm 28.86}$
Proprietary models											
embed-multilingual-v3.0	1024	**81.59**	**88.87**	**68.29**	**71.24**	40.69	**88.34**	**81.36**	**96.87**	**77.77**	**77.22**$_{\pm 15.39}$
jina-embeddings-v3	1024	80.25	87.72	65.98	69.56	**44.38**	81.83	78.94	96.51	72.49	75.30$_{\pm 14.02}$
voyage-3	1024	78.26	87.36	62.59	61.50	36.50	54.22	66.19	87.29	74.62	67.61$_{\pm 15.46}$
text-embedding-3-small	1536	75.76	86.53	61.95	47.54	30.24	40.82	67.57	83.60	71.25	62.81$_{\pm 18.36}$

(a) Human-crafted datasets

Model	Dim.	Clf	M. Clf	Pr. Clf	STS	Clust	Btxt	Rtrvl	In. Rtrvl	Rrnk	Avg.
Number of datasets (→)		(9)	(2)	(6)	(1)	(0)	(6)	(2)	(3)	(0)	(29)
multilingual-e5-large-instruct (560M)	1024	65.76	93.34	64.24	**80.61**	-	**92.13**	43.04	60.00	-	71.30$_{\pm 16.96}$
Qwen3-Embedding-8B (8B)	4096	65.45	96.05	60.38	79.19	-	86.67	**61.99**	62.35	-	73.15$_{\pm 13.13}$
bge-multilingual-gemma2 (9B)	3584	64.19	**97.45**	**70.37**	78.44	-	84.81	56.81	**63.07**	-	**73.59**$_{\pm 13.15}$
multilingual-e5-large (560M)	1024	65.79	93.78	63.00	72.67	-	84.57	42.48	56.07	-	68.34$_{\pm 15.96}$
bge-m3 (568M)	4096	64.12	94.87	65.76	76.84	-	83.29	27.70	48.76	-	65.91$_{\pm 20.76}$
GritLM-7B (7B)	4096	**66.83**	95.98	62.50	69.54	-	90.38	37.93	58.96	-	68.87$_{\pm 18.12}$
e5-mistral-7b-instruct (7B)	4096	65.81	94.43	62.16	68.97	-	87.72	45.87	41.18	-	66.59$_{\pm 18.22}$
Qwen3-Embedding-0.6B (595M)	1024	62.89	94.34	58.17	68.95	-	80.57	55.39	56.30	-	68.09$_{\pm 13.48}$
multilingual-mpnet-base (278M)	768	62.87	94.99	67.01	74.76	-	75.91	29.14	40.98	-	63.67$_{\pm 20.61}$
LaBSE (471M)	768	64.52	93.48	60.60	69.13	-	90.89	21.12	26.34	-	60.87$_{\pm 26.24}$
multilingual-MiniLM-L12 (118M)	768	59.81	93.24	62.21	65.30	-	73.23	24.94	37.01	-	59.39$_{\pm 20.94}$
Gemma-SEA-LION-v3-9B-IT (9B)	3584	63.11	97.12	56.86	51.31	-	2.03	4.12	4.31	-	39.84$_{\pm 34.25}$
Sailor2-8B-Chat (8B)	3584	62.72	96.58	56.18	47.11	-	5.28	1.18	0.86	-	38.56$_{\pm 34.35}$
Proprietary models											
embed-multilingual-v3.0	1024	**64.52**	94.97	**63.15**	76.86	-	**87.86**	44.68	55.49	-	69.65$_{\pm 16.55}$
jina-embeddings-v3	1024	64.41	94.61	60.96	**80.38**	-	82.28	**48.38**	**59.98**	-	**70.14**$_{\pm 14.86}$
voyage-3	1024	64.13	94.69	57.55	62.92	-	80.80	28.42	53.27	-	63.11$_{\pm 19.43}$
text-embedding-3-small	1536	59.77	**95.62**	58.51	61.84	-	84.55	39.99	42.63	-	63.27$_{\pm 18.91}$

(b) Machine-generated datasets

Table 13: Task-model performance view, where each cell reports scores averaged over all evaluated languages, separated into human-crafted datasets (Top) and machine-generated datasets (Bottom). "-" indicates that no dataset is available for the corresponding task.

Model	Btxt	Clf	Clust	In. Rtrvl	M. Clf	Pr. Clf	Rtrvl	Rrnk	STS
Indonesian	75.98	81.00	43.55	-	92.49	76.28	72.71	69.28	46.43
Thai	70.99	75.67	38.49	82.98	77.00	70.31	74.69	71.86	68.32
Vietnamese	76.36	80.62	38.56	-	96.49	71.86	56.80	-	-
Burmese	48.83	85.10	28.09	-	69.71	67.10	55.44	-	61.19
Filipino	69.56	75.60	52.69	-	-	50.63	-	-	-
Khmer	54.21	74.39	23.83	-	-	-	-	-	-
Malay	75.98	77.21	-	-	-	-	-	-	-
Lao	51.79	75.91	18.23	-	-	-	-	-	-
Tamil	61.23	84.35	38.75	-	-	67.14	46.27	-	37.21
Tetum	31.09	-	-	-	-	-	-	-	-

