BILINGUAL PUBLISHING GROUP

Forum for Linguistic Studies
https://journals.bilpubgroup.com/index.php/fls

ARTICLE

Advancing Thai Sentence Embedding: Benchmark Development

Panuthep Tasawong* , ***Peerat Limkonchotiwat*** , ***Wuttikorn Ponwitayarat*** , ***Surapon Nonesung*** , ***Sitiporn Sae Lim*** , ***Chayapat Uthayopas*** , ***Can Udomcharoenchaikit*** , ***Sarana Nutanong***

School of Information Science and Technology, Vidyasirimedhi Institute of Science and Technology (VISTEC), Rayong 21210, Thailand

ABSTRACT

Sentence embedding is the task of capturing textual information in contextualized vectors, which has attracted considerable attention in recent years due to its effectiveness in a wide range of downstream NLP applications, such as classification, retrieval, and semantic search. Despite substantial progress, particularly for English, the study of sentence embeddings in resource-constrained languages like Thai remains underexplored. Existing Thai benchmarks are limited in scope, as they primarily evaluate models on text classification, leaving other important tasks insufficiently examined. To address this gap, we introduce the Thai Sentence Embedding Benchmark, a comprehensive evaluation suite covering diverse tasks including semantic textual similarity (STS), text classification, pairwise classification, and retrieval. We systematically collect and reformat high-quality Thai texts into embedding-based tasks, ensuring robust and standardized evaluation. Furthermore, we propose a new dataset, Thai STS, specifically designed to fill a crucial gap in evaluating semantic similarity in Thai. Beyond benchmarking, we present new Thai sentence embeddings trained under four different sentence embedding frameworks designed for low-resource settings, with three model sizes spanning monolingual and multilingual encoder-based architectures. This variety enables meaningful insights into the trade-offs between scale, architecture, and resource constraints. Through extensive experiments, we evaluate a broad spectrum of

*CORRESPONDING AUTHOR:
Panuthep Tasawong, School of Information Science and Technology, Vidyasirimedhi Institute of Science and Technology (VISTEC), Rayong 21210, Thailand; Email: panuthep.t_s20@vistec.ac.th

ARTICLE INFO
Received: 10 September 2025 | Revised: 27 September 2025 | Accepted: 29 September 2025 | Published Online: 19 November 2025
DOI: https://doi.org/10.30564/fls.v7i12.12023

CITATION
Limkonchotiwat, P., Tasawong, P., Ponwitayarat, W., et al., 2025. Advancing Thai Sentence Embedding: Benchmark Development. Forum for Linguistic Studies. 7(12): 1380–1397. DOI: https://doi.org/10.30564/fls.v7i12.12023

embedding models, including newly developed large language models (LLMs), smaller language models (SLMs), and off-the-shelf API-based systems. Our findings highlight both strengths and persistent challenges across tasks, providing guidance for future work. All datasets, models, and code are released under the Apache-2.0 License to support open, reproducible, and community-driven progress in Thai NLP community.

Keywords: Sentence Embedding Evaluation; Text Classification; Retrieval; Semantic Textual Similarity

1. Introduction

Sentence embedding is a foundational task of many Natural Language Processing (NLP) applications, such as text classification, clustering, and information retrieval. The task of sentence embedding is to transform a sentence into a fixed-length vector that encapsulates the semantics of the sentence. Sentence embedding models utilize transformer-based Pre-trained Language Models (PLMs) like BERT [1] and RoBERTa [2] that were trained on a large-scale corpus of unlabeled data using unsupervised pre-training methods, i.e., Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). In addition, unsupervised fine-tuning techniques, such as SimCSE [3], have been developed to enhance the performance of sentence embedding models without relying on labeled data. Recently, researchers have proposed techniques to convert generative Large Language Models (LLMs) into sentence embedding models, e.g., E5 Mistral 7B [4], GritLM 7B [5], and gte-Qwen2 7B [6]. These models have achieved new state-of-the-art performance in MTEB [7], a standard English sentence embedding benchmark.

To evaluate the effectiveness of sentence embedding, researchers have created various benchmarks, SentEval [8], BEIR [9], and MTEB [7], which involve a suite of tasks designed to evaluate the performance of sentence embedding models on three different aspects.

- **Intrinsic Evaluation Performance.** Measures the quality of embeddings produced by the sentence embedding model by assessing their performance on specific NLP tasks that are related to the embedding space itself, such as semantic textual similarity (STS), and natural language inference (NLI).
- **Downstream Task Performance.** Evaluates how well the sentence embedding model performs on various downstream NLP tasks, such as text classification and retrieval.
- **Efficiency.** Assesses the computational efficiency of the model, including inference speed.

These standard benchmarks serve two vital purposes: (i) They enable researchers to systematically study the strengths and weaknesses of different models, leading to the development of better models. (ii) They facilitate industry practitioners in identifying the most suitable models for specific downstream tasks, ensuring accurate model selection for real-world applications. However, these established benchmarks focus solely on English, while a more resource-constrained language like Thai must contend with less comprehensive benchmarks.

The major challenge in the development of Thai sentence embedding is the lack of a comprehensive benchmark. The existing benchmark for Thai sentence embedding, namely Thai Text Classification Benchmark [10], only assesses the text classification performance, neglecting the intrinsic evaluation and other downstream task performance. Addressing the gap between Thai and English sentence embedding benchmarks is crucial for advancing the development of NLP applications in the Thai language.

In this paper, we propose a new *Thai sentence embedding benchmark*. The benchmark consists of 8 datasets covering 4 tasks: (i) STS, (ii) Text classification, (iii) Pair classification, and (iv) Retrieval QA. We aggregate existing Thai and multilingual datasets for text classification, pair classification, and retrieval tasks. We found that there is no dataset for the STS task in the Thai language. Thus, we created one by translating from the STS-B dataset [11]. We evaluate 36 sentence embedding models using our *Thai sentence embedding benchmark* to study the strengths and weaknesses of each sentence embedding model in the Thai language. Specifically, we assess the performance of newly developed multilingual and Thai sentence embedding models (LLMs and SLMs) and off-the-shelf embedding APIs. In addition, we evaluate the effectiveness of variant unsupervised fine-tuning techniques, such as SimCSE, on multiple pre-trained models.

The experimental results show that no individual model dominates all tasks. Models that exhibit strong performance in text classification and retrieval tasks tend

to demonstrate weaker performance in STS and pair classification tasks, and vice versa. Interestingly, we observe no LLM-based sentence embedding models among the top performers in the *Thai sentence embedding benchmark*. The GRITLM 7B and gte-Qwen2 7B, which are considered among the top performers in MTEB, exhibit the poorest performance compared to other LLMs in our benchmark. These results reveal the disparity in language generalization among the LLM-based sentence embedding models. Further, we find that non-LLM sentence embedding models (i.e., the model with less than 600M parameters) can achieve performance on par with LLMs through unsupervised fine-tuning methods. Our benchmarking study reveals valuable insights into the strengths and weaknesses of various sentence embedding models and embedding API services. We hope our work will facilitate model selection and advance the future development of Thai NLP. We make all of our benchmark's source code, datasets, and fine-tuned models publicly available.

Our contributions can be summarized as follows:

- We propose a new *Thai sentence embedding benchmark*, the first comprehensive benchmark for evaluating Thai sentence embedding. The benchmark covers diverse datasets across 4 tasks: (i) STS, (ii) Text classification, (iii) Pair classification, and (iv) Retrieval QA, for intrinsic and downstream task performance evaluation.
- We assess the performance of 36 newly developed multilingual and Thai sentence embedding models, including LLM-based, SLM-based, and off-the-shelf embedding APIs. We evaluate the effectiveness of different unsupervised fine-tuning methods on multiple pre-trained models. Notably, this paper is the first to evaluate the performance of LLM-based sentence embedding models on diverse Thai sentence embedding tasks.
- The experimental findings reveal (i) the best embedding model for each Thai sentence embedding task, (ii) the disparity of top LLMs' performance on Thai and English sentence embedding, and (iii) the success and failure cases of different unsupervised fine-tuning techniques on Thai sentence embedding tasks.
- The source code, datasets, and fine-tuned models used in this study are publicly available for convenient assessment of future methods.

2. Related Work

2.1. Sentence Embedding

Training sentence embeddings often involves contrastive learning with unlabeled text data. Contrastive learning consists of three major components: (i) anchor, (ii) positive, and (iii) negative. The contrastive learning objective is to maximize the similarity between anchor and positive samples while minimizing the similarity between anchor and negative samples. In general, the negative samples are obtained using in-batch negative samples.

Gao et al. [3] propose a simple contrastive learning method called SimCSE. The anchor and positive samples of SimCSE are identical instances but undergo different dropout strategies. Wongso et al. [12] applied contrastive learning for Indonesian with the training pipeline from SimCSE [3]. Their experiment demonstrated that unsupervised contrastive learning yields comparable performance to supervised learning. Additionally, Wang et al. [13] extended SimCSE to a cross-lingual setting called mSimCSE, where the anchor and positive samples are written in different languages. The experimental results demonstrated that mSimCSE outperformed fully supervised multilingual sentence embedding in cross-lingual retrieval benchmarks. These studies show that leveraging similar architectures, such as BERT [1] or RoBERTa [2], across different languages can lead to effective and transferable techniques for sentence embedding.

