LUSIFER: Language Universal Space Integration for Enhanced Multilingual Embeddings with Large Language Models

Hieu Man[1], **Nghia Trung Ngo**[1], **Viet Dac Lai**[2],
Ryan A. Rossi[2], **Franck Dernoncourt**[2], **Thien Huu Nguyen**[1]
[1]Dept. of Computer Science, University of Oregon, OR, USA
[2]Adobe Research, USA
{hieum,nghian,thienn}@uoregon.edu
{viet.lai,ryrossi,franck.dernoncourt}@adobe.com

Abstract

Recent advancements in large language models (LLMs) based embedding models have established new state-of-the-art benchmarks for text embedding tasks, particularly in dense vector-based retrieval. However, these models predominantly focus on English, leaving multilingual embedding capabilities largely unexplored. To address this limitation, we present LUSIFER, a novel zero-shot approach that adapts LLM-based embedding models for multilingual tasks without requiring multilingual supervision. LUSIFER's architecture combines a multilingual encoder, serving as a language-universal learner, with an LLM-based embedding model optimized for embedding-specific tasks. These components are seamlessly integrated through a minimal set of trainable parameters that act as a connector, effectively transferring the multilingual encoder's language understanding capabilities to the specialized embedding model. Additionally, to comprehensively evaluate multilingual embedding performance, we introduce a new benchmark encompassing 5 primary embedding tasks, 123 diverse datasets, and coverage across 14 languages. Extensive experimental results demonstrate that LUSIFER significantly enhances the multilingual performance across various embedding tasks, particularly for medium and low-resource languages, without requiring explicit multilingual training data.

1 Introduction

Text embeddings, which provide dense vector representations of textual content (Mikolov et al., 2013; Devlin et al., 2019), have become fundamental building blocks in modern natural language processing. These embeddings encode semantic information and serve as an important component for numerous downstream applications, ranging from information retrieval and document reranking to classification, clustering, and semantic textual similarity assessment. Recently, the significance of high-quality embeddings has been further amplified by their crucial role in retrieval-augmented generation (RAG) systems (Lewis et al., 2020b). RAG architectures enable large language models (LLMs) to dynamically access and integrate external or proprietary knowledge without the need for model parameter updates, substantially enhancing their adaptability and accuracy (Wang et al., 2023; Liu et al., 2024b; Gao et al., 2024).

The evolution of embedding models has witnessed remarkable advancements, progressing from static word embeddings (Robertson et al., 2009) through contextualized representations (Reimers and Gurevych, 2019; Gao et al., 2021b; Ni et al., 2021a) to state-of-the-art LLM-based embedding models (Wang et al., 2024b) that harness the sophisticated semantic understanding capabilities of large language models. These developments have substantially enhanced performance across various embedding tasks (Luo et al., 2024), achieving unprecedented accuracy in semantic similarity and retrieval applications. However, a critical limitation remains: the predominant focus on English in LLM-based embedding models has created a significant disparity in multilingual capabilities. This gap is especially pronounced in medium and low-resource languages, where English-centric models exhibit substantial performance degradation due to insufficient language-specific training data (Wang et al., 2020; Thakur et al., 2024). While recent advances in multilingual embedding models, particularly those leveraging multilingual pre-trained architectures, have demonstrated promising results in multilingual embedding tasks (Li et al., 2023; Wang et al., 2024c; Chen et al., 2024), their reliance on explicit multilingual supervision for embeddings constrains their applicability primarily to languages with abundant training resources, leaving the challenge of true language-agnostic representation largely unaddressed.

To address this challenge, we present LUSIFER,

a novel zero-shot approach that adapts English LLM-based embedding models for multilingual tasks without requiring explicit multilingual supervision. Drawing inspiration from recent advances in multimodal integration (Liu et al., 2024a; Lu et al., 2024), LUSIFER employs a unique architecture that bridges the gap between multilingual understanding and specialized embedding capabilities. At its core, LUSIFER leverages the robust multilingual representations from XLM-R (Conneau et al., 2020) and introduces a learnable connector mechanism to interface with English-optimized LLM embedding models. This approach enables LUSIFER to effectively transfer the multilingual understanding of XLM-R to the target LLM while inheriting advanced embedding capabilities of the LLM. In this way, LUCIFER can achieve effective multilingual representation capabilities without requiring explicit multilingual training data.

We conduct comprehensive evaluations of LUSIFER through extensive experiments across 123 diverse datasets spanning 14 languages, focusing on five fundamental embedding tasks: Classification, Clustering, Reranking, Retrieval, and Semantic Textual Similarity (STS). Our experimental results demonstrate that LUSIFER substantially enhances the performance of English-centric LLM-based embedding models, achieving average improvements of 3.19 points across all tasks, with particularly significant gains observed for medium and low-resource languages (up to 22.15 improvement). To validate LUSIFER's broader applicability and cross-lingual capabilities, we extend our evaluation to cross-lingual tasks using four comprehensive datasets that encompass over 100 languages, including several critically low-resource languages. LUSIFER significantly outperforms existing English-centric embedding models by 5.75 on average in cross-lingual scenarios. These results demonstrate the effectiveness of our approach in enhancing multilingual representation capabilities without explicit multilingual supervision.

The theoretical foundation for LUSIFER's effectiveness lies in its ability to create a language-agnostic universal space through the integration of a multilingual encoder (Pires et al., 2019; Libovický et al., 2020). We hypothesize that this universal space serves as a bridge between different languages, enabling the target language model to process semantic information independently of the input language. By mapping these language-neutral representations to the target model's input space, we conjecture that the target LLM can grasp the semantics of these representations, thereby improving the quality of output embeddings across multiple languages. This mechanism allows the model to become less dependent on the specific language of the input, enabling it to better capture semantic information for embedding tasks in languages it rarely encountered during pretraining. Our empirical analysis using t-SNE visualization supports this hypothesis.

2 Related Work

2.1 English-centric Embedding Models

Text embedding models have experienced significant advancement in recent years, driven by the evolution of pre-trained language models. Early successes with BERT-based architectures, as demonstrated in Sentence-BERT (Reimers and Gurevych, 2019), SimCSE (Gao et al., 2021b), and DPR (Karpukhin et al., 2020), established the foundation for modern embedding approaches. The field has since progressed to leverage LLMs, with recent works (Wang et al., 2024b; Muennighoff et al., 2024; BehnamGhader et al., 2024; Lee et al., 2024; Man et al., 2024) demonstrating substantial improvements in embedding quality and task performance through the enhanced representational capacity of LLMs (Luo et al., 2024). However, these advances primarily benefit high-resource language applications, as most state-of-the-art LLM-based embedding models are derived from English-centric foundation models (Jiang et al., 2023; Touvron et al., 2023) and trained predominantly on English or high-resource language datasets (Wang et al., 2024b). This bias has resulted in a significant performance gap between high-resource and low-resource languages, limiting the global applicability of these models. Our proposed method, LUSIFER, addresses this limitation by enabling effective multilingual representation without multilingual training data.

2.2 Zero-shot Multilingual Embedding

Multilingual Embedding has evolved through several distinct methodological approaches, each addressing the fundamental challenge of bridging language gaps in embedding tasks. Early successful approaches relied on translation models to enable multilingual understanding (Liu et al., 2020; Shi et al., 2021; Zhang and Misra, 2022). While effective, these methods introduced operational com-

plexity by requiring external translation systems, limiting their practical deployment and scalability.