(a) Human-crafted datasets

Model	Btxt	Clf	Clust	In. Rtrvl	M. Clf	Pr. Clf	Rtrvl	Rrnk	STS
Indonesian	74.84	67.61	-	34.82	-	59.58	-	-	75.65
Thai	-	59.11	-	57.88	-	-	25.87	-	71.99
Vietnamese	-	67.96	-	42.75	-	55.49	-	-	77.15
Burmese	-	66.15	-	-	-	-	-	-	68.03
Filipino	-	36.12	-	-	90.06	-	-	-	68.41
Khmer	-	39.36	-	-	100.00	66.41	-	-	63.74
Malay	-	55.84	-	-	-	71.50	46.27	-	75.39
Lao	-	59.81	-	-	-	64.36	-	-	59.11
Tamil	-	-	-	-	-	-	-	-	67.78
Tetum	75.08	99.81	-	-	-	-	-	-	-

(b) Machine-generated datasets

Table 14: Language-task performance view, where each cell reports scores averaged over all evaluated models, separated into human-crafted datasets (Top) and machine-generated datasets (Bottom). "-" indicates that no dataset is available for the corresponding language-task combination.

Task	Text	Label
Language Identification	sadiissss kae pasti neng tegal yakin inyong	Jawa Ngapak
	kowe gurung ngerti betapa bahagianya dichat arek e seh wkwkwk	Jawa Timur
Sentiment	Namiss ko yung pusa ko nung bata pa ako, lagi ko sya katabi matulog and malambing.	Positive
	Sa tiktok ang pugad nila. Dapat ma ban na ang tiktok app	Negative
Topic Classification	மீண்டும் செல்வராகவன் படத்தில் நடிப்பீர்களா?	tamil-cinema
	இத்தகையோரை பாதுகாக்கவே இத்தகைய ஸ்கூட்டர் வடிவமைக்கப் பட்டுள்ளது.	business
Toxic Language Detection	Yang blm pada move on mending bsk piknik aja mumpung long weekend hehehe	Non hate speech
	Karna dia ga punya program hanya modal bacot doang	Hate speech

Figure 6: Classification examples.

Task	Text	Label
Sentiment	saran ku dan pengalaman ku , mending beli mobil niaga L300 atau canter . irit dan bandel .	[fuel (positive), machine (positive), others (neutral), part (neutral), price (neutral), service (neutral)]
	Sudah dari dulu Toyota selalu kasih produk super mahal dengan fitur pas pasan	[fuel (neutral), machine (neutral), others (neutral), part (negative), price (negative), service (neutral)]
Topic Classification	เกี่ยวกับซิมนะคะ พอดีซื้อซิมมาใหม่ไม่สามารถเปิดใช้งานได้	["report", "phone_issues"]
	เบอร์โทร โดน ระงับใช้บริการ ต้องทำอย่างไงค่ะ	["enquire", "suspend"]
Toxic Language Detection	Prabowo Sudah Kalah Menyebut Bantuan Jokowi Hanya Pencitraan Adalah Ratapan Pilu'	[Hate speech, Hate speech Individual, Hate speech Week]
	Wah bangke emang nih truk'	[Hate speech, Abusive, Hate speech Individual, Hate speech Week]

Figure 7: Multi-label Classification examples.

Task	Sentence 1	Sentence 2	Label
Textual Entailment	Làm sao anh biết ? Tất cả đây là thông tin của họ lần nữa .	Thông tin này thuộc về họ .	Contradiction
	Conceptually kem skimming có hai kích thước cơ bản - sản phẩm và địa lý .	Sản phẩm và địa lý là những gì làm cho kem skimming làm việc .	Entailment
	Vui vẻ dành cho người lớn và trẻ em .	Vui vì chỉ có trẻ con .	Neutral

Figure 8: Pair Classification examples.

Task	Sentence 1	Sentence 2	Score
Multilingual STS	လူတစ်ယောက်သည် ဘေ့စ်ဘောအသင်းတွင် ရှိနေသည်။	လူတစ်ဦးသည် အသင်းတစ်သင်းတွင် ဘတ်စကတ်ဘော ကစားနေသည်။	2.4
	Istilah benda hitam pertama kali diperkenalkan oleh Gustav Kirchhoff tahun 1860.	Istilah "benda hitam" pertama kali diperkenalkan oleh Gustav Robert Kirchhoff pada tahun 1862.	5
	ชายคนหนึ่งในรถสีเขียวกำลังทำความสะอาดถนนในเมือง	ชายและหญิงสามคนข้ามถนนในเมืองที่พลุกพล่าน	1.7
Cross-lingual STS	This triggered a revolution in the earth sciences.	இக் கோட்பாடு புவி அறிவியல் துறைகளில் புரட்சிகரமான மாற்றங்களை ஏற்படுத்திற்று.	4
	The up-regulation of miR-146a was also detected in cervical cancer tissues.	miR-146a ၏အသုံးအနှုန်းသည်သားအိမ်ခေါင်းကင်ဆာ တွင်ထိန်းချုပ်နိုင်သည်ကိုတွေ့ရှိရသည်။	4
	A person is on a baseball team.	มีคนเล่นบาสเก็ตบอลในทีม	2.4

Figure 9: STS examples.