Recently, researchers used multilingual sentence embedding to achieve reasonable performance in high- and low-resource languages. Notably, Chen et al. [14] proposed BGE-M3, a model trained on extensive parallel datasets spanning over 100 languages to achieve state-of-the-art multilingual text retrieval. However, the performance of Thai within this model is notably lower than other languages.

2.2. Benchmark for Sentence Embedding

To assess the effectiveness of sentence embedding techniques, researchers have developed benchmarks to study the performance and robustness of these models.

Conneau and Kiela [8] proposed an evaluation toolkit for sentence embedding called SentEval. The toolkit consists of two main tasks for sentence embedding works: (i) downstream tasks (STS and text classification datasets) and (ii) probing tasks (e.g., evaluating what linguistic properties are encoded in sentence embedding). Thakur et al. [9] introduced BEIR, a benchmark for diverse information retrieval tasks. The benchmark encompasses 9 tasks across 18 datasets, serving to evaluate the robustness of sentence embedding models. Muennighoff et al. [7] proposed a massive text embedding benchmark called MTEB. The benchmark consists of 58 datasets covering 112 languages from 8 embedding tasks. Although there are various benchmarks and evaluation tools for sentence embedding in multilingual settings, Thai is not included as the main task, except for the bi-text mining task.

Furthermore, there are attempts to create a Thai benchmark for sentence embedding. Charin and Phasathorn [10] proposed a Thai text classification benchmark. The benchmark only evaluates the text classification, while other tasks, i.e., retrieval and text understanding, are omitted.

3. Thai Sentence Embedding Benchmark

3.1. Desired Properties

Drawing from the desiderata of previous benchmarks [7], we seek to build a Thai Sentence Embedding Benchmark on the same set of desiderata:

- **Diversity.** The benchmark shall provide a diverse range of tasks for evaluating sentence embedding models in various use cases.
- **Simplicity.** The benchmark shall provide a simple API for evaluating new models.
- **Extensibility.** The benchmark shall provide a simple API for adding new datasets.
- **Reproducibility.** The evaluation results of the benchmark shall be reproducible.

3.2. Tasks and Evaluation

The Thai Sentence Embedding Benchmark offers the first comprehensive evaluation of Thai text across various tasks, unlike the existing Thai benchmark [10] that evaluated Thai sentence embedding only in text classification datasets. Our benchmark comprises a diverse set of evaluation tasks as shown in **Table 1**. In addition, we also assess the runtime of each model to ascertain the trade-off between performance and computational demand. **Figure 1** illustrates an overview of each task.

Semantic Textual Similarity (STS). In this task, we assess the generalizability of language understanding using Semantic Textual Similarity (STS). STS allows us to test generalizability using Wikipedia data for training and non-Wikipedia data for testing. The main objective is to determine the similarity between pairs of sentences, where similarity scores range from 0 to 5 (0 denoting dissimilarity and 5 indicating high similarity). To evaluate, the provided embedding model generates embeddings for each sentence pair, and similarity is computed using cosine similarity.

However, we found that there are no available STS datasets in Thai. To address this, we translated the STS-B dataset [11] (both development and test data) using Google NMT. Then, we employ experienced Thai-English translators to review the translation. They (i) check for grammatical correctness, (ii) ensure that the intended meaning is preserved, and (iii) adjust terminology or phrasing where necessary to improve naturalness. This process ensures that the translations remain both accurate and fluent. Furthermore, we use Spearman's correlation coefficient, based on cosine similarity, as the primary evaluation metric.

Text Classification. To evaluate the adaptability of embedding models, we adopt the transfer learning task from Thai text classification benchmarks [10] for sentence embedding. We begin by embedding both the training and test data using the provided model. Subsequently, we employ a logistic regression model trained on the training data. We then use the trained classifier model to categorize the sentiment or class of the test data. This allows us to test the adaptability through the transfer learning process. For this evaluation, we employ seventeen high-quality Thai text classification datasets: CyberbullyingLGBT, Depression [15], Emoji, General-Amy [16], Generated Review [17], Krathu500, LimeSoda [18], MassiveIntent [19], MassiveScenario [19], MultiLingual-Sentiment [20], ReviewShopping

[16], SIB200 [21], SEATran-translationese Resampled [22], TCAS61 [16], The40 ThaiChild-renStories [23], Wisesight Sentiment [24], and Wongnai-Review. The evaluation metrics provided are accuracy and F1 scores.

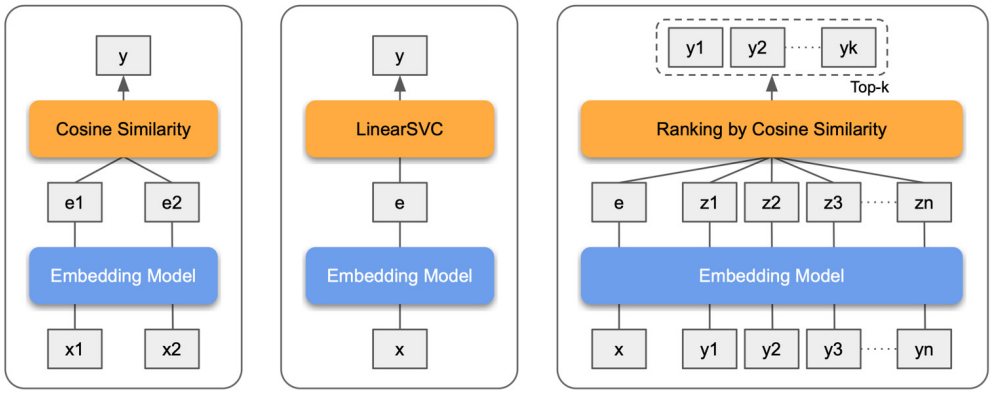
Pair Classification. To assess the generalizability of the provided model, we conduct zero-shot pair text classification. Unlike traditional classification tasks, this evaluation does not involve a linear classification head or any training data. In this task, the objective is to generate embeddings for pairs of sentences such that similar pairs have a cosine similarity of 1 while dissimilar pairs have a cosine similarity of 0. Following the approach outlined by Liu et al. [25], we utilize the provided embedding model to embed sentence pairs and then calculate the cosine similarity between these embeddings. Then, we compute the Average Precision (AP) score by comparing the cosine similarity with the gold labels. We employ XNLI [26] for this task. To adapt XNLI for our task, we map entailment labels to 1, contradiction labels to 0, and omit neural labels from the data.

Retrieval. In this task, we assess the provided model's generalizability for a retrieval task, specifically utilizing the retrieval QA setting [27–30] to evaluate its effectiveness in scenarios akin to the retrieval-augmented generation (RAG) framework. In this setting, given a query and a set of documents within each dataset, the objective is to associate each query with relevant documents from the datasets. The provided model is tasked with embedding all queries and documents, and cosine similarity is utilized to rank them from most similar to least similar. We use IAppWiki [31], MLDR [14], MIRACL [32], ThaiWikiQA [33], TyDiQA [34], WangchanXLegalThaiCCLRAG [35], and XQuAD [27] as the retrieval datasets. We employ recall@1 and MRR@10 metrics to evaluate the retrieval and ranking performance.

Table 1. Dataset statistics of train/dev/test of each dataset.

Dataset	Task	#Train	#Dev	#Test
STS-B	Semantic Understanding	-	1499	1379
CyberbullyingLGBT (CBLGBT)	Text classification	20,000	-	-
Depression (DEP)	Text classification	25,100	3340	5020
Emoji (EMO)	Text classification	128	-	55
GeneralAmy (GAMY)	Text classification	90	-	-
Generated Review (GRENTH)	Text classification	141,369	15,708	17,453
Krathu500 (K500)	Text classification	5700	-	-
LimeSoda (LS)	Text classification	2700	300	2770
MassiveIntent (MINT)	Text classification	11,500	2030	2970
MassiveScenario (MSCN)	Text classification	11,500	2030	2970
MultiLingualSentiment (MLS)	Text classification	8100	1150	2340
ReviewShopping (RSHOP)	Text classification	128	-	-
SIB200 (SIB200)	Text classification	701	99	204
SEATranslationeseResampled (SEATR)	Text classification	15,000	-	5980
TCAS61 (TCAS61)	Text classification	123	-	-
The40ThaiChildrenStories (40TCS)	Text classification	1960	-	-
WisesightSentiment (WSENT)	Text classification	21,628	2404	2671
WongnaiReview (WREV)	Text classification	36,000	4000	6203
XNLI	Pair classification	-	1660	3340
IAppWiki	Retrieval	5760	742	739
MLDR	Retrieval	1970	200	200
MIRACL	Retrieval	2972	733	992
ThaiWikiQA	Retrieval	17,000	-	-
TyDiQA	Retrieval	3809	763	-
WangchanXLegalThaiCCLRAG (ThaiCCLRAG)	Retrieval	8210	-	3740
XQuAD	Retrieval	-	-	1190

Notes: We use the test set for experiments. If a test set is not available, we split 30% from the training set, except for MIRACL and TyDiQA, which use the development set instead.