The emergence of multilingual pre-trained language models, particularly XLM-R (Conneau et al., 2020), opened new possibilities for multilingual transfer. Recent works have demonstrated promising results by fine-tuning such models with contrastive learning objectives on multilingual data (Wang et al., 2024c; Chen et al., 2024; Sturua et al., 2024). However, these approaches face two key limitations: they require substantial multilingual training data, and moreover, they do not exploit the sophisticated semantic representations afforded by contemporary English-centric LLM architectures, which have demonstrated superior performance in capturing nuanced semantic relationships.

Recent advances in aligning multilingual and English-centric representations could offer a solution. By combining independently pre-trained representations, a paradigm that has shown remarkable success in multimodal alignment research (Alayrac et al., 2022; Liu et al., 2024a; Lu et al., 2024), these works bridge the gap between visual encoders and language models to enhance visual comprehension. As such, similar principles can be applied to align multilingual representations with LLM-based semantic spaces. While related efforts have explored aligning multiple LLMs for improved reasoning capabilities in multilingual settings (Bansal et al., 2024; Yoon et al., 2024), these approaches primarily target generation tasks and typically require large-scale alignment data. Our work extends these efforts by focusing on embedding tasks and leveraging a minimal set of parameters to align multilingual and English-centric representations, enabling enhanced multilingual representation capabilities without requirement for large-scale multilingual training data.

2.3 Multilingual Embedding Benchmarks

The evaluation landscape for multilingual embedding models has historically been fragmented across various benchmarks, each with significant limitations. While existing benchmarks have made valuable contributions, they often exhibit constrained scope: MINERS (Winata et al., 2024) provides evaluation across multiple languages but is limited to classification and STS tasks with only 11 datasets; XNLI (Conneau et al., 2018), XQuAD (Artetxe et al., 2020), and SIB-200 (Adelani et al., 2024) offer broad language coverage but focus exclusively on classification tasks; and MTEB (Muennighoff et al., 2023), despite its diverse task selection, primarily addresses high-resource languages. To address these limitations, we introduce a comprehensive evaluation framework that encompasses 5 fundamental embedding tasks—Classification, Clustering, Reranking, Retrieval, and STS—across an extensive collection of 123 datasets spanning 14 languages. This holistic approach enables systematic evaluation across both task and language dimensions, providing unprecedented insights into models' multilingual capabilities. Furthermore, our benchmark extends beyond traditional multilingual evaluation by incorporating cross-lingual tasks, featuring coverage of over 100 languages, including critically low-resource languages that have been historically underrepresented in existing benchmarks. This extensive coverage allows for a more nuanced understanding of embedding models' performance across the global linguistic landscape.

3 Methodology

Previous works demonstrate that representations of multilingual encoder models exhibit inherent language-agnostic properties, facilitating zero-shot multilingual transfer (Pires et al., 2019; Libovický et al., 2020). Building upon this foundation, we propose LUSIFER, an embedding framework that aligns a multilingual encoder model with a target English-centric LLM's representational space, enabling the target to encode semantics across multiple languages without extensive multilingual training. This section details our architectural design and two-stage training process for LUSIFER.

3.1 Model Architecture

The core development of LUSIFER lies in its novel approach to enabling multilingual encoding of target LLMs through efficient representation mapping. As illustrated in Figure 1, LUSIFER's architecture consists of three key components: (1) a multilingual encoder that functions as a language-universal learner, capturing semantic information for diverse languages, (2) a language-agnostic connector that serves as a minimal parametric bridge between representations, and (3) a target LLM optimized for embedding-specific tasks. The multilingual encoder processes input from various languages into a shared semantic space, while the connector, designed with minimal trainable parameters, aligns these universal representations

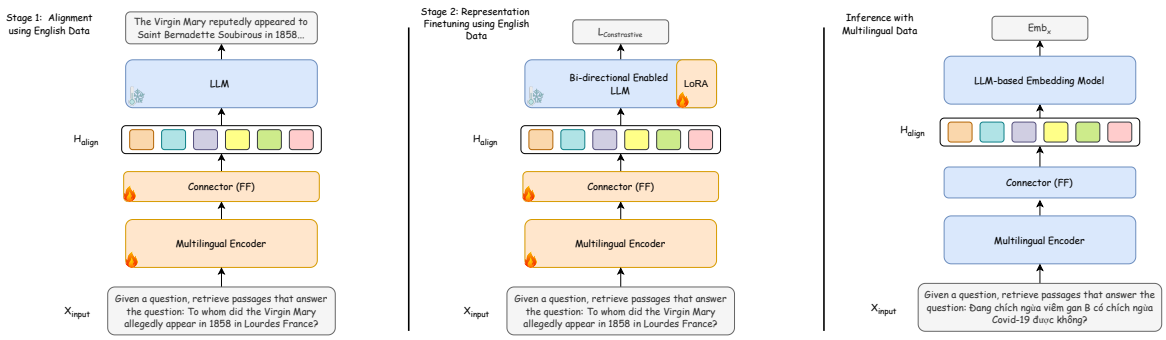


Figure 1: Overview of LUSIFER. **Left**: Align a multilingual encoder with the target English-centric LLM only using English data and a minimal set of trainable parameter. **Center**: End-to-end representation finetune through contrastive learning on English text-embedding tasks using LoRA. **Right**: During inference, LUSIFER successfully processes text-embedding tasks across multiple languages.

with the target LLM's native representational space. This alignment enables the target LLM embedding model to effectively leverage multilingual understanding without requiring extensive multilingual training data or architectural modifications.

Following successful approaches in multimodal alignment (Alayrac et al., 2022; Liu et al., 2024a; Lu et al., 2024), we implement the connector as a 2-layers feed-forward network, $\mathbf{FF}$, augmented with a single trainable token appended to the multilingual encoder's hidden states. Formally, given input tokens $\mathbf{X}_{input}$ (with necessary padding), the multilingual encoder's hidden states $\mathbf{H}_{enc}$ are transformed to align with the target LLM's representational space. The resulting aligned hidden states $\mathbf{H}_{align}$ maintain dimensionality compatibility with the target LLM's hidden states while extending the sequence length by one ($|\mathbf{X}_{input}| + 1$): $\mathbf{H}_{align} = [\mathbf{FF}(\mathbf{H}_{enc}); \mathbf{t}]$, where $\mathbf{FF}$ is the feed-forward network to align the multilingual encoder's hidden states with dimension $\mathbf{d}_e$ to the target LLM's hidden states with dimension $\mathbf{d}_t$, and $\mathbf{t} \in \mathbb{R}^{d_t}$ is the trainable token. Moreover, we employ a masking mechanism to mask any original padding tokens in $\mathbf{H}_{enc}$ to prevent their influence on the target LLM's processing, ensuring the model focuses on meaningful tokens.

3.2 Training Pipeline

LUSIFER employs a two-stage training process to achieve optimal multilingual representation capabilities. Both stages only require training on English data, leveraging the multilingual encoder's inherent language-agnostic properties and embedding advantages of LLMs to facilitate zero-shot multilingual transfer.