Task	Text	Cluster
Topic Clustering	ဤရာသီ၏ ရောဂါဖြစ်ပွားမှု ကနဦးလူနာများသည် ဇူလိုင်လ အနှောင်းပိုင်းတွင် ပေါ်လာကြသည်။	health
	အပင်များသည် အစာချက်ခြင်းကို နေမှတဆင့်ပြုလုပ်သည်။ အရိပ်လဲပေးပါသည်။	science/technology
	ဤစာရွက်စာတမ်းများကို ဂုဏ်ပြုရန် နိုင်ငံခြား အစိုးရများ၏ စိတ်ထက်သန်မှုမှာ ပြောင်းလဲနိုင်ပါသည်။	politics
	ဝါရှင်တန်၏ အတ္တလန်နာသရက်ရှာကို ၅-၃ ဖြင့် အနိုင်ရသော ပွဲတွင် ၂ ဂိုးသွင်းပြီး ၂ ဂိုး ဖန်တီးပေးခဲ့သည်။	sports

Figure 10: Clustering examples.

Task	First set sentence	Second set sentence
Cross-lingual pairing	Paris is the most beautiful city in the world	Paris adalah kota tercantik di dunia.
Dialect pairing	Andrea Maisi đã mở tỉ số cho Ý ở phút thứ tư với một quả try.	แอมเดรย มาซี่ ได้เปิดการทำคะแนนในนาทีที่สี่ให้แก่อิตาลี.
Written-forms pairing	โรซาลีเล่าว่า "คุณต้องรู้ถึงอันตรายต่าง ๆ และดูว่าคุณพอจะทำอะไรได้บ้าง ที่ปรึกษาของฉัน ตอนที่เราดู การปะทุขนาดเล็กของภูเขาไฟเอตนา มีเศษวัตถุขนาดเล็กตกลงมา เขาจะบอกเราให้เข้าไปเก็บตัวอย่าง คุณต้องได้รับการฝึกอย่างดี อย่าวิ่ง ให้อยู่กับที่ และมองขึ้นด้านบน ถ้ามีวัตถุขนาดใหญ่ตกลงมาใส่ ก็จะได้หลบออกด้านข้าง" สำหรับผู้ที่ต้องการชมภูเขาไฟคุกรุ่น โรซาลี แนะนำให้ไป วานูอาตู มีภูเขาไฟชื่อ ยาซูร์ ซึ่งมีการปะทุขนาดเล็ก คล้ายกับดอกไม้ไฟ มีความสวยงาม และเป็นภูเขาไฟที่ไปง่าย เธอบอกว่า สามารถขับรถขึ้นไปเกือบถึงปากปล่อง จากนั้นก็มีบันไดคอนกรีต ที่สามารถเดินขึ้นไปได้ และยังมีม้านั่งให้นั่งเล่นด้วย นอกจากภูเขาไฟบนโลก เธอยังพบภูเขาไฟที่คุกรุ่น 71 ลูก บนดวงจันทร์ไอโอของดาวพฤหัสฯ ด้วย	โรซาลี โลเปส นักภูเขาไฟวิทยาของนาซาโปรดปรานการปีนภูเขาไฟที่ยังคุกรุ่นอยู่ เพื่อไปชมการปะทุขนาดเล็ก โดยเธอได้เยือนภูเขาไฟที่คุกรุ่นในทุกทวีปทั่วโลกมาแล้ว 63 ลูก

Figure 11: Bitext mining examples.

Task	Query	Relevant Document
Article Retrieval	Hà Nội: Đưa vào hoạt động trạm biến áp 110kV Bắc Thành Công	Việc đầu tư dự án 'Xây dựng mới Trạm 110kV Bắc Thành Công và nhánh rẽ' sẽ góp phần giảm được tổn hao công suất và điện năng của lưới điện trong khu vực, nâng cao chất lượng điện năng.
Long Document Retrieval	มะแว้งต้นมีประโยชน์อย่างไรในเชิงสรรพคุณ?	มะแว้งต้น ประโยชน์ดีๆ สรรพคุณเด่นๆ และข้อมูลงานวิจัย\nหน้าแรก > บทความทั้งหมด > มะแว้งต้น\nชื่อสมุนไพร มะแว้งต้น\nชื่ออื่นๆ/ชื่อท้องถิ่น มะแคว้งขม, มะแคว้งดำ, มะแคว้ง (ภาคเหนือ) ,หมากแข้ง , หมากแข้งขม (ภาคอีสาน) , มะแว้ง (ภาคกลาง) , แว้งกาม (สงขลา,สุราษฎร์ธานี,ภาคใต้) , สะกั้งแค (กะเหรี่ยง-แม่ฮ่องสอน) , หมากแซ้งคง (ไทยใหญ่ – แม่ฮ่องสอน , ฉาน) , เทียนเฉีย ,ซื่อเทียนเฉีย (จีนกลาง)\nชื่อวิทยาศาสตร์ Solanum indicum L. (มีหนาม) Solanum sanitwongsei (ไร้หนาม)\nชื่อพ้องทางวิทยาศาสตร์ Solanum violaceum (มีหนาม)\nชื่อสามัญ Sparrow's Brinjal , Indian nightshade\nถิ่นกำเนิดมะแว้งต้น\nมีการคาดการณ์กันว่าถิ่นกำเนิดดั้งเดิมของมะแว้งต้นนั้นอยู่ในเขตร้อนของทวีปเอเชียซึ่งอาจอยู่ในประเทศ แถบเอเชียใต้ เช่น อินเดีย บังคลาเทศ เนปาล ฯลฯ รวมถึงประเทศแถบเอเชียตะวันออกเฉียงใต้ เช่น ไทย, พม่า , ลาว ,กัมพูชา ฯลฯ ...
Question Answering	Dimana Jamie Richard Vardy lahir?	Jamie Richard Vardy (lahir dengan nama Gill; 11 January 1987) adalah pemain sepak bola Inggris yang bermain di klub Premiere League Leicester City dan tim nasional Inggris. Ia bermain sebagai striker, namun juga bisa bermain di sayap.