(**a**) STS & Pair Classification. (**b**) Text Classification. (**c**) Retrieval.

Figure 1. Visualization of each sentence embedding task in the Thai Sentence Embedding Benchmark.

3.3. Benchmark Usage

Our benchmark is designed to seamlessly integrate with HuggingFace's models and datasets, enabling easy evaluation of new models by simply providing the HuggingFace model name (e.g., *BAAI/bge-m3*), through our simple API 2. Similarly, new datasets can also be easily added by specifying the task name and the HuggingFace dataset name (e.g., *miracl/miracl*) in a configuration file. Additionally, we also provide a script to reproduce our benchmarking results.

4. Experimental Setup

4.1. Models

We evaluate three types of sentence embedding models: (i) SLMs, (ii) LLMs, and (iii) Off-the-shelf embedding APIs. All models (except embedding APIs) are openly available on HuggingFace. To ensure reproducibility, we provide a list of URIS for all the models used in our experiments in the Data Availability Statement section.

4.1.1. SLMs

- *XLMR* [36]. A multilingual version of the RoBERTa [2] model pre-trained on 2.5TB of filtered CommonCrawl data [37] containing 100 languages. The model has a vocabulary size of 250,002, where 4274 are Thai subwords. The model has two versions, XLMR-base and XLMR-large, which allow us to study the impact of increasing model size directly.
- *WangchanBERTa* [38] A Thai version of the RoBERTa-base model pre-trained on 78.5GB of Thai assorted texts. The model has a vocabulary size of 25,005, where 22,200 are Thai subwords.
- *PhayaThaiBERT* [39] An improvement over WangchanBERTa to understand code-switching and unassimilated loanwords by expanding Wangchan-BERTa's vocabulary to support foreign words and continual pre-training. The model was pre-trained on 156.3GB of Thai assorted texts and has a vocabulary size of 249,262, where 22,200 are Thai subwords.
- *MPNet-multilingual* [40] We employ a common use of Thai text embedding. This model was trained on multilingual corpora and supported over 50 languages.
- *DistilUSE-multilingual* [40] We also evaluate a small multilingual text embedding. This model was trained by a knowledge distillation framework from Reimers and Gurevych [41], leveraging massive multilingual datasets.
- *BGE-M3* [14] A multilingual sentence embedding model that supports more than 100 languages. The model was pre-trained on massive unsupervised multilingual data and then fine-tuned specifically for retrieval tasks using large supervised multilingual datasets.

4.1.2. LLMs

- *E5 Mistral-7b-instruct* [4]: A language embedding model, based on Mistral-7B v0.1 with 7 billion parameters, has been improved through instruction fine-tuning with a diverse set of multilingual data-

sets, using synthetic data and 13 public datasets with standard contrastive loss. This enhancement has given the model multilingual capabilities. However, since Mistral-7B-v0.1 was primarily trained on English data, it is recommended that this model be primarily used for English text processing tasks.

- *gte-Qwen2-7B-instruct* [6]: This model belongs to the General Text Embedding (GTE) family, developed by Alibaba with 7 billion parameters. It is based on Qwen2 [27] and incorporates several key advancements, such as the integration of bidirectional attention mechanisms and comprehensive training across a vast, multilingual text corpus spanning diverse domains and scenarios. This training leverages both weakly supervised and supervised data.
- *GritLM-7B* [5]: A generative representational instruction-tuned language model with 7 billion parameters, which combines text representation (embedding) and text generation in a single model, achieving state-of-the-art performance in both areas.
- *Llama-3-8B* and *Llama-3-8B-Instruct* [42] models, originally developed by Meta, were pre-trained on over 15 trillion tokens with 8 billion parameters on next token prediction. For the instruction-tuned version, this model utilizes both supervised fine-tuning (SFT) and Direct Preference Optimization (DPO) [43] for its alignment training.
- *Llama-3.1-8B* and *Llama-3.1-8B-Instruct* [43]: These models were enhanced from the Llama-3 version, incorporating several key features: multilingual support, an expanded context window, advanced capabilities for synthetic data generation, and fine-tuning for tool use. The instruction-tuned version applies both the SFT and DPO algorithms.
- Llama-3-Typhoon-v1.5-8B-instruct [44] This model is part of a series of Thai large language models (LLMs) based on Llama3, developed by SCB10X specifically for the Thai language with 8 billion parameters. It utilizes supervised fine-tuning (SFT) for alignment training.

4.1.3. Off-the-Shelf Embedding APIs

- *Cohere-embed-multilingual-v2.0 and v3.0.* evaluate an off-the-shelf multilingual embedding model. The inference speed was measured by running each model on our datasets using a 1x Nvidia H100 80GB (**Figure 2**).

API service from Cohere that supports more than 100 languages and multiple downstream tasks.

- *Openai-text-embedding-3-large.* We also evaluate an off-the-shelf multilingual embedding API service from OpenAI.

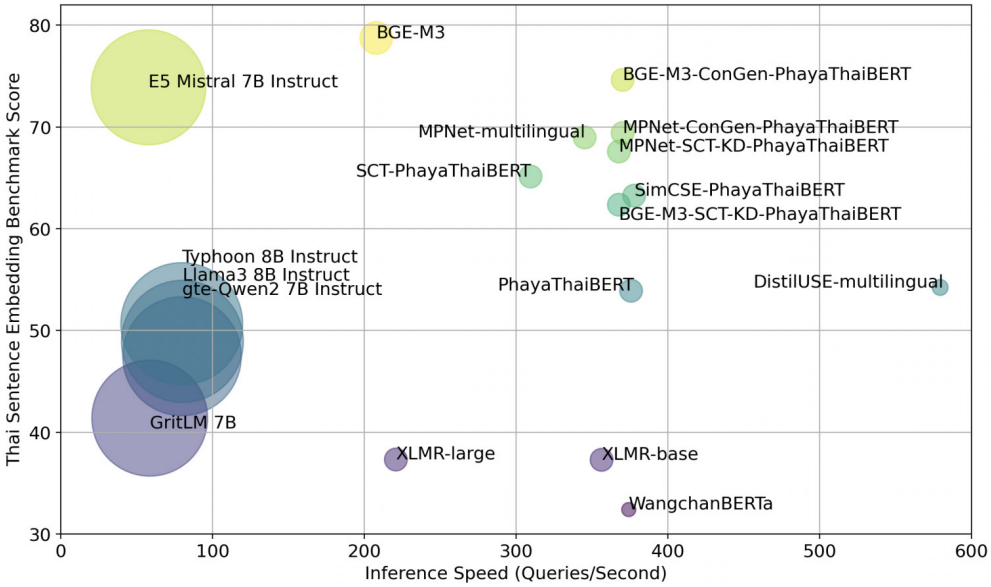


Figure 2. Performance, inference speed, and size of model parameters (size of the circles) of each sentence embedding model.

4.2. Unsupervised Fine-Tuning Methods

In addition, we assess the effectiveness of newly developed unsupervised fine-tuning techniques for improving sentence embedding models in the Thai language.

- *SimCSE* [3] A contrastive learning method that utilizes different random dropouts as the data augmentation scheme.
- *SCT* [45] A cross-view training framework that is designed for small PLMs. The cross-view pipeline improves the guidance by adding complex tasks to the training process, e.g., comparing a representation with a large scale of negative samples.
- *SCT-KD* [45] An unsupervised knowledge distillation that changes from self-supervised to knowledge distillation by replacing the representation from itself with a larger model.
- *ConGen* [46] An unsupervised distillation method uses many negative representations produced from a teacher model as the knowledge distillation reference.

For each method, we trained three pre-trained language models, i.e., WangchanBERTa, PhayaThai-BERT, and XLMR-base. The training was conducted using the original publicly available codes from their papers. For the distillation methods, ConGen and SCT-KD, we employ MPNet-multilingual as a teacher model. We also study the effect of changing the teacher model from MPNet-multilingual to BGE-M3. We employed 1 million sentences from Wikipedia provided by Phatthiyaphaibun et al. [16] as a training dataset and STS-B development set for validation, following the SimCSE [3] methodology. We use grid search to find hyperparameters, including learning rate, epoch, and batch size (**Table 2**).

Table 2. Hyper-parameter setting.

Parameter	Value
Learning rate	5×10^{-4}, 3×10^{-4}, 1×10^{-4}, 5×10^{-5}, 1×10^{-5}
Batch size	32, 64, 128, 256, 512
Temperature	0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07
Instance queue (ConGen and SCT)	128, 1024, 16, 384, 65, 536, 131, 072

5. Experimental Results

Figure 2 illustrates the overall performance and inference speed of each sentence embedding model. **Table 3** reports the results for each sentence embedding task. **Table 4** shows the detailed results across various text classification datasets. **Table 5** provides the detailed results on different retrieval datasets.

Table 3. Results of each model on STS (Spearman's rank correlation), text classification (F1), pair classification (Average Precision), and retrieval QA (R@1 / MRR@10).