Stage 1: Alignment Training. The initial training stage aligns the multilingual encoder's representations with the target LLM's embedding space. Specifically, we optimize the connector parameters θ_c and the multilingual encoder parameters θ_e while keeping the target LLM's parameters fixed, ensuring stable convergence. The training employs two complementary objectives: (1) A masked reconstruction task where we randomly mask $k\%$ of input tokens such that $\mathbf{X}_{input} = \texttt{mask}(\mathbf{X}, k)$, training the model to recover the original sequence $\mathbf{X}_{lm} = \mathbf{X}$. (2) An autoregressive completion task that focuses on next-token prediction, where the model learns to generate the target sequence $\mathbf{X}_{lm}$ conditioned on the input context $\mathbf{X}_{input}$. The training objective for both tasks is formulated as language modeling objective to generate the target sequence $\mathbf{X}_{lm}$ given the input sequence $\mathbf{X}_{input}$. This objective enables local token-level alignment through masked reconstruction task where the model learns to predict the masked tokens by leveraging the context. In addition, it exploits global semantic alignment through autoregressive completion task that encourages the model to capture semantic information of the input sequences to generate the target sequence. As such, our training strategy learns to align the multilingual encoder's representations with the target LLM's embedding space while preserving important semantic information of multilingual input sequences. Our training process is conducted using the standard cross-entropy loss function.

Stage 2: Representation Finetuning. The second stage improves text representations through a contrastive learning process, effectively teaching the model to distinguish between positive and

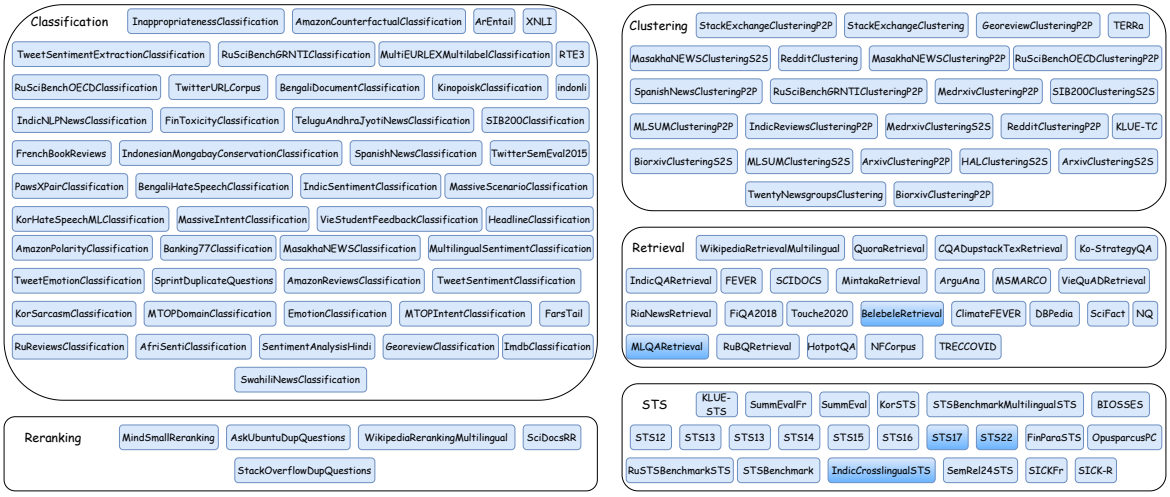


Figure 2: Overview of tasks and datasets in our benchmark. Crosslingual datasets are marked with a blue shade.

negative examples. Our approach leverages both in-batch negatives sampled from the current training batch and hard-negative examples specifically curated to enhance model training. Additionally, we incorporate bidirectional attention mechanisms within the target LLM, following recent advances in LLM's representation learning (Muennighoff et al., 2024; BehnamGhader et al., 2024; Lee et al., 2024; Man et al., 2024). This bidirectional context modeling significantly enhances the quality of learned representations by enabling the model to capture both forward and backward dependencies in the input sequence. During this stage, we finetune all components of LUSIFER, including the target LLM, the multilingual encoder, and the connector parameters, to optimize the model's representation quality for embedding-specific tasks. The goal of this stage is to improve the quality of text representations by leveraging the advanced embedding capabilities of the target LLM while maintaining the multilingual understanding provided by the multilingual encoder.

4 Experiment

In this section, we first introduce the benchmark datasets and evaluation metrics in Section 4.1. Then, we describe the experimental setup, including the model implementation, training data, and training details in Section 4.2. Afterward, we present the main results in Section 4.3, and analyze the effectiveness of LUSIFER's components in Section 4.6. Finally, we visualize the LUSIFER's representations in multilingual space to obtain insights into its lingual-agnostic capabilities in Section 4.7.

4.1 Benchmark

Figure 2 illustrates the tasks and datasets in our benchmark. Following (Muennighoff et al., 2023), our benchmark includes five fundamental embedding tasks, with the evaluation protocol for each task adapted from the respective original papers. The benchmark involves 123 diverse datasets, including 48 Classification datasets, 24 Clustering datasets, 24 Retrieval datasets, 22 Semantic Textual Similarity STS datasets, and 5 Reranking datasets. The main metrics for each task are as follows: Classification: Accuracy, Clustering: V-measure (Rosenberg and Hirschberg, 2007), Retrieval: nDCG@10, STS: Pearson correlation based on cosine similarity (Reimers et al., 2016), and Reranking: MAP. Following (Lai et al., 2023), our benchmark covers 14 languages including 5 high-resource languages: English (en), Spanish (es), Russian (ru), French (fr), Vietnamese (vi); 6 medium-resource languages: Persian (fa), Indonesian (id), Arabic (ar), Finnish (fi), Korean (ko), Hindi (hi); 3 low-resource languages: Bengali (bn), Telugu (te), Swahili (sw).

Additionally, we evaluate models on cross-lingual retrieval tasks where the models need to perform text embedding tasks with queries and documents in different languages. These tasks feature 5 datasets, including Belebele (Bandarkar et al., 2024), MLQA (Lewis et al., 2020a), STS17, STS22 (Agirre et al., 2016), and IndicCrosslingualSTS (Ramesh et al., 2022), covering over 100 languages, including critically low-resource languages.

Baselines	En	Es	Ru	Fr	Vi	Fa	Id	Ar	Fi	Ko	Hi	Bn	Te	Sw	Avg.
Jina-embeddings-v3* (Sturua et al., 2024)	59.84	61.23	62.88	58.94	66.74	78.35	58.51	64.71	73.57	64.96	64.19	61.54	68.96	49.20	63.83
mGTE-base* (Zhang et al., 2024)	60.40	59.65	61.02	56.20	65.81	73.46	56.55	61.97	68.96	61.22	60.81	58.24	63.58	52.57	61.46
BGE-M3* (Chen et al., 2024)	60.09	60.60	62.37	57.34	70.69	78.97	58.78	64.12	75.60	64.72	64.61	65.31	69.85	54.20	64.80
Multilingual-E5-large* (Wang et al., 2024d)	61.91	61.97	62.91	59.40	71.30	78.08	55.21	63.41	76.53	66.55	63.75	63.67	67.32	51.55	64.54
UDEVER-Bloom-7B* (Zhang et al., 2023)	55.83	56.39	59.73	54.38	64.32	68.70	48.97	55.02	67.60	58.54	55.96	55.13	61.00	47.41	57.78
SimCSE (Gao et al., 2021b)	51.92	51.81	24.90	46.95	31.18	37.12	39.27	29.46	41.64	26.23	25.17	21.54	26.71	38.36	35.16
Contriever (Izacard et al., 2022)	49.29	44.26	26.55	44.05	33.03	39.66	38.33	32.36	45.76	26.47	23.27	22.61	22.64	39.26	34.82
GTE-large (Li et al., 2023)	62.29	51.66	33.49	50.13	38.88	44.67	43.07	30.27	51.98	27.02	20.38	22.97	22.75	41.40	38.64
BGE-en-1.5 (Xiao et al., 2023)	63.27	51.65	32.79	50.84	38.50	49.73	43.28	30.81	51.16	31.11	25.28	26.34	23.02	41.96	39.98
E5-large (Wang et al., 2024a)	60.12	52.41	26.81	51.00	37.99	39.47	43.86	31.32	53.59	28.84	24.57	23.48	22.03	43.25	38.48
ST5-XXL (Ni et al., 2021c)	58.81	60.35	44.42	58.50	41.81	24.66	53.43	25.30	52.46	15.43	18.07	17.10	21.63	38.81	37.91
GTR-XXL (Ni et al., 2021b)	58.12	54.39	41.94	53.21	37.96	24.67	50.08	25.14	53.88	15.23	17.35	15.92	22.12	40.57	36.47
E5-Mistral (Wang et al., 2024b)	**66.64**	**61.84**	**61.30**	**59.65**	58.58	72.55	58.25	54.43	66.97	62.82	56.23	55.10	47.15	50.61	59.44
LUSIFER **(Ours)**	57.20	60.14	59.82	59.24	**67.69**	**76.17**	**59.70**	**55.60**	**72.83**	**65.23**	**62.37**	**58.43**	**69.30**	**53.12**	**62.63**