Figure 12: Retrieval examples.

Task	Query	Instruction	Relevant Document
Instruction Question Answering	Stellar คืออะไร	Stellar: เครือข่ายโอนเงินไร้พรมแดน เทคโนโลยีการโอนเงินระหว่างธนาคารในตอนนี้ถือว่าล้ำหน้ามาก ๆ นะครับ ทุกวันนี้เราสามารถโอนเงินจากบัญชีของเราไปยังบัญชีของธนาคารอื่น ๆ ได้อย่างสะดวก รวดเร็ว และไม่มีค่าธรรมเนียมใด ๆ ซึ่งไม่ใช่ทุกประเทศบนโลกนี้จะมีสิ่งอำนวยความสะดวกเหมือนกับประเทศไทยนะครับ ในประเทศอื่น ๆ การโอนเงินระหว่างบัญชียังมีค่าธรรมเนียม จะถูกจะแพงก็แล้วแต่ประเทศไป และยังใช้เวลาประมาณนึงอีกด้วยครับ ...	Stellar คือ เครือข่ายการโอนเงินแบบกระจายศูนย์ (decentralized) ที่มีเป้าหมายจะเป็นช่องทางการจ่ายเงินที่เร็ว ปลอดภัย ไร้พรมแดน และมีค่าธรรมเนียมที่ต่ำ ด้วยการใช้งานเทคโนโลยีบล็อกเชนทำให้ Stellar สามารถเชื่อมต่อทั้งบุคคลธรรมดากับองค์กร (เช่น ธนาคาร) และทำให้ผู้ใช้งานเหล่านี้สามารถส่งผ่านสินทรัพย์ไป-มาได้อย่างรวดเร็ว Stellar มีเป้าหมายที่จะ disrupt ระบบการจ่ายเงินที่ใช้กันอยู่ทุกวันนี้ ลองคิดถึงการโอนเงินข้ามประเทศดูทุกวันนี้การโอนเงินข้ามประเทศมีค่าธรรมเนียมการโอนที่แพง ...

Figure 13: Instruction Retrieval examples.

Task	Query	Positive	Negative
Article Reranking	kapankah Radin Inten II dilahirkan?	Radin Inten II (Lampung, 1834 - Lampung, 5 Oktober 1858) adalah seorang pahlawan nasional Indonesia.\nNamanya diabadikan sebagai sebuah Bandara Radin Inten II dan perguruan tinggi IAIN Raden Intan di Lampung.	Akhirnya, Waleson menemukan cara lain. Ia berhasil memperalat Radin Ngerapat. Maka pengkhianatan pun terjadi. Radin Ngerapat mengundang Radin Inten II untuk mengadakan pertemuan. Dikatakannya bahwa ia ingin membicarakan bantuan yang diberikannya kepada Radin Inten II. Tanpa curiga, Radin Inten II memenuhi undangan itu. Pertemuan diadakan malam tanggal 5 Oktober 1856 di suatu tempat dekat Kunyanya. Radin Inten II ditemani oleh satu orang pengikutnya. Radin Ngerapat disertai pula oleh beberapa orang. Akan tetapi, di tempat yang cukup tersembunyi, beberapa orang serdadu Belanda sudah disiapkan untuk bertindak bila diperlukan. Radin Ngerapat mempersilahkan Radin Inten II dan pengiringnya memakan makanan yang sengaja dibawanya terlebih dahulu.

Figure 14: Reranking examples.