Task (→) Metric (→) Dataset# (→)	STSSpear.1	TextCLF F1 17	PairCLF AP 1	Retrieval R@1/MRR@10 7	Average 26
SLMs					
XLMR-base 0.28B	47.49	71.52	57.62	3.92 / 5.90	37.29
XLMR-large 0.56B	41.80	71.19	54.56	7.74 / 11.31	37.32
WangchanBERTA 0.11B	22.91	58.04	52.96	11.65 / 16.45	32.40
DistilUSE-multilingual 0.14B	68.45	68.71	65.94	30.21 / 37.83	54.23
PhayaThaiBERT 0.28B	54.20	72.81	59.67	36.86 / 46.00	53.91
MPNet-multilingual 0.28B	**82.92**	73.73	**84.14**	48.01 / 56.05	68.97
BGE-M3 (dense only) 0.57B	80.62	<u>75.77</u>	79.02	**76.81 / 81.54**	**78.75**
Self-supervised SLMs					
SimCSE-XLMR-base 0.28	68.85	65.49	61.87	40.20 / 47.07	56.70
SimCSE-WangchanBERTa 0.11B	64.65	68.15	59.14	40.11 / 47.89	55.99
SimCSE-Phaya ThaiBERT 0.28B	71.54	70.08	63.35	52.04 / 59.24	63.25

Table 3. *Cont.*

Task (→) Metric (→) Dataset# (→)	STSSpear.1	TextCLF F1 17	PairCLF AP 1	Retrieval R@1/MRR@10 7	Average 26
SCT-XLMR-base 0.28B	71.54	70.34	66.49	37.72 / 46.20	58.46
SCT-WangchanBERTa 0.11B	73.86	72.38	67.04	45.91 / 54.97	62.83
SCT-PhayaThaiBERT 0.28B	76.93	72.38	65.87	51.00 / 59.4	65.13
Knowledge distillation SLMs (MPNet-multilingual as teacher model)					
MPNet-SCT-KD-XLMR-base 0.28B	80.91	72.60	79.78	47.14 / 55.81	67.25
MPNet-SCT-KD-WangchanBERTa 0.11B	80.14	72.07	77.04	44.30 / 53.12	65.33
MPNet-SCT-KD-PhayaThaiBERT 0.28B	80.48	73.03	77.84	49.01 / 57.58	67.59
MPNet-ConGen-XLMR-base 0.28B	81.76	72.81	81.47	51.27 / 59.66	69.39
MPNet-ConGen-Wangchan BERTA 0.11B	81.59	73.43	82.43	50.16 / 58.48	69.22
MPNet-ConGen-PhayaThaiBERT 0.28B	81.75	73.30	81.01	51.32 / 59.63	69.40
Knowledge distillation SLMs (BGE-M3 as teacher model)					
BGE-M3-SCT-KD-XLMR-base 0.28B	68.56	71.98	65.58	30.84 / 38.75	55.14
BGE-M3-SCT-KD-WangchanBERTa 0.11B	69.49	71.80	64.16	37.70 / 45.18	57.67
BGE-M3-SCT-KD-PhayaThaiBERT 0.28B	70.59	71.87	64.80	40.34 / 48.03	59.13
BGE-M3-ConGen-XLMR-base 0.28B	78.63	73.59	76.31	69.11 / 74.91	74.51
BGE-M3-ConGen-WangchanBERTa 0.11B	78.63	73.46	76.31	67.67 / 73.74	73.96
BGE-M3-ConGen-PhayaThaiBERT 0.28B	79.75	74.57	76.13	70.08 / 75.53	75.21
LLMS					
E5 Mistral 7B Instruct	79.72	73.57	68.04	71.02 / 77.16	73.90
gte-Qwen2 7B Instruct	53.77	74.16	61.73	20.56 / 27.12	47.47
GritLM 7B	48.57	69.42	56.40	13.57 / 18.97	41.39
Llama3 8B	51.50	71.71	57.76	27.65 / 35.62	48.85
Llama3 8B Instruct	53.55	71.57	58.04	26.90 / 34.59	48.93
Llama3.1 8B	52.20	71.97	71.97	26.84 / 34.66	48.76
Llama3.1 8B Instruct	53.65	72.03	57.47	25.58 / 33.53	48.45
Typhoon 8B Instruct	56.85	72.29	58.05	28.89 / 37.21	50.66
Off-the-shelf embedding APIs					
Cohere-embed-multilingual-v2.0	72.78	72.71	72.71	68.24 / 73.76	69.90
Cohere-embed-multilingual-v3.0	81.44	**76.21**	73.28	71.34 / 76.00	75.65
Openai-text-embedding-3-large	73.85	73.48	67.33	67.20 / 73.61	71.09

Notes: **Bold** and Underline indicate the best and second performers.

Table 4. The F1 of each model on seventeen Thai text classification datasets.

Dataset (→) Metric (→)	CBLGBT F1.	DEP F1.	EMO F1.	GAMY F1.	GRENTH F1.	K500 F1.	LS F1.	MINT F1.	MSCN F1.	MLS F1.	RSHOP F1.	SIB200 F1.	SEATR F1.	TCAS61 F1.	40TCS F1.	WSENT F1.	WREV F1.	Avg F1.
SLMs																		
XLMR-base 0.28B	99.78	75.97	71.45	85.23	55.91	84.96	5.80	77.54	85.12	84.34	92.30	86.09	31.87	88.89	71.18	66.57	52.78	71.52
XLMR-large 0.56B	99.83	77.46	65.27	70.21	57.72	85.49	5.96	78.87	86.47	84.96	**100.00**	84.68	32.25	88.89	71.31	66.78	54.04	71.19
WangchanBERTA 0.11B	91.47	75.64	32.78	55.68	52.97	79.43	5.38	41.04	51.49	75.38	71.23	68.76	31.85	86.24	53.93	60.46	52.95	58.04
DistilUSE-multilingual 0.14B	92.39	74.55	85.63	81.00	48.89	84.45	5.67	73.83	83.00	80.57	87.18	75.16	33.78	85.95	72.18	63.78	40.11	68.71
PhayaThaiBERT 0.28B	99.67	77.52	76.15	85.23	56.63	84.08	5.96	79.65	85.95	**87.24**	97.44	82.88	31.72	94.54	70.00	69.56	53.51	72.81
MPNet-multilingual 0.28B	98.27	76.03	90.28	**96.31**	56.34	85.35	5.73	80.23	87.53	84.12	**100.00**	86.00	33.61	83.33	75.86	67.60	46.74	73.73
BGE-M3 (dense only) 0.57B	99.75	**78.40**	89.13	92.61	58.36	86.12	5.99	81.91	88.94	87.20	97.43	**92.12**	33.95	97.31	**77.44**	68.85	52.65	75.77