Table 1: Comparative analysis of model performance across multiple languages and tasks. The table presents average metrics for each model, with the highest score for each language emphasized in bold. * denotes the models trained on extensive multilingual data.

Baselines	MLQARetrieval	BelebeleRetrieval	STS17	STS22	IndicCrosslingual	Avg.
SimCSE (Gao et al., 2021b)	7.41	18.35	39.71	37.95	0.18	20.72
Contriever (Izacard et al., 2022)	9.75	22.94	34.55	41.72	0.03	21.80
GTE-large (Li et al., 2023)	16.99	31.82	37.57	53.79	1.59	28.35
BGE-en-1.5 (Xiao et al., 2023)	16.64	31.19	40.40	50.77	1.11	28.02
E5-large (Wang et al., 2024a)	17.04	31.12	37.90	54.31	1.83	28.44
ST5-XXL (Ni et al., 2021c)	20.82	41.68	56.19	59.02	1.76	35.89
GTR-XXL (Ni et al., 2021b)	20.19	38.02	50.83	60.11	2.74	34.38
E5-Mistral (Wang et al., 2024b)	31.54	54.75	**81.12**	**71.37**	21.92	52.14
LUSIFER **(Ours)**	**36.68**	**57.81**	81.09	70.49	**43.40**	**57.89**

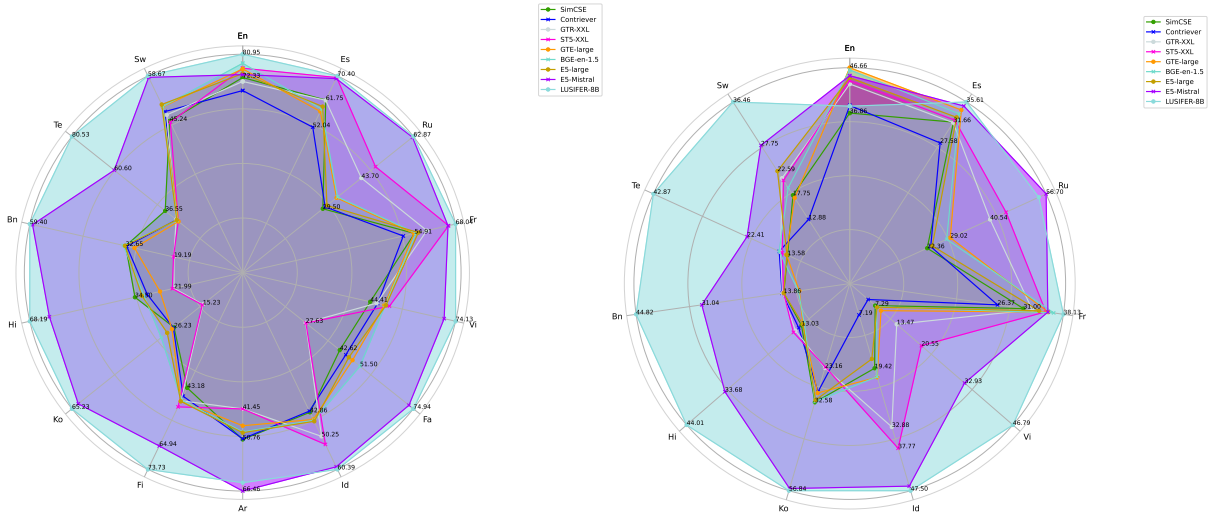
Table 2: Cross-lingual evaluation results. The table presents average metrics for each model over all languages of the datasets, with the highest score for each language emphasized in bold.

4.2 Experimental Setup

Implementation Details. LUSIFER encompasses three key components: a multilingual encoder, a connector, and a target LLM. We employ XLM-R-large (Conneau et al., 2020) as the multilingual encoder, Mistral-7B (Jiang et al., 2023) as the English-centric target LLM, and a 2-layer feed-forward network with one trainable token as the connector. To facilitate efficient training, we leverage the LoRA framework (Hu et al., 2022) for training of LUSIFER's components. Furthermore, we employ GradCache (Gao et al., 2021a), gradient checkpointing, mixed precision training, and FSDP (Zhao et al., 2023) to minimize GPU memory requirements. The LUSIFER architecture and its training code are built on top of the Hugging Face Transformers (Wolf et al., 2020) and Pytorch Lightning libraries (Falcon and team, 2024). We detail the training hyper-parameters for each stage in Table 4 of Appendix A.

Training Data. We only train LUSIFER on a diverse public English datasets. For alignment training, we use the combination of the English Wikipedia and questions-answering datasets. Specifically, we use subset of Wikitext-103 (Merity et al., 2017) and MSMARCO (Bajaj et al., 2018) for the masked reconstruction and autoregressive completion tasks, respectively. For representation finetuning, we adopt the retrieval datasets as follows: MS MARCO (Bajaj et al., 2018), NQ (Kwiatkowski et al., 2019), PAQ (Lewis et al., 2021), HotpotQA (Yang et al., 2018), SNLI (Bowman et al., 2015), SQuAD (Rajpurkar et al., 2016), ArguAna (Wachsmuth et al., 2018), FiQA (Maia et al., 2018) and FEVER (Thorne et al., 2018). To address the lack of hard negatives in these datasets, we leverage an encoder-based model (Wang et al., 2024a) to select the hard negatives on those datasets. Refer to Table 5 for the number of samples used in each dataset.

Baselines. We evaluate LUSIFER's performance across the five fundamental embedding tasks on the benchmark datasets. We make comparisons with a variety of baseline models for embedding tasks which only trained/finetuned on mainly English data. Baselines include the following categories: dense retrieval models with Small Language Model (SLM) backbone: SimCSE (Gao et al., 2021b), Contriever (Izacard et al., 2022), GTE-large (Li et al., 2023), BGE-en-1.5 (Xiao et al., 2023), E5-large (Wang et al., 2024a); and dense retrieval models with Large Language Model



(a) Classification tasks

(b) Clustering tasks

Figure 3: Performance comparison of LUSIFER and baseline models on Classification and Clustering tasks.