Type	Name	Languages	Domains	Sample creation	Annotations creators	License
Classification	ABUSIVE (Ibrohim and Budi, 2018)	['ind']	['Social', 'Written']	found	human-annotated	CC BY-SA 4.0
	AbusiveNewsComment (Kiasati Desrul and Romadhony, 2019)	['ind']	['Social', 'Web', 'News', ...]	found	human-annotated	CC BY-SA 4.0
	BookmebusReviews	['khm']	['Reviews', 'Written']	found	human-annotated	
	Clickbait* (William and Sari, 2020)	['ind']	['News', 'Written']	found	expert-annotated	
	CodeMixed (Tho et al., 2021)	['ind']	['Social', 'Web']	found	manual curation	CC BY 3.0
	CyberbullyingLGBT	['tha']	['Social', 'Written']	found	derived	
	Depression (Hämäläinen et al., 2021)	['tha']	['Social', 'Web', 'News', ...]	found	human-annotated	CC BY-NC-ND 4.0
	EMoTES3K (Catapang and Visperas, 2023)	['fil']	['Morality', 'Written']	found	human-annotated	Apache license 2.0
	Emoji	['tha']	['Social', 'Written']	found	human-annotated	GPL-3.0
	EmoT (Mei Silviana Saputri and Adriani, 2018)	['ind']	['Social', 'Written']	found	human-annotated	MIT
	EmotionOpinion (Riccosan et al., 2022)	['ind']	['Social', 'Written']	found	human-annotated	CC BY-SA 4.0
	EmotCMT (Yulianti et al., 2021)	['ind']	['Social', 'Written']	found	derived	MIT
	Fakenews (Cruz et al., 2020b)	['fil']	['News', 'Written']	found	human-annotated	
	GeneralAmy (Phatthiyaphaibun et al., 2023)	['tha']	['Social', 'Written']	found	human-annotated	CC BY 3.0
	GeneratedReviewsENTH (Lowphansirikul et al., 2022)	['tha']	['conversation', 'Web', 'Written', ...]	found	human-annotated	CC BY-SA 4.0
	GKLMIPSentiment (Jiang et al., 2021b)	['mya']	['Social', 'Web', "Written]	found	derived	
	GooglePlayReview	['ind']	['Reviews', 'Written']	found	human-annotated	CC BY 4.0
	HateSpeech (Alfina et al., 2017)	['ind']	['Social', 'Written']	found	human-annotated	
	HateSpeech*	['fil']	['Social', 'Written']	found	human-annotated	Apache license 2.0
	HoaxNews (Pratiwi et al., 2017)	['ind']	['News', 'Written']	found	human-annotated	CC BY 4.0
	HSDNofaaulia (Aulia and Budi, 2019)	['fil']	['Social', 'Written']	found	human-annotated	
	IMDB (Maas et al., 2011)	['ind']	['Reviews', 'Written']	found	human-annotated	
	Indonglish (Astuti et al., 2023)	['ind']	['Social', 'Written']	found	expert-annotated	
	JaDiIde (Hidayatullah et al., 2020)	['ind']	['Social', 'Written']	found	derived	
	Karonese (Sitepu et al., 2024)	['ind']	['Social', 'Web']	found	derived	
	KhineMyanmarNews (Khine et al., 2017)	['mya']	['News', 'Written']	found	derived	GPL-3.0
	Krathu500	['tha']	['Social', 'Web', 'News', ...]	found	human-annotated	
	LazadaReview	['fil']	['Reviews', 'Written']	found	derived	
	LEMSentiment (Koto et al., 2020b)	['ind']	['Social', 'Review', 'Written']	found	human-annotated	CC BY-SA 4.0
	LimeSoda (Payoungkhamdee et al., 2021)	['tha']	['Healthcare', 'Written']	found	human-annotated	CC BY 4.0
	MADLAD400 (Kudugunta et al., 2023)	['tet']	['Web']	found	derived	ODC-BY
	MassiveIntent* (FitzGerald et al., 2022)	['ind', 'tha', 'vie', ...]	['Spoken']	found	human-annotated	CC BY 4.0
	MassiveScenario* (FitzGerald et al., 2022)	['ind', 'tha', 'vie', ...]	['Spoken']	found	human-annotated	CC BY 4.0
	Minang (Koto and Koto, 2020)	['ind']	['Encyclopaedic', 'Written']	found	derived	MIT
	MultiLingualSentiment* (Mollanorozy et al., 2023)	['ind', 'tha', 'vie']	['Reviews', 'Written']	found	derived	
	MurasuNews	['tam']	['News', 'Written']	found	derived	CC0
	News (Khine et al., 2017)	['mya']	['News', 'Written']	found	derived	GLP-3.0
	News	['zsm']	['News', 'Written']	found	derived	
	News	['khm']	['Encyclopaedic', 'Web', 'News', ...]	found	derived	
	News	['tam']	['News', 'Written']	found	derived	CC BY-SA 4.0
	News (Phatthiyaphaibun, 2025)	['lao']	['News', 'Written']	found	derived	
	NewsDataset	['ind']	['News', 'Written']	found	derived	
	NusaX (Winata et al., 2023)	['ind']	['Social', 'Economics', 'Healthcare', ...]	found	expert-annotated	CC BY-SA 4.0
	PhoATIS (Dao et al., 2021)	['vie']	['Spoken']	found	expert-annotated	
	PHElectionsSA	['fil']	['Social']	found	human-annotated	
	PHElectionsTD	['fil']	['Social']	found	human-annotated	
	Profanity (Galinato et al., 2023)	['fil']	['Social']	found	human-annotated	
	ReviewShopping (Phatthiyaphaibun et al., 2023)	['tha']	['Reviews', 'Written']	found	human-annotated	CC BY 3.0
	SIB200 (Adelani et al., 2023)	['ind', 'tha', 'vie', ...]	['News', 'Written']	found	expert-annotated	CC BY-SA 4.0
	SEATranslationeseResampled (Lovenia et al., 2024)	['ind', 'tha', 'vie', ...]	['News', 'Social', 'Culture', ...]	found	derived	Apache license 2.0