Table 4. *Cont.*

Dataset (→) Metric (→)	CBLGBT F1.	DEP F1.	EMO F1.	GAMY F1.	GRENTH F1.	K500 F1.	LS F1.	MINT F1.	MSCN F1.	MLS F1.	RSHOP F1.	SIB200 F1.	SEATR F1.	TCAS61 F1.	40TCS F1.	WSENT F1.	WREV F1.	Avg F1.
Self-supervised SLMs																		
SimCSE-XLMR-base 0.28	97.84	71.19	69.37	70.45	46.42	80.07	5.23	72.93	81.88	79.76	82.05	76.19	35.53	83.61	58.90	63.45	38.46	65.49
SimCSE-WangchanBERTa 0.11B	99.27	75.35	71.57	74.07	52.44	81.53	5.31	73.47	81.03	83.03	89.61	73.57	34.83	83.61	62.18	67.01	50.68	68.15
SimCSE-PhayaThaiBERT 0.28B	99.37	75.42	85.29	69.95	50.75	83.03	4.98	77.07	84.37	85.40	**100.00**	77.41	34.36	91.74	62.48	67.73	42.03	70.08
SCT-XLMR-base 0.28B	99.24	75.81	81.36	74.15	55.26	82.15	5.46	78.48	85.91	83.54	**100.00**	80.75	31.11	83.33	66.74	66.45	46.09	70.34
SCT-Wangchan-BERTa 0.11B	99.75	77.20	83.38	81.48	56.29	82.37	5.30	78.31	86.17	86.34	94.87	79.93	31.98	97.29	68.52	70.67	50.70	72.38
SCT-PhayaThaiBERT 0.28B	99.80	77.75	86.94	74.07	56.23	83.53	5.46	80.03	86.48	86.59	**100.00**	80.25	35.35	94.54	71.10	**69.81**	50.27	72.38
Knowledge distillation SLMs (MPNet-multilingual as teacher model)																		
MPNet-SCT-KD-XLMR-base 0.28B	98.75	75.63	84.77	96.28	56.58	82.44	5.53	77.06	85.81	84.73	**100.00**	80.18	31.40	91.92	69.06	67.34	46.68	72.60
MPNet-SCT-KD-WangchanBERTa 0.11B	98.35	75.45	88.71	92.59	55.93	82.51	5.45	76.02	83.91	85.92	97.43	75.77	31.75	91.74	70.55	66.71	46.34	72.07
MPNet-SCT-KD-PhayaThaiBERT 0.28B	98.32	76.49	85.72	96.28	56.50	83.05	5.39	76.76	85.20	85.78	**100.00**	77.77	31.96	94.54	72.38	68.54	46.77	73.03
MPNet-ConGen-XLMR-base 0.28B	98.25	75.85	87.26	88.83	56.66	84.71	5.75	80.05	87.10	86.00	**100.00**	82.54	28.46	88.89	73.46	67.44	46.59	72.81
MPNet-Con-Gen-Wangchan BERTA 0.11B	98.20	76.67	87.12	85.19	57.49	84.91	5.81	79.62	87.59	85.58	**100.00**	83.75	30.73	94.54	74.19	67.59	49.39	73.43
MPNet-Con-Gen-PhayaThaiBERT 0.28B	98.59	76.51	81.35	96.28	57.21	84.77	5.86	79.85	87.46	86.30	**100.00**	82.34	31.32	89.07	73.54	68.15	47.54	73.30
Knowledge distillation SLMs (BGE-M3 as teacher model)																		
BGE-M3-SCT-KD-XLMR-base 0.28B	99.43	75.32	76.69	88.92	54.44	84.20	6.05	78.66	86.60	83.37	**100.00**	77.10	32.52	91.74	71.26	67.50	49.84	71.98
BGE-M3-SCT-KD-WangchanBERTa 0.11B	99.78	76.26	83.25	81.53	51.47	82.51	5.54	79.53	85.97	84.65	97.43	76.14	33.32	94.54	69.83	69.06	49.84	71.80
BGE-M3-SCT-KD-PhayaThaiBERT 0.28B	99.83	75.53	90.14	81.38	46.42	82.98	5.74	80.55	87.79	85.01	**100.00**	81.54	30.40	89.19	70.70	65.69	48.89	71.87
BGE-M3-ConGen-XLMR-base 0.28B	99.75	77.05	85.67	85.23	58.33	84.15	5.91	79.18	86.51	85.54	97.43	86.75	30.42	97.29	77.20	67.00	47.67	73.59
BGE-M3-Con-Gen-Wangchan-BERTa 0.11B	99.75	77.05	85.67	85.23	57.79	84.15	5.91	79.18	85.54	85.54	97.43	86.75	30.42	97.29	72.20	68.68	49.25	73.46
BGE-M3-Con-Gen-PhayaThaiBERT 0.28B	99.80	77.69	89.10	92.59	58.37	84.48	5.79	79.73	87.56	86.85	97.43	85.38	31.40	97.31	75.98	68.92	49.22	74.57
LLMs																		
E5 Mistral 7B Instruct	99.75	76.86	81.80	**96.31**	58.13	86.15	5.87	78.00	86.60	84.72	97.44	86.57	29.93	89.07	70.29	68.33	**54.91**	73.57
gte-Qwen2 7B Instruct	99.60	77.05	**90.66**	85.19	55.52	85.16	**6.17**	80.94	87.61	85.08	**100.00**	85.64	**48.04**	88.89	68.00	66.62	50.52	74.16
GritLM 7B	99.88	75.75	75.01	73.93	56.46	85.70	5.74	78.98	86.45	81.50	92.31	77.37	30.95	83.33	62.79	64.11	49.93	69.42
Llama3 8B	99.90	75.71	76.37	74.15	59.28	85.95	5.55	80.08	87.94	82.69	**100.00**	86.25	31.61	89.07	68.13	64.03	52.30	71.71
Llama3 8B Instruct	**99.93**	75.71	74.72	77.84	**59.30**	85.44	5.88	80.35	87.57	83.34	**100.00**	86.78	31.41	86.24	69.80	65.04	52.22	71.86
Llama3.1 8B	**99.93**	75.59	76.48	77.84	58.56	86.29	5.75	79.93	87.41	81.55	97.43	86.25	30.40	94.54	69.49	63.54	52.44	71.97
Llama3.1 8B Instruct	99.90	75.66	74.84	81.48	58.67	86.36	5.47	80.92	87.64	81.76	**100.00**	84.72	29.66	91.86	70.37	63.77	51.45	72.03
Typhoon 8B Instruct	99.90	76.25	76.60	81.48	58.67	**87.07**	5.71	80.75	87.73	81.64	97.43	83.28	31.29	94.54	68.45	65.86	52.20	72.29

Table 4. *Cont.*

Dataset (→) Metric (→)	CBLGBT F1.	DEP F1.	EMO F1.	GAMY F1.	GRENTH F1.	K500 F1.	LS F1.	MINT F1.	MSCN F1.	MLS F1.	RSHOP F1.	SIB200 F1.	SEATR F1.	TCAS61 F1.	40TCS F1.	WSENT F1.	WREV F1.	Avg F1.
Off-the-shelf embedding APIs																		
Cohere-embed-multilingual-v2.0	96.00	77.97	89.01	81.38	56.60	83.16	5.46	79.03	86.99	84.99	**100.00**	82.80	35.46	97.29	64.56	67.24	48.08	72.71
Cohere-embed-multilingual-v3.0	99.65	77.66	88.99	**96.31**	58.73	86.69	5.98	**83.04**	**89.64**	86.94	97.03	91.58	35.06	**100.00**	77.14	68.66	52.48	**76.21**
Openai-text-embedding-3-large	99.90	76.92	81.60	77.78	56.86	85.23	6.07	81.42	88.49	85.34	**100.00**	88.11	33.61	97.29	71.08	68.75	50.76	73.48

Note: **Bold** and Underline indicate the best and second performers.

Table 5. The Recall@1 and MRR@10 of each model on seven Thai retrieval datasets.

Dataset (→)	IAppWiki		MLDR-th		MIRACL-th		ThaiWikiQA		TyDiQA-th		ThaiCCLRAG		XQuAD-th		Avg.	
Metric (→)	**R@1**	**MRR@10**	**R@1**	**MRR@10**	**R@1**	**MRR@10**	**R@1**	**MRR@10**	**R@1**	**MRR@10**	**R@1**	**MRR@10**	**R@1**	**MRR@10**	**R@1**	**MRR@10**
SLMs																
XLMR-base 0.28B	6.90	9.77	1.00	1.62	2.32	3.02	2.79	3.68	3.41	5.06	6.92	9.52	4.12	8.64	3.92	5.90
XLMR-large 0.56B	11.37	17.05	1.50	3.01	4.37	6.27	6.19	8.35	4.59	7.10	11.78	15.34	14.37	22.03	7.74	11.31
WangchanBERTA 0.11B	18.00	23.89	3.50	6.38	6.14	10.07	11.96	16.14	12.58	19.74	7.61	10.27	21.76	28.66	11.65	16.45
DistilUSE-multilingual 0.14B	50.47	59.07	3.50	6.74	17.74	27.78	38.16	44.54	32.50	42.20	19.96	26.30	49.16	58.19	30.21	37.83
PhayaThaiBERT 0.28B	56.56	66.28	7.50	10.32	26.19	38.00	26.21	34.39	47.44	58.37	33.05	42.07	61.09	72.56	36.86	46.00
MPNet-multilingual 0.28B	65.76	72.19	10.50	14.84	38.20	49.65	54.02	60.93	54.39	63.12	41.92	52.02	71.26	79.63	48.01	56.05
BGE-M3 (dense only) 0.57B	**93.64**	**94.89**	25.00	30.73	**79.54**	**86.62**	**92.83**	**94.88**	88.99	93.36	**67.22**	**76.03**	90.42	**94.27**	**76.81**	**81.54**
Self-supervised SLMs																
SimCSE-XLMR-base 0.28	70.50	75.58	2.00	4.77	34.92	47.53	59.95	65.53	58.06	64.72	14.64	21.09	41.34	50.25	40.20	47.07
SimCSE-WangchanBERTa 0.11B	71.85	77.25	4.00	6.61	19.92	32.02	66.26	71.39	47.58	57.57	19.48	26.83	51.68	63.55	40.11	47.89
SimCSE-Phaya ThaiBERT 0.28B	81.46	84.70	5.00	9.46	43.11	57.19	73.96	78.39	71.17	78.07	35.05	43.95	54.54	62.90	52.04	59.24
SCT-XLMR-base 0.28B	59.68	66.50	4.50	10.12	28.51	40.84	47.28	54.79	49.28	58.62	19.50	27.31	55.29	65.23	37.72	46.20
SCT-WangchanBERTa 0.11B	67.39	74.81	7.50	11.72	34.52	47.26	53.46	61.50	56.23	66.60	34.68	45.28	67.56	77.62	45.91	54.97
SCT-PhayaThaiBERT 0.28B	73.34	79.09	9.50	14.52	37.52	51.06	63.88	71.09	63.17	71.53	42.29	53.13	67.31	76.00	51.00	59.49
Knowledge distillation SLMs (MPNet-multilingual as teacher model)																
MPNet-SCT-KD-XLMR-base 0.28B	66.44	74.29	9.00	14.74	40.38	51.68	53.08	60.56	56.36	65.18	35.83	46.04	68.91	78.19	47.14	55.81
MPNet-SCT-KD-WangchanBERTa 0.11B	65.90	73.48	7.00	11.59	36.97	48.94	50.31	57.74	54.26	64.01	31.85	41.90	63.78	74.18	44.30	53.12
MPNet-SCT-KD-PhayaThaiBERT 0.28B	69.96	76.77	8.50	14.24	45.16	56.57	55.57	62.99	58.06	67.32	34.84	45.22	71.01	79.94	49.01	57.58
MPNet-ConGen-XLMR-base 0.28B	72.67	78.67	8.50	15.24	43.11	55.51	57.77	64.86	60.29	68.56	44.78	54.75	71.76	80.01	51.29	59.66
MPNet-ConGen-Wangchan BERTA 0.11B	70.09	76.77	10.00	15.21	41.75	53.72	54.93	62.24	57.93	66.81	41.97	52.13	74.45	82.46	50.16	58.48
MPNet-ConGen-PhayaThaiBERT 0.28B	71.72	78.40	9.50	15.72	44.34	55.77	57.68	64.66	59.63	67.89	44.24	54.54	72.10	80.45	51.32	59.63

