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

Hieu Man*, Nghia Trung Ngo*, Viet Dac Lai^, Ryan A. Rossi^, Franck Dernoncourt^, Thien Huu Nguyen*
*University of Oregon, ^Adobe Research

SIGIR 2025 Padova ITALY

INTRODUCTION

Why LLM for Embedding tasks?

- LLMs pretrained with **all input tokens,** much **more sample-efficient** than encoder-only models.
- LLMs **excel at instruction**, an ideal choice for building **universal text embedding models**.

Limitations:

- Current LLM-based embeddings excel in English but **underperform in multilingual scenarios.**
- LLMs' **causal attention mechanism** restricts the model's attention to only preceding tokens.
- Lack of **comprehensive evaluations benchmark** for multilingual text embedding.

➔ LUSIFER: **zero-shot approach** to adapting English-focused LLM for multilingual text embedding tasks.

➔ LUSIFER's benchmark: **123 diverse datasets in 14 languages**, focusing on five fundamental embedding tasks

FRAMEWORK

LUSIFER includes the three key components:

- Multilingual encoder as language-universal learner
- Connector with minimal trainable parameters
- Target LLM optimized for embedding tasks

Two-Stage Training:

- Stage 1: Cross-Lingual Representation Alignment: **Establishes a universal semantic space** connecting multilingual encoder with English-centric LLMs
  - **Masked Reconstruction**: Predicts original tokens from corrupted inputs using cross-entropy loss.
  - **Autoregressive Completion**: Generates answers for QA
- Stage 2: Representation Finetuning: **enhances embedding quality through contrastive learning** while maintaining cross-lingual alignment
  - **Bidirectional Attention**: Integrates forward and backward context modeling to improve sequence representations
  - Representation learning with in-batch negatives **contrastive loss**

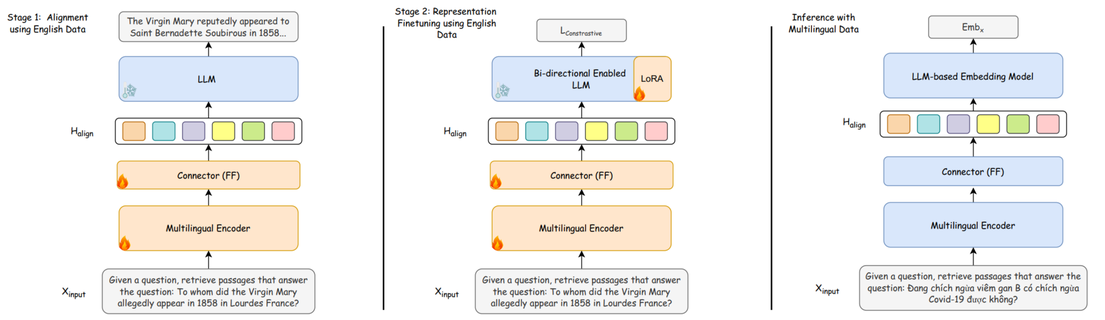


Figure 1: Overview of LUSIFER. Left: Align a multilingual encoder with the target English-centric LLM only using English data and a minimal set of trainable parameter. Center: End-to-end representation finetune through contrastive learning on English text-embedding tasks using LoRA. Right: During inference,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

MAIN RESULTS

Baselines	En	Es	Ru	Fr	Vi	Fa	Id	Ar	Fi	Ko	Hi	Bn	Te	Sw	Avg.
Jina-embeddings-v3* [55]	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* [73]	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* [10]	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* [64]	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* [72]	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 [16]	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 [20]	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 [29]	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 [68]	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 [61]	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 [45]	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 [44]	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 [62]	**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 [16]	7.41	18.35	39.71	37.95	0.18	20.72
Contriever [20]	9.75	22.94	34.55	41.72	0.03	21.80
GTE-large [29]	16.99	31.82	37.57	53.79	1.59	28.35
BGE-en-1.5 [68]	16.64	31.19	40.40	50.77	1.11	28.02
E5-large [61]	17.04	31.12	37.90	54.31	1.83	28.44
ST5-XXL [45]	20.82	41.68	56.19	59.02	1.76	35.89
GTR-XXL [44]	20.19	38.02	50.83	60.11	2.74	34.38
E5-Mistral [62]	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.

LUSIFER'S BENCHMARK

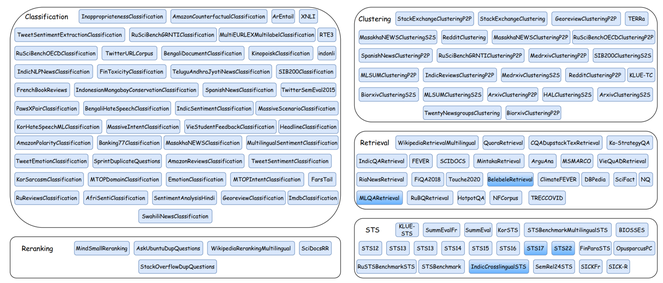


Figure 2: Overview of tasks and datasets in our benchmark. Crosslingual datasets are marked with a blue shade.

EXPERIMENTS

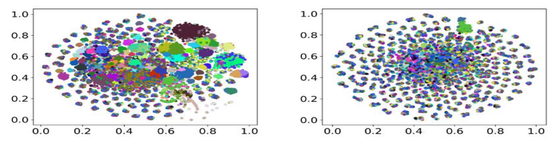
Ablation Study

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 5: 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.

- Should allow the multilingual encoder and the LLM to be finetuned during training
- Two-stage training approach complement each other, especially the importance of the representation finetuning stage

Visualization



(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.

- Present a more mixed distribution of languages, with overlapping clusters across different languages

➔ Lingual-agnostic capability, bridge the gaps between representation spaces of different languages

REFERENCES

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang,Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. Preprint, arXiv:2401.00368.

Parishad BehnamGhader, Vaibhav Adlakha, MariusMosbach, Dzmitry Bahdanau, Nicolas Chapados, andSiva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. Preprint,arXiv:2404.05961.

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

Luyu Gao, Yunyi Zhang, Jiawei Han, and Jamie Callan.2021a. Scaling deep contrastive learning batch size under memory limited setup. In Proceedings of the6th Workshop on Representation Learning for NLP(RepL4NLP-2021),