	SentEmoMobileApps (Riccosan and Saputra, 2023)	['ind']	['Reviews', 'Written']	found	human-annotated	
	SentimentAnalysis (Fe, 2019)	['ind']	['Social', 'Written']	found	derived	CC BY-NC-ND 4.0
	ShopeeReviews* (Purwarianti and Crisdayanti, 2019)	['fil']	['Social', 'Written']	found	human-annotated	MLP-2.0
	SMSA	['ind']	['Reviews', 'Written']	found	derived	MIT
	SpamidPair (Chrismanto et al., 2022)	['ind']	['Social', 'Written']	found	human-annotated	CC BY 4.0
	SpamReviews (Van Dinh et al., 2022)	['vie']	['Reviews', 'Written']	found	human-annotated	CC BY-NC 4.0
	StudentFeedback (Nguyen et al., 2018b)	['vie']	['Reviews', 'Written']	found	human-annotated	MIT
	TCAS61 (Phatthiyaphaibun et al., 2023)	['tha']	['Social', 'Written']	found	human-annotated	CC BY 3.0
	The40ThaiChildrenStories (Pasupa et al., 2016)	['tha']	['Encyclopaedic', 'Written']	found	human-annotated	
	ThuraMyanmarNews (Aung et al., 2024)	['mya']	['News', 'Written']	found	derived	MIT
	TiktokHatespeech (Hernandez Urbano Jr et al., 2021)	['fil']	['Social', 'Written']	found	human-annotated	CC BY-SA 4.0
	Tweets (Juan et al., 2022)	['zsm']	['Reviews', 'Written']	found	derived	
	TyphoonYolandaTweets	['fil']	['Social', 'Written']	found	human-annotated	CC BY 4.0
	UITViCTSD (Nguyen et al., 2021a)	['vie']	['Social', 'Written']	found	human-annotated	
	UITViHSD (Luu et al., 2021)	['vie']	['Social', 'Written']	found	human-annotated	
	UITViSFD (Luc Phan et al., 2021)	['vie']	['Social', 'Written']	found	human-annotated	
	UITVION (Fujita and Perez-Meana, 2021)	['vie']	['Social', 'Written']	found	human-annotated	
	UITVSMEC (Ho et al., 2020)	['vie']	['Social', 'Written']	found	human-annotated	
	VaccinesTweets	['ind']	['Social', 'Written']	found	human-annotated	
	ViOCD (Nguyen et al., 2021b)	['vie']	['Reviews', 'Written']	found	human-annotated	
	VLSP2016Sentiment (Nguyen et al., 2018a)	['vie']	['Reviews', 'Written']	found	human-annotated	
	WisesightSentiment (Suriyawongkul et al., 2019)	['tha']	['Social', 'News', 'Written']	found	expert-annotated	CC0-1.0
	WongnaiReviews	['tha']	['Reviews', 'Written']	found	derived	LGPL-3.0
Multi-label Classification	BurmesePrachathai67k (Phatthiyaphaibun et al., 2023)	['mya']	['News', 'Web', 'Written']	created	human-annotated	Apache license 2.0
	CASA (Arfinda Ilmania and Purwarianti, 2018)	['ind']	['Reviews', 'Written']	found	human-annotated	MIT
	Dengue (Livelo and Cheng, 2018)	['fil']	['Social', 'Written']	found	derived	GLP-3.0
	GKLMIPNews (Jiang et al., 2021a)	['khm']	['News', 'Written']	found	derived	
	HateSpeech (Ibrohim and Budi, 2019)	['ind']	['Social', 'Written']	found	human-annotated	CC BY-SA 4.0
	HoASA (A. N. Azhar and Sutiono, 2019)	['ind']	['Reviews', 'Written']	found	human-annotated	MIT
	Netifier (Izzan et al., 2025)	['ind']	['Social', 'Written']	found	human-annotated	CC BY-SA 4.0
	Prachathai67k (Phatthiyaphaibun et al., 2023)	['tha']	['News', 'Web', 'Written']	found	derived	Apache license 2.0
	TrueVoiceIntent	['tha']	['Conversation']	found	derived	
	VLSP2018SAHotel (Dang et al., 2022)	['vie']	['Reviews', 'Written']	found	human-annotated	
	VLSP2018SARestaurant (Dang et al., 2022)	['vie']	['Reviews', 'Written']	found	human-annotated	
Pair Classification	BurmeseXNLI (Conneau et al., 2018)	['mya']	['Non-fiction', 'Fiction', 'Government']	created	human-annotated	CC BY-NC 4.0
	IDKMRCNLI	['ind']	['Encyclopaedic', 'News', 'Written']	found		
	IndicXNLI* (Aggarwal et al., 2022)	['tam']	['Non-fiction', 'Fiction', 'Government']	found	expert-annotated	CC BY-NC 4.0
	IndoNLI* (Mahendra et al., 2021)	['ind']	['Encyclopaedic', 'Web', 'News', ...]	found	expert-annotated	CC BY-SA 4.0
	MultilingualNLI26lang2mil7 (Laurer et al., 2022)	['ind', 'vie']	['Non-fiction', 'Fiction', 'Government']	found	machine-translated and reviewed	
	MyXNLI (Htet and Dras, 2024)	['mya']	['Non-fiction', 'Fiction', 'Government']	found	human-annotated	CC BY-NC 4.0
	NewsPHNLI (Cruz et al., 2020a)	['fil']	['News', 'Written']	found	human-annotated	GPL-3.0
	PAWS	['fil']	['Web']	found	human-annotated	
	SQuADNLI	['ind']	['Encyclopaedic', 'News', 'Written']	found		
	TyDIQANLI	['ind']	['Encyclopaedic', 'News', 'Written']	found		
	WReTE (Setya and Mahendra, 2018)	['tha']	['Encyclopaedic', 'Web', 'News']	found	expert-annotated	MIT
	XNLI* (Conneau et al., 2018)	['tha', 'vie']	['Non-fiction', 'Fiction', 'Government']	found	expert-annotated	CC BY-NC 4.0
	XNLITranslated (Conneau et al., 2018)	['khm', 'zsm', 'lao']	['Non-fiction', 'Fiction', 'Government']	machine-translated and verified	machine-translated and reviewed	CC BY-NC 4.0