Table 5. *Cont.*

Dataset (→)	IAppWiki		MLDR-th		MIRACL-th		ThaiWikiQA		TyDiQA-th		ThaiCCLRAG		XQuAD-th		Avg.	
Metric (→)	**R@1**	**MRR@10**	**R@1**	**MRR@10**	**R@1**	**MRR@10**	**R@1**	**MRR@10**	**R@1**	**MRR@10**	**R@1**	**MRR@10**	**R@1**	**MRR@10**	**R@1**	**MRR@10**
Knowledge distillation SLMs (BGE-M3 as teacher model)																
BGE-M3-SCT-KD-XLMR-base 0.28B	42.63	52.51	5.00	7.75	21.69	30.99	38.54	45.84	34.08	43.49	25.27	33.24	48.69	57.42	30.84	38.75
BGE-M3-SCT-KD-WangchanBERTa 0.11B	54.67	63.06	5.00	6.91	29.33	37.70	52.87	59.63	40.76	50.27	25.70	33.86	55.55	64.84	37.70	45.18
BGE-M3-SCT-KD-PhayaThaiBERT 0.28B	56.97	63.76	3.00	7.22	33.42	43.93	54.24	60.90	46.53	54.77	30.54	38.78	57.65	66.85	40.34	48.03
BGE-M3-ConGen-XLMR-base 0.28B	91.75	93.54	15.00	20.06	68.35	78.25	85.73	88.98	83.75	88.59	57.17	67.12	82.02	87.83	69.11	74.91
BGE-M3-ConGen-WangchanBERTa 0.11B	91.75	93.54	15.00	20.06	64.26	74.63	85.73	88.98	77.59	83.54	57.17	67.12	82.18	88.28	67.67	73.74
BGE-M3-ConGen-PhayaThaiBERT 0.28B	92.02	93.20	13.50	19.12	70.40	79.33	85.42	88.73	83.36	88.29	60.03	69.59	85.80	90.48	70.08	75.53
LLMs																
E5 Mistral 7B Instruct	92.02	93.91	18.50	27.06	69.85	79.16	82.73	86.41	87.29	91.61	63.16	72.34	83.61	89.63	71.02	77.16
gte-Qwen2 7B Instruct	33.83	42.32	4.50	7.45	21.15	30.71	11.50	14.93	36.04	44.18	6.79	10.19	30.08	40.04	20.56	27.12
GritLM 7B	18.81	27.43	0.50	1.84	9.41	14.73	14.73	14.11	14.55	20.28	16.56	21.08	24.37	33.35	13.57	18.97
Llama3 8B	36.94	46.66	4.50	7.54	20.74	30.98	21.35	27.36	39.84	49.86	17.82	24.02	52.35	62.95	27.65	35.62
Llama3 8B Instruct	27.74	37.28	5.00	7.65	23.06	33.38	21.56	27.83	43.12	52.18	13.22	18.22	54.62	65.58	26.90	34.59
Llama3.1 8B	38.84	48.54	4.00	7.51	17.74	26.23	21.65	27.75	33.81	43.31	20.95	27.94	50.92	61.37	26.84	34.66
Llama3.1 8B Instruct	39.51	49.99	3.50	6.92	17.33	26.25	18.99	24.79	32.63	42.68	15.79	22.10	51.34	61.95	25.58	33.53
Typhoon 8B Instruct	34.51	43.15	5.50	8.54	22.10	35.06	22.81	28.79	48.10	58.00	15.76	22.06	53.45	64.88	28.89	37.21
Off-the-shelf embedding APIs																
Cohere-embed-multilingual-v2.0	91.20	92.77	11.00	15.94	66.98	77.58	86.32	89.24	85.45	90.33	54.21	62.69	82.52	87.78	68.24	73.76
Cohere-embed-multilingual-v3.0	92.29	94.03	6.00	8.74	78.04	85.71	89.79	92.62	**91.09**	**94.36**	51.40	62.33	**90.76**	94.23	71.34	76.00
Openai-text-embedding-3-large	91.07	93.21	**29.00**	**34.09**	70.94	79.87	72.89	77.71	83.75	89.14	48.22	58.62	74.54	82.61	67.20	73.61

Note: **Bold** and Underline indicate the best and second performers.

The key findings from these results are summarized as follows:

- **Overall Performance.** BGE-M3 achieves the highest overall score (78.75), while DistillUSE-multilingual offers the fastest inference (579 Queries/Second). Cohere-embed-multilingual-v3.0 ranks second overall (75.65) and stands out as the most effective off-the-shelf API model.
- **Task-Specific Strengths.** No model dominates all tasks: BGE-M3 leads in retrieval, Cohere-embed-multilingual-v3.0 in text classification, and MPNet-multilingual in STS and pair classification.
- **LLMs vs. SLMs.** Contrary to the prevailing assumption that larger models yield superior embeddings (as seen in English benchmarks like MTEB), our results show the opposite trend for Thai. Small-to-medium models consistently outperform LLM-based embeddings: BGE-M3 surpasses the strongest LLM (E5 Mistral 7B Instruct) by 4.85 points while being almost 4 times faster. This result underscores both the limited Thai coverage in large model training data and the need for more inclusive multilingual representation.
- **Unsupervised Fine-tuning Effectiveness.** All evaluated methods (SimCSE, SCT, SCT-KD, ConGen) consistently improve model performance, with ConGen yielding the largest gains. Remarkably, ConGen boosts PhayaThaiBERT to perform nearly on par

with BGE-M3 while being only half its size. These results demonstrate that small-to-medium models, when enhanced through unsupervised fine-tuning, can serve as both effective and efficient solutions.

- **Text Classification Performance.** Performance is fragmented. Cohere-embed-multilingual-v3.0 leads on four datasets (GAMY, MINT, MSCN, and TCAS61) and BGE-M3 on three (DEP, SIB200, and 40TCS), but no model dominates overall. Importantly, all models perform poorly on factual detection (LS dataset), showing that pretrained embeddings struggle to distinguish real from fake information.
- **Retrieval Performance.** BGE-M3 leads most datasets (IAppWiki, MIRACL-th, ThaiWikiQA, ThaiCCLRAG, and XQUAD-th), but retrieval remains difficult in specific settings. Long-document retrieval (MLDR-th) is a major weakness, with all models underperforming except Openai-text-embedding-3-large. All models yield moderate results on Legal-domain retrieval (ThaiCCLRAG), pointing to the need for domain-specific fine-tuning.
- **STS and Pair Classification Performance.** MPNet-multilingual excels on these proxy tasks, but its strength does not carry over to downstream applications such as retrieval. This misalignment reflects a broader limitation of relying on STS and pair classification as predictors of real-world performance. This finding aligns with previous studies in the English language [47,48].

In summary, the experiments demonstrate that small-to-medium models outperform LLM-based embeddings, unsupervised fine-tuning provides substantial efficiency and accuracy gains, and key challenges remain in Thai factual detection and long-document retrieval.

6. Conclusion and Future Works

In conclusion, we present the Thai Sentence Embedding Benchmark to bridge the gap between English and Thai sentence embedding evaluations. Our benchmark encompasses four distinct tasks to evaluate the performance and efficiency of embedding models: (i) semantic textual similarity (STS), (ii) text classification, (iii) pairwise text classification, and (iv) retrieval. We conducted extensive experiments with 36 embedding models, including the latest LLM-based sentence embeddings, and evaluated the effectiveness of four state-of-the-art unsupervised fine-tuning methods. The experimental results reveal that there is no single model that dominates others in all tasks within our benchmark. Interestingly, all newly developed LLM-based embedding models (e.g., E5 Mistral 7B Instruct) perform poorly in Thai compared to smaller models like BGE-M3, which is approximately 12 times smaller. This discrepancy likely stems from the distribution of Thai language examples in their training dataset. In future research, we aim to expand the scope of our benchmark to include other languages in Southeast Asia, such as Lao, Vietnamese, and Burmese, which are underrepresented in existing sentence embedding benchmarks. To achieve this, we plan to integrate data from open-source, multilingual projects like SEACrowd [22] into our benchmark, thereby enriching the diversity and inclusivity of our dataset. Through these efforts, we aim to contribute to the continuous improvement of NLP research in the Thai language and beyond.