(LLM) backbone: GTR-XXL (Ni et al., 2021b), ST5-XXL (Ni et al., 2021c), E5-Mistral (Wang et al., 2024b). Moreover, we include the following state-of-the-art multilingual embedding models which are trained on extensive multilingual data for reference: Jina-embeddings-v3 (Sturua et al., 2024), mGTE-base (Zhang et al., 2024), BGE-M3 (Chen et al., 2024), Multilingual-E5-large (Wang et al., 2024d), and UDEVER-Bloom-7B (Zhang et al., 2023).

4.3 Main Results

Table 1 presents the main results of LUSIFER and baseline models on the benchmark datasets. LUSIFER achieves state-of-the-art performance in 10 out of 14 languages, with an average score of 62.63 across all languages, a 3.19 points improvement over the previous best-performing baseline, E5-Mistral (59.44) (Wang et al., 2024b). Note that E5-Mistral is essentially the Mistral model fine-tuned on extensive proprietary synthetic data and supplemented with some multilingual data for training. Our results demonstrate that LUSIFER significantly enhances the multilingual capabilities of English-centric embedding LLM by aligning it with a multilingual encoder, enabling effective multilingual representation without requiring explicit multilingual training data. The improvements are particularly pronounced for medium and low-resource languages, with Telugu (te) showing the largest gain of 22.15 points over E5-Mistral. This highlights LUSIFER's effectiveness in improving representation capabilities for traditionally under-represented languages. Additionally, LUSIFER significantly outperforms the embedding models with SLM backbones, such as E5-large (38.48) and BGE-en-1.5 (39.98) which are trained on English data only, thus further demonstrating the benefits of combining multilingual encoder and LLM's English-centric for text-embedding tasks in multilingual settings. Furthermore, even without explicit multilingual supervision, LUSIFER achieves competitive performance (62.63) compared to state-of-the-art multilingual models that require extensive multilingual training data, such as BGE-M3 (64.80) (Chen et al., 2024) and Multilingual-E5-large (64.54) (Wang et al., 2024d). These results further demonstrate the benefits of LUCIFER for multilingual representation learning while avoiding expensive multilingual data for text embeddings.

4.4 Cross-Lingual Evaluation

Table 2 presents the results of LUSIFER and baseline models on the cross-lingual tasks. LUSIFER achieves the highest average score of 57.89, outperforming the previous best-performing baseline, E5-Mistral (52.14), by 5.75 points. Notably, LUSIFER demonstrates significant improvements in low-resource languages, as evidenced by its performance on the IndicCrosslingual dataset, where it achieves a score of 43.40, substantially higher than the next best baseline, E5-Mistral (21.92). These results underscore LUSIFER's effectiveness in enhancing cross-lingual capabilities through efficient multilingual representation alignment, enabling the model to process text-embedding tasks across mul-

Baselines	En	Es	Ru	Fr	Vi	Fa	Id	Ar	Fi	Ko	Hi	Bn	Te	Sw	Avg.
LUSIFER **(Full)**	**57.20**	**60.14**	**59.82**	**59.24**	**67.69**	**76.17**	**59.70**	**55.60**	**72.83**	**65.23**	**62.37**	**58.43**	**69.30**	**53.12**	**62.63**
LUSIFER **(Connector Only)**	35.53	33.98	42.95	33.54	35.68	57.86	35.55	27.60	48.72	34.45	47.57	41.85	46.50	34.66	44.18
LUSIFER **(Frozen Multilingual Encoder)**	50.99	58.77	58.30	52.73	62.24	75.88	58.11	41.66	70.75	59.53	62.48	55.53	66.24	49.12	58.74
LUSIFER **(Alignment Only)**	43.32	38.94	45.12	36.75	41.96	64.60	38.38	33.07	52.78	38.08	53.06	47.84	48.34	40.03	44.45
LUSIFER **(Representation Finetuning Only)**	49.71	58.76	58.08	51.01	62.11	74.01	57.32	40.95	68.47	57.81	59.74	53.53	63.39	47.03	57.28

Table 3: Ablation study results of LUSIFER's components. The table presents average metrics for each model, with the highest score for each language emphasized in bold.

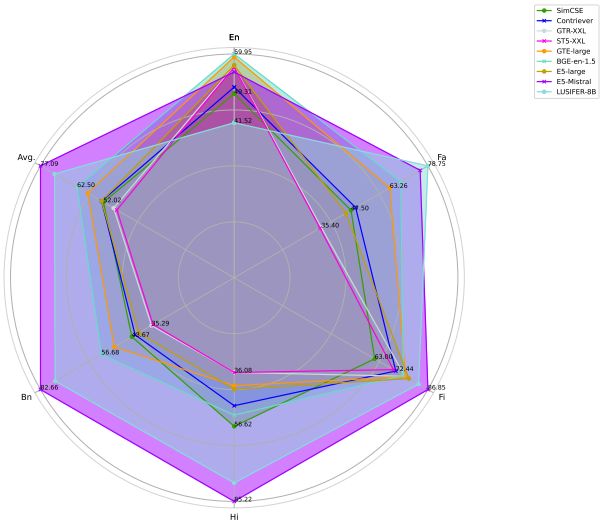


Figure 4: Performance comparison of LUSIFER and baseline models on Reranking tasks.

tiple languages effectively.

4.5 Task-Specific Performance

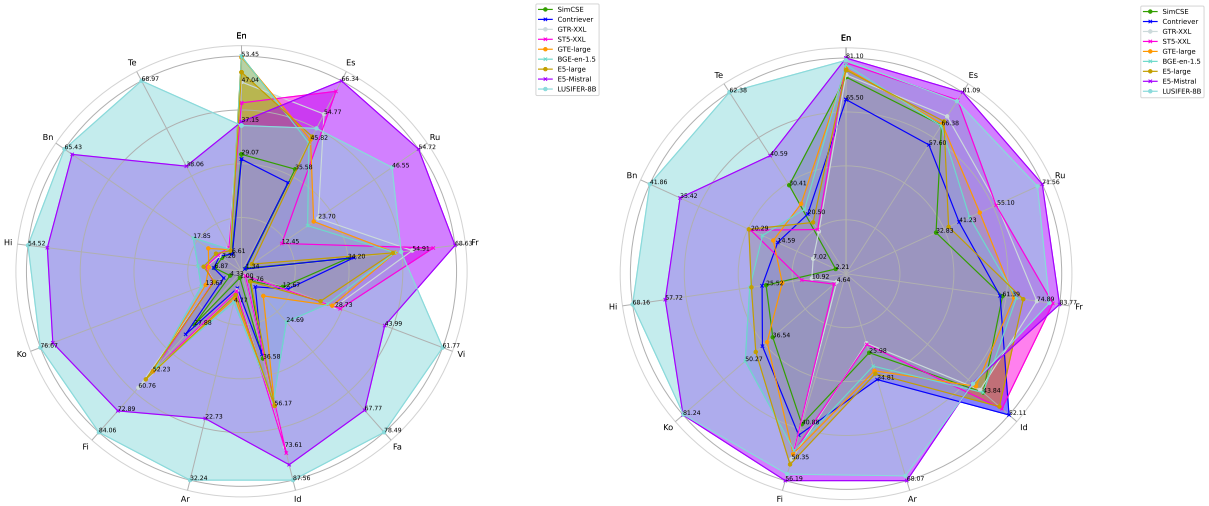
Figure 3, 4, 5 present the performance comparison of LUSIFER and baseline models on Classification, Clustering, Reranking, Retrieval, and STS tasks. LUSIFER consistently outperforms the baseline models across 4 out of 5 tasks, with the largest improvements observed in Clustering and Retrieval tasks, especially in the medium and low-resource languages. However, the performance of LUSIFER in the Reranking tasks is slightly worse than the baseline models. This discrepancy may be attributed to the task's complexity and the information loss in the alignment process between the multilingual encoder and the target LLM. Nevertheless, LUSIFER's strong performance across a variety of tasks and languages highlights its ability to enhance multilingual representations without relying on explicit multilingual training data.