Table 15: The datasets included in SEA-BED (part 1).

Type	Name	Languages	Domains	Sample creation	Annotations creators	License
STS	Biosses (Soğancıoğlu et al., 2017)	['tha', 'mya']	['Medical']	created	human-annotated	GPL-3.0
	BiossesCrosslingual (Soğancıoğlu et al., 2017)	['tha', 'mya']	['Medical']	created	human-annotated	GPL-3.0
	IndicCrosslingual* (Ramesh et al., 2022)	['tam']	[News, Non-fiction, Web, ...]	found	expert-annotated	CC0-1.0
	SemRel2024* (Ousidhoum et al., 2024a)	['ind']	['Spoken', 'Written']	found	human-annotated	
	STS17 (Cer et al., 2017)	['tha', 'mya']	['News', 'Web', 'Written']	created	human-annotated	
	STS17Crosslingual (Cer et al., 2017)	['tha', 'mya']	['News', 'Web', 'Written']	created	human-annotated	
	STS22 (Chen et al., 2022)	['tha', 'mya']	['News', 'Written']	created	human-annotated	
	STS22Crosslingual (Chen et al., 2022)	['tha', 'mya']	['News', 'Written']	created	human-annotated	
	STS24 (Ousidhoum et al., 2024b)	['tha', 'mya']	['Spoken', 'Written']	created	human-annotated	
	STS24Crosslingual (Ousidhoum et al., 2024b)	['tha', 'mya']	['Spoken', 'Written']	created	human-annotated	
	STSBenchmark (Cer et al., 2017)	['ind', 'tha', 'vie', ...]	['News', 'Web', 'Written']	machine-translated and verified	machine-translated and reviewed	CC BY-SA 4.0
Clustering	EMoTES3K (Catapang and Visperas, 2023)	['fil']	['Morality', 'Written']	found	human-annotated	Apache license 2.0
	MurasuNews	['tam']	['News', 'Written']	found	derived	CC0
	News (Phatthiyaphaibun, 2025)	['lao']	['News', 'Written']	found	derived	
	News (Jiang et al., 2022)	['khm']	['News', 'Written']	found	derived	
	News (Chandra, 2020)	['ind']	['News', 'Written']	found	derived	
	News	['tam']	['News', 'Written']	found	derived	CC BY-SA 4.0
	News (Khine et al., 2017)	['mya']	['News', 'Written']	found	derived	
	SIB200 (Adelani et al., 2023)	['ind', 'tha', 'vie', ...]	['News', 'Written']	found	expert-annotated	CC BY-SA 4.0
	UITVION (Fujita and Perez-Meana, 2021)	['vie']	['Social', 'Written']	found	human-annotated	
	ViOCD (Nguyen et al., 2021b)	['vie']	['Reviews', 'Written']	found	human-annotated	
BitextMining	ALT (Riza et al., 2019)	['ind', 'tha', ...]	['News', 'Written']	found	expert-annotated	CC BY 4.0
	BibleNLP* (Akerman et al., 2023)	['ind', 'tha', 'vie', ...]	['Religious', 'Written']	found	expert-annotated	CC BY 4.0
	Flores* (Goyal et al., 2022)	['ind', 'tha', 'vie', ...]	['Non-fiction', 'Encyclopaedic', 'Written']	found	human-annotated	
	Embassy (Phatthiyaphaibun, 2020)	['tha', 'lao']	['Government', 'News']	found	human-annotated	CC0-1.0
	IN22Conv* (Gala et al., 2023)	['tam']	['Social', 'Spoken', 'Fiction', ...]	found	expert-annotated	CC BY 4.0
	IN22Gen* (Gala et al., 2023)	['tam']	['Web', 'Legal', 'Government', ...]	found	expert-annotated	CC BY 4.0
	IndoGeneral (Guntara et al., 2020)	['ind']	['General', 'Writen']	found	derived	CC BY-SA 4.0
	IndoIdentic (Gala et al., 2023)	['ind']	['News', 'Spoken', 'Web', ...]	found	derived	
	IndoNLG (Cahyawijaya et al., 2021)	['ind']	['religion']	found	derived	
	IndoNews (Guntara et al., 2020)	['ind']	['News', 'Written']	found	derived	CC BY-SA 4.0
	IndoReligious (Guntara et al., 2020)	['ind']	['Religion', 'Writen']	found	derived	CC BY-SA 4.0
	Liputan6 (Koto et al., 2020a)	['ind']	['News', 'Written']	found	human-annotated	CC BY-SA 4.0
	MADLAD400 (Kudugunta et al., 2023)	['tet']	['Web']	found	derived	ODC-BY
	NTREX* (Federmann et al., 2022)	['ind', 'tha', 'vie', ...]	['News', 'Written']	found	expert-annotated	CC BY-SA 4.0
	NusaxMiners* (Winata et al., 2023)	['ind']	['Reviews', 'Written']	found	human-annotated	CC BY-SA 4.0
	QED (Lamm et al., 2020)	['ind', 'tha', 'vie', ...]	['Education', 'Social', 'Spoken', ...]	found	human-annotated	CC BY-SA