Author Contributions

P.L., P.T. and W.P. contributed equally to this work. Conceptualization, P.L.; methodology, P.L. and P.T.; software, P.L., P.T. and W.P.; validation, P.L. and P.T.; formal analysis, P.L. and P.T.; experimentation, P.T., W.P., S.N., S.S.L., and C.U.; investigation, P.L. and P.T.; resources, S.N.; data curation, P.L. and W.P.; writing—original draft preparation, P.L.; writing—review and editing, P.L., P.T., C.U. and S.N.; visualization, P.T.; supervision, S.N.; project administration, P.T. All authors have read and agreed to the published version of the manuscript.

Funding

This work received no external funding.

Institutional Review Board Statement

The study does not require ethical approval.

Informed Consent Statement

Not applicable.

Data Availability Statement

Source code for benchmarking is publicly available at https://github.com/mrpeerat/Thai-Sentence-Vector-Benchmark for reproducibility.

All models and datasets used in this work are publicly available on Hugging Face. This section provides links to the corresponding repositories.

Models

SLMS

- XLMR-base: https://huggingface.co/FacebookAI/xlm-roberta-base
- XLMR-large: https://huggingface.co/FacebookAI/xlm-roberta-large
- WangchanBERTa: https://huggingface.co/airesearch/wangchanberta-base-att-spm-uncased
- Phaya ThaiBERT: https://huggingface.co/clicknext/phayathaibert
- MPNet-multilingual: https://huggingface.co/sentence-transformers /paraphrase-multilingual-mpnet-base-v2
  DistilUSE-multilingual: https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2
- BGE-M3: https://huggingface.co/BAAI/bge-m3
- SimCSE-XLMR-base: https://huggingface.co/kornwtp/simcse-model-XLMR
- SimCSE-WangchanBERTa: https://huggingface.co/kornwtp/simcse-model-wangchanberta
- SimCSE-PhayaThaiBERT: https://huggingface.co/kornwtp/simcse-model-phayathaibert
- SCT-XLMR-base: https://huggingface.co/kornwtp/SCT-model-XLMR
- SCT-WangchanBERTa: https://huggingface.co/kornwtp/SCT-model-wangchanberta
- SCT-PhayaThaiBERT: https://huggingface.co/kornwtp/SCT-model-phayathaibert
- MPNet-ConGen-XLMR-base: https://huggingface.co/kornwtp/ConGen-model-XLMR
- MPNet-ConGen-WangchanBERTa: https://huggingface.co/kornwtp/ConGen-model-wangchanberta
- MPNet-ConGen-PhayaThaiBERT: https://huggingface.co/kornwtp/ConGen-model-phayathaibert
- MPNet-SCT-KD-XLMR-base: https://huggingface.co/kornwtp/SCT-KD-model-XLMR
- MPNet-SCT-KD-WangchanBERTa: https://huggingface.co/kornwtp/SCT-KD-model-phayathaibert
- MPNet-SCT-KD-PhayaThaiBERT: https://huggingface.co/kornwtp/SCT-KD-model-phayathaibert
- BGE-M3-ConGen-XLMR-base: https://huggingface.co/kornwtp/ConGen-BGE_M3-model-XLMR
- BGE-M3-ConGen-WangchanBERTa: https://huggingface.co/kornwtp/ConGen-BGE_M3-model-wangchanberta
- BGE-M3-ConGen-PhayaThaiBERT: https://huggingface.co/kornwtp/ConGen-BGE_M3-model-phayathaibert
- BGE-M3-SCT-KD-XLMR-base: https://huggingface.co/kornwtp/SCT-KD-BGE-M3-model-XLMR
- BGE-M3-SCT-KD-WangchanBERTa: https://huggingface.co/kornwtp/SCT-KD-BGE-M3-model-wangchanberta
- BGE-M3-SCT-KD-PhayaThaiBERT: https://huggingface.co/kornwtp/SCT-KD-BGE-M3-model-phayathaibert

LLMs

- E5 Mistral 7B Instruct: https://huggingface.co/intfloat/e5-mistral-7b-instruct
- gte-Qwen2 7B Instruct: https://huggingface.co/Alibaba-NLP/gte-Qwen2-7B- instruct
- GritLM 7B: https://huggingface.co/GritLM/GritLM-7B
- Llama3 8B: https://huggingface.co/meta-llama/Meta-Llama-3-8B
- Llama3 8B Instruct: https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
- Llama3.1 8B: https://huggingface.co/meta-llama/Meta-Llama-3.1-8B
- Llama3.1 8B Instruct: https://huggingface.co/meta-llama/Meta-Llama-3.1- 8B-Instruct
- Typhoon 8B Instruct: https://huggingface.co/scb10x/llama-3-typhoon-v1.5- 8b-instruct

Datasets

Semantic Textual Similarity (STS)

- STS-B: https://huggingface.co/kornwtp/stsbenchmark-tha-sts

Text Classification

- CyberbullyingLGBT: https://huggingface.co/kornwtp/cyberbullying-lgbt-tha-classification
- Depression: https://huggingface.co/kornwtp/depression-tha-classification
- Emoji: https://huggingface.co/kornwtp/emoji-tha-classification
- GeneralAmy: https://huggingface.co/kornwtp/general-amy-tha-classification
- Generated Review: https://huggingface.co/airesearch/generated_reviews_enth
- Krathu500: https://huggingface.co/kornwtp/krathu500-tha-classification
- LimeSoda: https://huggingface.co/kornwtp/limesoda-tha-classification
- MassiveIntent: https://huggingface.co/kornwtp/massive-intent-tha-classification
- MassiveScenario: https://huggingface.co/kornwtp/massive-scenario-tha-classification
- MultiLingualSentiment: https://huggingface.co/kornwtp/multilingual-sentiment-tha-classification
- ReviewShopping: https://huggingface.co/kornwtp/review-shopping-tha-classification
- SIB200: https://huggingface.co/kornwtp/sib200-tha-clustering
- SEATranslationeseResampled: https://huggingface.co/kornwtp/sea-translationese-resampled-tha-classification
- TCAS61: https://huggingface.co/kornwtp/tcas61-tha-classification
- The40ThaiChildrenStories: https://huggingface.co/kornwtp/the40thai-children-stories-tha-classification
- Wisesight: https://huggingface.co/pythainlp/wisesight_sentiment
- Wongnai: https://huggingface.co/datasets/Wongnai/wongnai_reviews

Pair Classification

- XNLI: https://huggingface.co/datasets/kornwtp/xnli-tha-pairclassification

Retrieval

- IAppWiki: https://huggingface.co/datasets/kornwtp/iapp-wikiqa-tha-qaretrieval
- MLDR: https://huggingface.co/kornwtp/mldr-tha-qaretrieval
- MIRACL: https://huggingface.co/miracl/miracl
- ThaiWikiQA: https://huggingface.co/kornwtp/thai-wikiqa-tha-qaretrieval
- TyDiQA: https://huggingface.co/chompk/tydiqa-goldp-th
- WangchanXLegalThaiCCLRAG: https://huggingface.co/kornwtp/wangchanx-legalrag-tha-qaretrieval
- XQuAD: https://huggingface.co/google/xquad

Acknowledgments

We would like to express our gratitude to our colleagues and collaborators for their invaluable feedback and discussions throughout the development of this work, as well as for their assistance with journal formatting. We are especially grateful to all expert translators who contributed their time, without whom this study would not have been possible.

Conflicts of Interest

The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

[1] Devlin, J., Chang, M.-W., Lee, K., et al., 2019. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.

[2] Liu, Y., Lin, W., Shi, Y., et al., 2021. RoBERTa: A Robustly Optimized BERT Pre-training Approach with Post-Training. In Proceedings of the 20th China National Conference on Computational Linguistics, Hohhot, China, 13–15 August 2021; pp. 1218–1227. Available from: http://www.cips-cl.org/static/antholo-

gy/CCL-2021/CCL-21-108.pdf

[3] Gao, T., Yao, X., Chen, D., 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 6894–6910.

[4] Wang, L., Yang, N., Huang, X., et al., 2024. Improving Text Embeddings with Large Language Models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 11–16 August 2024; pp. 11897–11916.

[5] Muennighoff, N., Su, H., Wang, L., et al., 2024. Generative Representational Instruction Tuning. arXiv preprint. arXiv:2402.09906v3. DOI: https://doi.org/10.48550/arXiv.2402.09906

[6] Li, Z., Zhang, X., Zhang, Y., et al., 2023. Towards General Text Embeddings with Multi-Stage Contrastive Learning. arXiv preprint. arXiv:2308.03281v1. DOI: https://doi.org/10.48550/arXiv.2308.03281

[7] Muennighoff, N., Tazi, N., Magne, L., et al., 2023. MTEB: Massive Text Embedding Benchmark. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Dubrovnik, Croatia, 2–6 May 2023; pp. 2014–2037.