4.6 Ablation Study

To evaluate the effectiveness of LUSIFER's components and training procerdure, we conduct an ablation study to analyze the impact of each component on the model's performance. We compare the performance of LUSIFER with the following ablated versions: (1) LUSIFER with only finetuning connector in both alignment training and representation finetuning stages, (2) LUSIFER with frezzing the multilingual encoder while training the connector and the target LLM in both stages, (3) LUSIFER with only alignment training, i.e., alignment training without representation finetuning, (4) LUSIFER with only representation finetuning without alignment training. Table 3 presents the results of the ablation study. The full LUSIFER model achieves the highest average score of 62.63 across all languages, outperforming the ablated versions. Notably, the alignment training and representation finetuning stages both contribute to the model's performance, with the representation finetuning stage showing a more substantial impact on the model's performance. These results underscore the importance of each component in LUSIFER's architecture and training process, highlighting the model's effectiveness in enhancing multilingual representation capabilities.

4.7 Model Representation Visualization

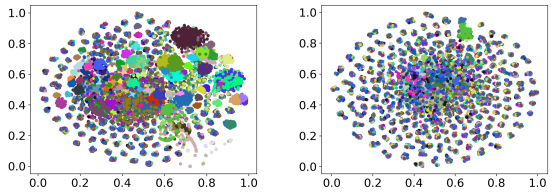
Figure 6 shows 2D scatter plots of representations from different models for 200 randomly sampled examples from the SIB200 dataset, visualized using t-SNE. The points are colored by the language of the samples. The t-SNE representation of E5-Mistral demonstrates a clearer separation between languages, with distinct clusters for each language. In contrast, the visualization of LUSIFER presents a more mixed distribution of languages, with overlapping clusters across different languages. This observation provides insights into LUSIFER's lingual-agnostic capabilities, highlighting the model's ability to bridge the gaps between representation spaces of different languages. These results suggest that LUSIFER's alignment strategy enables the model to comprehend semantics across multiple languages effectively, facilitating zero-shot multilingual transfer. Overall, our experiments confirm the advantages of the representation alignment strategies in LUCIFER to ef-



(a) Retrieval tasks

(b) STS tasks

Figure 5: Performance comparison of LUSIFER and baseline models on Retrieval and STS tasks.



(a) E5-Mistral (language)

(b) LUSIFER (language)

Figure 6: t-SNE representation of 200 randomly samples from the SIB200 dataset. The points are colored by the languages.

fectively enable zero-shot multilingual transfer for LLM-based embedding methods.

5 Conclusion

In this work, we propose LUSIFER, a novel framework that enables effective multilingual representation without explicit multilingual training data. LUSIFER aligns a multilingual encoder with a target English-centric LLM through a minimal set of trainable parameters, facilitating zero-shot multilingual transfer. Our experimental results demonstrate that LUSIFER achieves state-of-the-art performance across diverse languages and tasks, outperforming existing baseline models. Moreover, LUSIFER significantly enhances cross-lingual capabilities, enabling the model to process text-embedding tasks across multiple languages effectively. Our work provides a promising direction for enhancing multilingual representation capabilities in English-centric embedding models, enabling global applicability without requiring extensive multilingual training data. In future work, we plan to explore additional alignment strategies and further investigate the impact of LUSIFER's components on multilingual representation quality.

References

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. *Preprint*, arXiv:2309.07445.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pages 4623–4637, Online. Association for Computational Linguistics.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 749–775. Association for Computational Linguistics.

Rachit Bansal, Bidisha Samanta, Siddharth Dalmia, Nitish Gupta, Sriram Ganapathy, Abhishek Bapna, Prateek Jain, and Partha Talukdar. 2024. LLM augmented LLMs: Expanding capabilities through composition. In *The Twelfth International Conference on Learning Representations*.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *Preprint*, arXiv:2404.05961.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

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. *Preprint*, arXiv:2402.03216.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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.

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

William Falcon and The PyTorch Lightning team. 2024. Pytorch lightning.

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan. 2021a. Scaling deep contrastive learning batch size under memory limited setup. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 316–321, Online. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. 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.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Preprint*, arXiv:2112.09118.

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. *Preprint*, arXiv:2310.06825.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *Preprint*, arXiv:2405.17428.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020a. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. PAQ: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *Preprint*, arXiv:2308.03281.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Jiapeng Liu, Xiao Zhang, Dan Goldwasser, and Xiao Wang. 2020. Cross-lingual document retrieval with smooth learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3616–3629, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. 2024b. Chatqa: Surpassing gpt-4 on conversational qa and rag. *arXiv preprint arXiv:2401.10225*.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. 2024. Ovis: Structural embedding alignment for multimodal large language model. *Preprint*, arXiv:2405.20797.

Kun Luo, Minghao Qin, Zheng Liu, Shitao Xiao, Jun Zhao, and Kang Liu. 2024. Large language models as foundations for next-gen dense retrieval: A comprehensive empirical assessment. *Preprint*, arXiv:2408.12194.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, and Thien Huu Nguyen. 2024. Ullme: A unified framework for large language model embeddings with generation-augmented learning. *Preprint*, arXiv:2408.03402.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *Preprint*, arXiv:1310.4546.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. *Preprint*, arXiv:2402.09906.

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

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y Zhao, Yi Luan, Keith B Hall, Ming-Wei Chang, et al. 2021a. Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:2112.07899*.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2021b. Large dual encoders are generalizable retrievers. *Preprint*, arXiv:2112.07899.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. 2021c. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *Preprint*, arXiv:2108.08877.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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, Philip Beyer, and Iryna Gurevych. 2016. Task-oriented intrinsic evaluation of semantic textual similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 87–96, Osaka, Japan. The COLING 2016 Organizing Committee.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic. Association for Computational Linguistics.

Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. 2021. Cross-lingual training of dense retrievers for document retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 251–253, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Nandan Thakur, Jianmo Ni, Gustavo Hernández Ábrego, John Wieting, Jimmy Lin, and Daniel Cer. 2024. Leveraging llms for synthesizing training data across many languages in multilingual dense retrieval. *Preprint*, arXiv:2311.05800.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

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

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

Boxin Wang, Wei Ping, Lawrence McAfee, Peng Xu, Bo Li, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Instructretro: Instruction tuning post retrieval-augmented pretraining. *arXiv preprint arXiv:2310.07713*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024a. Text embeddings by weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

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

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

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Genta Indra Winata, Ruochen Zhang, and David Ifeoluwa Adelani. 2024. Miners: Multilingual language models as semantic retrievers. *Preprint*, arXiv:2406.07424.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. 2024. LangBridge: Multilingual reasoning without multilingual supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7502–7522, Bangkok, Thailand. Association for Computational Linguistics.