	SCBMTEnTh2020 (Lowphansirikul et al., 2022)	['tha']	['conversation', 'Web', 'Government', ...]	found	human-annotated	CC BY-SA 4.0
	SoftwareDocumentation (Buschbeck and Exel, 2020)	['ind', 'tha', 'vie', ...]	['Web', 'Product']	found	expert-annotated	CC BY-NC 4.0
	TALPCo (Nomoto et al., 2018, 2019)	['ind', 'tha', 'vie', ...]	['Conversation', 'spoken']	found	human-annotated	CC BY-4.0
	Tatoeba* (Tiedemann, 2020)	['ind', 'tha', 'vie', ...]	['Written']	found	human-annotated	CC BY-2.0
	TED2020 (Reimers and Gurevych, 2020)	['ind', 'tha', 'vie', ...]	['Education', 'Social', 'Spoken', ...]	found	human-annotated	CC BY-NC-ND 4.0
	ThaiGov	['tha']	['Government', 'News']	found	human-annotated	PDDL
	USEmbassy (Phatthiyaphaibun et al., 2023)	['tha']	['News']	found	derived	CC0-1.0
	VSoLSCSum (Nguyen et al., 2016a)	['vie']	['Social', 'Written']	found	human-annotated	CC BY-4.0
	XLSum (Hasan et al., 2021)	['ind', 'tha', 'vie', ...]	['News', 'Written']	found	human-annotated	CC BY-NC-SA 4.0
Retrieval	ACIQuAD (Doxolodeo and Krisnadhi, 2024)	['ind']	['Encyclopaedic', 'Written']	found	expert-annotated	CC-BY 4.0
	Agricutlure1K (Min Si Thu,Khin Myat Noe)	['mya']	['Encyclopaedic', 'Written']	found	expert-annotated	CC BY-SA 4.0
	AskCovidDrBot (Aung and San, 2025)	['mya']	['Encyclopaedic', 'Written']	found	human-annotated	MIT
	ChatGPTOpenQA	['zsm']	['Encyclopaedic', 'Written']	found	LM-generated	CC BY-NC-SA 2.0
	ContextSearch (Nguyen et al., 2025)	['tha']	['STEM', 'Humanities', 'Social Sciences', ...]	found	human-annotated	MIT
	IAppWiki (Viriyayudhakorn and Polpanumas, 2021)	['tha']	['Encyclopaedic', 'Web', 'News']	found	expert-annotated	MIT
	IDKMRC (Putri and Oh, 2022)	['tnd']	['Encyclopaedic', 'Written']	found	human-annotated	CC BY-SA 4.0
	IndicQA* (Doddapaneni et al., 2022)	['tam']	['Web', 'Written']	machine-translated and verified	human-annotated	CC BY 4.0
	IndoNLG (Cahyawijaya et al., 2021)	['ind']	['Religion', 'Writen']	found	human-annotated	CC BY-SA 4.0
	IndoQA (Jakarta Artificial Intelligence Research, 2023)	['ind']	['Web']	found	expert-annotated	CC BY-ND 4.0
	MLDR (Chen et al., 2024)	['tha']	['Encyclopaedic', 'Written']	found	LM-generated	MIT
	MLQA* (Lewis et al., 2019)	['vie']	['Encyclopaedic', 'Written']	found	human-annotated	CC BY-SA 3.0
	MIRACL* (Zhang et al., 2023)	['ind', 'tha']	['Encyclopaedic', 'Written']	found	expert-annotated	Apache license 2.0
	Microbiology1K (Si Thu, 2024)	['mya']	['Encyclopaedic', 'Written']	found	human-annotated	CC BY-SA 4.0
	QASiNa (Rizqullah et al., 2023)	['ind']	['Religion', 'Writen']	found	human-annotated	MIT
	ThaiWikiQA (Trakultaweekoon et al., 2019)	['tha']	['Encyclopaedic', 'Written']	found	human-annotated	CC BY-NC-SA 3.0
	TyDiQA (Clark et al., 2020)	['ind', 'tha']	['Encyclopaedic', 'Written']	found	human-annotated	Apache license 2.0
	ViQuAD2_0 (Nguyen et al., 2022)	['vie']	['Encyclopaedic', 'Written']	found	expert-annotated	MIT
	WangchanXLegalThaiCCLRAG (Akarajaradwong et al., 2025)	['tha']	['Legal', 'Written']	found	human-annotated	MIT
	XQuAD (Artetxe et al., 2019)	['tha', 'vie']	['Web', 'Written']	found	human-annotated	CC BY-SA 4.0
Instruction Retrieval	AlpacaInstruct	['ind']	None	found	LM-generated	Apache license 2.0
	Vietnamese52KAlpaca (Nhiem, 2023)	['vie']	None	found	LM-generated	
	WangchanThaiInstruct	['tha']	['Medical', 'Finance', 'Legal', ...]	found	human-annotated	CC BY-SA 4.0
	WangchanXSyntheticInstructThai120k (Pengpun et al., 2024)	['tha']	['Encyclopaedic', 'Written']	found	LM-generated	MIT
Reranking	MIRACL (Zhang et al., 2023)	['ind', 'tha']	['Encyclopaedic', 'Written']	found	expert-annotated	Apache license 2.0

Table 16: The datasets included in SEA-BED (part 2).