[8] Conneau, A., Kiela, D., 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, 7–12 May 2018; pp. 1699–1704.

[9] Thakur, N., Reimers, N., Rücklé, A., et al., 2021. BEIR: A Heterogenous Benchmark for Zero-Shot Evaluation of Information Retrieval Models. arXiv preprint. arXiv:2104.08663. DOI: https://doi.org/10.48550/arXiv.2104.08663

[10] Charin, P., Phasathorn, S., 2020. PyThaiNLP Classification Benchmarks. Available from: https://github.com/PyThaiNLP/classification-benchmarks (cited 20 January 2025).

[11] Cer, D., Diab, M., Agirre, E., et al., 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-Lingual Focused Evaluation. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 1–14.

[12] Wongso, W., Ananto, J., David, S.S., et al., 2024. Indonesian Sentence Embeddings. Available from: https://github.com/LazarusNLP/indonesian-sentence-embeddings (cited 20 January 2025).

[13] Wang, Y.-S., Wu, A., Neubig, G., 2022. English Contrastive Learning Can Learn Universal Cross-Lingual Sentence Embeddings. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 9122–9133.

[14] Chen, J., Xiao, S., Zhang, P., et al., 2024. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, Bangkok, Thailand, 11–16 August 2024; pp. 2318–2335.

[15] Hämäläinen, M., Patpong, P., Alnajjar, K., et al., 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), online, 11 November 2021; pp. 20–25.

[16] Phatthiyaphaibun, W., Chaovavanich, K., Polpanumas, C., et al., 2023. PyThaiNLP: Thai Natural Language Processing in Python. In Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023), Singapore, 6 December 2023; pp. 25–36.

[17] Lowphansirikul, L., Polpanumas, C., Rutherford, A.T., et al., 2022. A Large English–Thai Parallel Corpus from the Web and Machine-Generated Text. Language Resources and Evaluation. 56(2), 477–499. DOI: https://doi.org/10.1007/s10579-021-09536-6

[18] Payoungkhamdee, P., Porkaew, P., Sinthunyathum, A., et al., 2021. LimeSoda: Dataset for Fake News Detection in Healthcare Domain. In Proceedings of the 2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Ayutthaya, Thailand, 22–23 December 2021; pp. 1–6. DOI: https://doi.org/10.1109/iSAI-NLP54397.2021.9678187

[19] FitzGerald, J., Hench, C., Peris, C., et al., 2023. MASSIVE: A 1M-Example Multilingual Natural Language Understanding Dataset with 51 Typologically-Diverse Languages. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada, 9–14 July 2023; pp. 4277–4302.

[20] Mollanorozy, S., Tanti, M., Nissim, M., 2023. Cross-Lingual Transfer Learning with Persian. In Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP, Dubrovnik, Croatia, 2–6 May 2023; pp. 89–95.

[21] Adelani, D.I., Liu, H., Shen, X., et al., 2024. SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects.

In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), St. Julian's, Malta, 17–22 March 2024; pp. 226–245.

[22] Lovenia, H., Mahendra, R., Maulana Akbar, S.M., et al., 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, Miami, FL, USA, 12–16 November 2024; pp. 5155–5203.

[23] Pasupa, K., Netisopakul, P., Lertsuksakda, R., 2016. Sentiment Analysis of Thai Children Stories. Artificial Life and Robotics. 21(3), 357–364. DOI: https://doi.org/10.1007/s10015-016-0283-8

[24] Suriyawongkul, A., Chuangsuwanich, E., Chormai, P., et al., 2019. Wisesight Sentiment. Available from: https://github.com/PyThaiNLP/wisesight-sentiment (cited 20 January 2025).

[25] Liu, F., Jiao, Y., Massiah, J., et al., 2022. Trans-Encoder: Unsupervised Sentence-Pair Modelling Through Self- and Mutual-Distillations. arXiv preprint. arXiv:2109.13059. DOI: https://doi.org/10.48550/arXiv.2109.13059

[26] Conneau, A., Rinott, R., Lample, G., et al., 2018. XNLI: Evaluating Cross-Lingual Sentence Representations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2475–2485.

[27] Yang, Y., Cer, D., Ahmad, A., et al., 2020. Multilingual Universal Sentence Encoder for Semantic Retrieval. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Stroudsburg, PA, USA, 5–10 July 2020; pp. 87–94.

[28] Karpukhin, V., Oğuz, B., Min, S., et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), online, 16–20 November 2020; pp. 6769–6781.

[29] Asai, A., Yu, X., Kasai, J., et al., 2021. One Question Answering Model for Many Languages with Cross-Lingual Dense Passage Retrieval. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), online, 6–14 December 2021.

[30] Limkonchotiwat, P., Ponwitayarat, W., Udomcharoenchaikit, C., et al., 2022. CL-ReLKT: Cross-Lingual Language Knowledge Transfer for Multilingual Retrieval Question Answering. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, USA, 10–15 July 2022; pp. 2141–2155.

[31] Viriyayudhakorn, K., Charin, P., 2021. iapp wiki_qa_squad. Available from: https://github.com/iapp-technology/iapp-wiki-qa-dataset (cited 20 January 2025).

[32] Zhang, X., Thakur, N., Ogundepo, O., et al., 2023. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. Transactions of the Association for Computational Linguistics. 11, 1114–1131. DOI: https://doi.org/10.1162/tacl_a_00595

[33] Trakultaweekoon, K., Thaiprayoon, S., Palingoon, P., et al., 2019. The First Wikipedia Questions and Factoid Answers Corpus in the Thai Language. In Proceedings of the 2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Chiang Mai, Thailand, 30 October–1 November 2019; pp. 1–4. DOI: https://doi.org/10.1109/iSAI-NLP48611.2019.9045143

[34] Clark, J.H.C., Choi, E., Collins, M., et al., 2020. A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. Transactions of the Association for Computational Linguistics. 8, 454–470. DOI: https://doi.org/10.1162/tacl_a_00317

[35] Akarajaradwong, P., Pothavorn, P., Chaksangchaichot, C., et al., 2025. NitiBench: Benchmarking LLM Frameworks on Thai Legal Question Answering Capabilities. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP), Suzhou, China, 10–14 December 2025; pp. 34292–34315.

[36] Conneau, A., Khandelwal, K., Goyal, N., et al., 2020. Unsupervised Cross-Lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, online, 5–10 July 2020; pp. 8440–8451.

[37] Wenzek, G., Lachaux, M.-A., Conneau, A., et al., 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 13–15 May 2020; pp. 4003–4012.

[38] Lowphansirikul, L., Polpanumas, C., Jantrakulchai, N., et al., 2021. WangchanBERTa: Pretraining Transformer-Based Thai Language Models. arXiv preprint. arXiv:2101.09635. DOI: https://doi.org/10.48550/arXiv.2101.09635

[39] Sriwirote, P., Thapiang, J., Timtong, V., et al., 2023. PhayaThaiBERT: Enhancing a Pretrained Thai

Language Model with Unassimilated Loanwords. arXiv preprint. arXiv:2311.12475. DOI: https://doi.org/10.48550/arXiv.2311.12475

[40] Reimers, N., Gurevych, I., 2019. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, 3–7 November 2019; pp. 3982–3992.

[41] Reimers, N., Gurevych, I., 2020. Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), online, 16–20 November 2020; pp. 4512–4525.

[42] Grattafiori, A., Dubey, A., Jauhri, A., et al., 2024. The Llama 3 Herd of Models. arXiv preprint. arXiv:2407.21783. DOI: https://doi.org/10.48550/arXiv.2407.21783

[43] Rafailov, R., Sharma, A., Mitchell, M., et al., 2023. Direct Preference Optimization: Your Language Model Is Secretly a Reward Model. arXiv preprint. arXiv:2305.18290. DOI: https://doi.org/10.48550/arXiv.2305.18290

[44] Pipatanakul, K., Jirabovonvisut, P., Manakul, P., et al., 2023. Typhoon: Thai Large Language Models. arXiv preprint. arXiv:2312.13951. DOI: https://doi.org/10.48550/arXiv.2312.13951

[45] Limkonchotiwat, P., Ponwitayarat, W., Lowphansirikul, L., et al., 2023. An Efficient Self-Supervised Cross-View Training for Sentence Embedding. Transactions of the Association for Computational Linguistics. 11, 1572–1587. DOI: https://doi.org/10.1162/tacl_a_00620.

[46] Limkonchotiwat, P., Ponwitayarat, W., Lowphansirikul, L., et al., 2022. ConGen: Unsupervised Control and Generalization Distillation for Sentence Representation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 6467–6480.

[47] Neelakantan, A., Xu, T., Puri, R., et al., 2022. Text and Code Embeddings by Contrastive Pre-Training. arXiv preprint. arXiv:2201.10005. DOI: https://doi.org/10.48550/arXiv.2201.10005

[48] Wang, K., Reimers, N., Gurevych, I., 2021. TSDAE: Using Transformer-Based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 671–688.