Bryan Zhang and Amita Misra. 2022. Machine translation impact in E-commerce multilingual search. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 99–109, Abu Dhabi, UAE. Association for Computational Linguistics.

Xin Zhang, Zehan Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, and Min Zhang. 2023. Language models are universal embedders. *arXiv preprint arXiv:2310.08232*.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *Preprint*, arXiv:2407.19669.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. 2023. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Preprint*, arXiv:2304.11277.

A Training Details

The code and dataset for training are available at https://github.com/hieum98/lusifer.

Hyperparameter	Alignment Training	Representation Finetuning
Batch size	256	256
Learning rate	1.5e-4	5e-5
Learning rate scheduler	cosine	cosine
Learning rate warm-up ratio	0.1	0.1
Weight decay	0.01	0.01
Grad norm clipping	1.0	1.0
Epochs	2	1
Optimizer	AdamW	AdamW
Float precision	bf16-mixed	bf16-mixed
LoRA rank	16	16
LoRA alpha	32	32
Random mask ratio	0.5	-
Number of hardnegatives	-	7

Table 4: Training hyperparameters for each stage.

Stage	Dataset	Number of Samples
Alignment Training	Wikitext-103 (Merity et al., 2017)	100,000
	MSMARCO (Bajaj et al., 2018)	100,000
Representation Finetuning	MS MARCO (Bajaj et al., 2018)	100,000
	FEVER (Thorne et al., 2018)	100,000
	PAQ (Lewis et al., 2021)	100,000
	SNLI (Bowman et al., 2015)	100,000
	HotpotQA (Yang et al., 2018)	97,800
	SQuAD (Rajpurkar et al., 2016)	97,400
	FiQA (Maia et al., 2018)	6,420
	NQ (Kwiatkowski et al., 2019)	3,420
	ArguAna (Wachsmuth et al., 2018)	1,280

Table 5: Number of samples used in each dataset for training. The number of negative samples is included in the total number of samples.

B Detailed Results

In this section, we provide detailed results of LUSIFER and E5-Mistral on all benchmark datasets for each language.

Es Datasets	E5-Mistral	LUSIFER
AmazonReviewsClassification	42.69	50.41
MassiveIntentClassification	69.67	68.93
MassiveScenarioClassification	74.63	73.41
MTOPIntentClassification	72.16	80.13
MultilingualSentimentClassification	87.91	91.01
TweetSentimentClassification	49.73	58.55
SpanishNewsClassification	89.5	87.81
PawsXPairClassification	61.19	62.82
XNLI	77.34	60.49
SpanishNewsClusteringP2P	42.28	43.85
MLSUMClusteringP2P	47.54	44.36
MLSUMClusteringS2S	47.11	41.56
SIB200ClusteringS2S	31.01	44.42
MultiEURLEXMultilabelClassification	6.16	3.87
BelebeleRetrieval	83.92	81.4
MintakaRetrieval	48.77	18.17
STS17	87.18	80.84
STS22	71.79	70.66
STSBenchmarkMultilingualSTS	84.31	79.89
Avg.	61.84	60.14

Table 6: Detailed results of E5-Mistral and LUSIFER on the Spanish benchmark datasets.

En Datasets	E5-Mistral	LUSIFER
AmazonCounterfactualClassification	78.69	72.45
AmazonPolarityClassification	95.91	94.3
AmazonReviewsClassification	55.79	55.46
Banking77Classification	88.23	87.33
EmotionClassification	49.77	74
ImdbClassification	94.78	92.52
MassiveIntentClassification	80.57	75.64
MassiveScenarioClassification	82.39	78
MTOPDomainClassification	96.12	96.81
MTOPIntentClassification	86.11	87.34
ToxicConversationsClassification	69.59	82.84
TweetSentimentExtractionClassification	63.72	72.74
SprintDuplicateQuestions	95.66	90.99
TwitterSemEval2015	81.62	68.49
TwitterURLCorpus	87.75	85.35
ArxivClusteringP2P	50.45	35.6
ArxivClusteringS2S	45.5	22.25
BiorxivClusteringP2P	43.53	39.93
BiorxivClusteringS2S	40.24	29.3
MedrxivClusteringP2P	38.19	41.2
MedrxivClusteringS2S	37.45	35.53
RedditClustering	57.71	39.94
RedditClusteringP2P	66.49	53.4
StackExchangeClustering	73.1	46.41
StackExchangeClusteringP2P	45.91	39.7
TwentyNewsgroupsClustering	54.31	38.5
AskUbuntuDupQuestions	66.98	60.56
MindSmallReranking	32.6	24.55
SciDocsRR	86.33	34.94
StackOverflowDupQuestions	54.91	46.04
ArguAna	61.88	74.15
ClimateFEVER	38.4	29.24
CQADupstackTexRetrieval	42.97	23.22
DBPedia	48.9	17.98
FEVER	87.8	82.77
FiQA2018	56.62	14.91
HotpotQA	75.7	49.04
MSMARCO	43.1	56.43
NFCorpus	38.59	5.48
NQ	63.5	42.95
QuoraRetrieval	89.62	89.1
SCIDOCS	16.27	5.53
SciFact	76.41	66.09
Touche2020	26.39	6.33
TRECCOVID	87.33	18.22
STS12	79.65	74.26
STS13	88.43	84.2
STS14	84.54	77.5
STS15	90.42	84.95
STS16	87.68	82.21
STS17	91.75	81.67
STS22	67.28	71.25
BIOSSES	82.64	84.22
SICK-R	80.76	78
STSBenchmark	88.6	84.18
SummEval	31.4	32.36
Avg.	67.69	57.20

Table 7: Detailed results of E5-Mistral and LUSIFER on the English benchmark datasets.

Ru Datasets	E5-Mistral	LUSIFER
GeoreviewClassification	46.92	43.79
HeadlineClassification	76.52	79.26
InappropriatenessClassification	59.35	63.15
KinopoiskClassification	60.67	60.57
MassiveIntentClassification	72.06	71.29
MassiveScenarioClassification	76.64	74.49
RuReviewsClassification	64.10	67.40
RuSciBenchGRNTIClassification	60.19	59.51
RuSciBenchOECDClassification	46.30	46.41
GeoreviewClusteringP2P	69.87	59.20
RuSciBenchGRNTIClusteringP2P	52.96	55.00
RuSciBenchOECDClusteringP2P	46.54	49.95
TERRa	57.45	54.24
RiaNewsRetrieval	71.39	49.61
RuBQRetrieval	38.04	43.48
RuSTSBenchmarkSTS	81.79	78.20
STS22	61.32	61.44
Avg.	61.30	59.82

Table 8: Detailed results of E5-Mistral and LUSIFER on the Russian benchmark datasets.

Fr Datasets	E5-Mistral	LUSIFER
AmazonReviewsClassification	43.36	49.96
MTOPIntentClassification	70.39	79.14
MassiveIntentClassification	71.12	70.88
MassiveScenarioClassification	74.68	73.96
TweetSentimentClassification	50.23	62.62
SIB200Classification	72.45	79.51
FrenchBookReviews	46.77	48.07
PawsXPairClassification	62.15	65.93
RTE3	88.45	87.62
XNLI	76.60	62.75
MasakhaNEWSClusteringP2P	50.96	48.59
MasakhaNEWSClusteringS2S	52.08	63.12
MLSUMClusteringP2P	42.69	42.70
MLSUMClusteringS2S	42.60	41.51
HALClusteringS2S	24.21	24.16
SIB200ClusteringS2S	29.94	43.30
MultiEURLEXMultilabelClassification	5.00	3.51
BelebeleRetrieval	84.66	83.76
MintakaRetrieval	52.60	18.88
OpusparcusPC	94.58	90.63
STS17	84.66	82.19
SICKFr	79.12	74.22
STS22	76.50	73.77
STSBenchmarkMultilingualSTS	83.98	78.42
SummEvalFr	31.38	31.91
Avg.	59.65	59.24

Table 9: Detailed results of E5-Mistral and LUSIFER on the French benchmark datasets.

Vi Datasets	E5-Mistral	LUSIFER
MassiveIntentClassification	66.36	71.38
MassiveScenarioClassification	70.69	74.82
MultilingualSentimentClassification	69.30	81.30
SIB200Classification	70.20	78.58
VieStudentFeedbackClassification	73.02	77.39
XNLI	71.32	61.30
SIB200ClusteringS2S	32.93	46.79
BelebeleRetrieval	79.20	85.51
MLQARetrieval	32.43	54.61
VieQuADRetrieval	20.35	45.20
Avg.	58.58	67.69

Table 10: Detailed results of E5-Mistral and LUSIFER on the Vietnamese benchmark datasets.

Fa Datasets	E5-Mistral	LUSIFER
MassiveScenarioClassification	76.37	77.94
MassiveIntentClassification	71.98	73.32
MultilingualSentimentClassification	80.07	80.54
FarsTail	63.49	67.98
WikipediaRerankingMultilingual	75.60	78.75
WikipediaRetrievalMultilingual	67.77	78.49
Avg.	72.55	76.17

Table 11: Detailed results of E5-Mistral and LUSIFER on the Farsi benchmark datasets.

Id Datasets	E5-Mistral	LUSIFER
IndonesianMongabayConservationClassification	24.72	25.27
MassiveIntentClassification	69.51	71.38
MassiveScenarioClassification	72.89	74.62
SIB200Classification	80.88	80.44
indonli	50.00	50.22
SIB200ClusteringS2S	46.46	47.50
BelebeleRetrieval	81.10	87.56
SemRel24STS	40.40	40.57
Avg.	58.25	59.70

Table 12: Detailed results of E5-Mistral and LUSIFER on the Indonesian benchmark datasets.

Ar Datasets	E5-Mistral	LUSIFER
TweetEmotionClassification	53.74	49.03
ArEntail	77.63	84.15
XNLI	68.00	58.58
MintakaRetrieval	17.15	16.59
MLQARetrieval	28.32	47.90
STS17	75.13	71.44
STS22	61.01	61.54
Avg.	54.43	55.60

Table 13: Detailed results of E5-Mistral and LUSIFER on the Arabic benchmark datasets.

Fi Datasets	E5-Mistral	LUSIFER
FinToxicityClassification	53.78	62.23
MassiveIntentClassification	64.15	70.77
MassiveScenarioClassification	67.79	75.02
MultilingualSentimentClassification	72.42	83.59
SIB200Classification	66.57	77.06
WikipediaRerankingMultilingual	86.85	82.65
BelebeleRetrieval	73.89	85.18
WikipediaRetrievalMultilingual	71.90	82.94
OpusparcusPC	91.41	91.63
FinParaSTS	20.97	17.24
Avg.	66.97	72.83

Table 14: Detailed results of E5-Mistral and LUSIFER on the Finnish benchmark datasets.

Ko Datasets	E5-Mistral	LUSIFER
MassiveIntentClassification	70.42	69.79
MassiveScenarioClassification	75.12	75.60
KorSarcasmClassification	57.64	55.28
SIB200Classification	72.70	77.89
KorHateSpeechMLClassification	8.49	7.54
PawsXPairClassification	53.10	54.97
KLUE-TC	60.58	63.95
SIB200ClusteringS2S	31.04	46.58
Ko-StrategyQA	63.81	68.66
BelebeleRetrieval	80.09	84.69
KLUE-STS	83.48	84.17
KorSTS	79.28	78.36
STS17	80.97	80.55
Avg.	62.82	65.23

Table 15: Detailed results of E5-Mistral and LUSIFER on the Korean benchmark datasets.

Hi Datasets	E5-Mistral	LUSIFER
MTOPIntentClassification	68.84	79.93
SentimentAnalysisHindi	58.98	73.92
MassiveIntentClassification	64.69	71.01
MassiveScenarioClassification	69.71	75.42
SIB200Classification	68.43	75.98
TweetSentimentClassification	37.70	40.78
XNLI	65.04	60.26
IndicReviewsClusteringP2P	40.04	42.40
SIB200ClusteringS2S	27.32	45.62
WikipediaRerankingMultilingual	85.22	78.17
BelebeleRetrieval	69.73	66.76
MintakaRetrieval	18.60	21.53
MLQARetrieval	35.37	54.54
WikipediaRetrievalMultilingual	74.62	75.25
IndicCrosslingualSTS	42.30	58.97
SemRel24STS	73.14	77.34
Avg.	56.23	62.37

Table 16: Detailed results of E5-Mistral and LUSIFER on the Hindi benchmark datasets.

Bn Datasets	E5-Mistral	LUSIFER
BengaliDocumentClassification	50.78	48.00
BengaliHateSpeechClassification	54.67	51.43
MassiveIntentClassification	59.51	66.65
MassiveScenarioClassification	64.57	70.91
XNLIV2	63.66	60.01
IndicReviewsClusteringP2P	38.20	45.68
SIB200ClusteringS2S	23.88	43.96
WikipediaRerankingMultilingual	82.66	76.39
BelebeleRetrieval	60.17	55.77
IndicQARetrieval	56.59	68.06
WikipediaRetrievalMultilingual	71.05	72.47
IndicCrosslingualSTS	35.42	41.86
Avg.	55.10	58.43

Table 17: Detailed results of E5-Mistral and LUSIFER on the Bengali benchmark datasets.

Te Datasets	E5-Mistral	LUSIFER
IndicNLPNewsClassification	89.46	98.90
IndicSentimentClassification	61.53	90.63
MassiveIntentClassification	47.34	68.69
MassiveScenarioClassification	51.67	74.17
SIB200Classification	46.23	74.56
TeluguAndhraJyotiNewsClassification	67.40	76.24
IndicReviewsClusteringP2P	34.02	43.62
SIB200ClusteringS2S	10.81	42.11
BelebeleRetrieval	42.46	80.32
IndicQARetrieval	33.67	57.61
IndicCrosslingualSTS	8.36	43.76
SemRel24STS	72.83	80.99
Avg.	47.15	69.30

Table 18: Detailed results of E5-Mistral and LUSIFER on the Telugu benchmark datasets.

Sw Datasets	E5-Mistral	LUSIFER
AfriSentiClassification	39.67	46.47
MasakhaNEWSClassification	72.96	74.79
MassiveIntentClassification	52.84	52.79
MassiveScenarioClassification	61.09	58.59
SwahiliNewsClassification	63.95	61.56
XNLI	58.86	57.82
MasakhaNEWSClusteringP2P	34.15	36.95
MasakhaNEWSClusteringS2S	21.34	35.97
Avg.	50.61	53.12

Table 19: Detailed results of E5-Mistral and LUSIFER on the Swahili benchmark datasets.