

## 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 Englishfocused 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

**Fwo-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
- Representation Finetuning: enhances 2: Stage 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

# 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

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



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

I	En	L3	Ru	11	•••	1a	10	AI	11	Ro	111	DII	10	
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	49.2
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	52.5
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	54.2
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	51.5
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	47.4
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	38.3
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.
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	41.4
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.
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.
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.
GIR-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.
E5-Mistral [62]	00.04	61.84	61.30	59.65	38.38	72.55	58.25	54.45	66.97	62.82	56.25	55.10	47.15	50.
LUSIFER (Ours)	57.20	60.14	59.82	59.24	67.69	76.17	<b>59.70</b>	55.60	72.83	65.23	62.37	58.43	69.30	53.
Cable 1: Comparative and   For each model, with the nultilingual data.   Baselines	nalysis ne high	of mod nest sco	lel perf ore for	forman each	langua	oss mul ge emp	ohasize	anguag d in bo	es and old. * d	tasks. lenotes	The tat	ole pres	sents a trained	vera l on
Cable 1: Comparative and   For each model, with the   nultilingual data.   Baselines	nalysis ne high MLQ	of mod nest sco QAReti	lel perfor ore for	forman each Bel	lebelel	ss mu ge emp Retrie	val	inguag d in bo STS17	es and old. * d	tasks. lenotes S22	Ine tat	cCross	sents a trained	vera l on al
Cable 1: Comparative and   For each model, with the nultilingual data.   Baselines   Baselines   SimCSE [16]	nalysis ne high MLQ	of mod nest sco QARetr 7.41	lel perfor	forman each Bel	lebelel	ss mu ge emp Retriev 35	val	inguag d in bo STS17 39.71	es and old. * d ' ST 37	tasks. lenotes S22 7.95	Ine tat	cCross	sents a trained	vera I on al
Cable 1: Comparative and for each model, with the nultilingual data.BaselinesBaselinesSimCSE [16]Contriever [20]	nalysis ne high MLQ	of mod nest sco 2AReti 7.41 9.75	lel perfor	forman each Bel	lebelel 18. 22.	ss mul ge emp Retriev 35 94	val	nguag d in bo STS17 39.71 34.55	es and old. * d ST 37 41	tasks. lenotes S22 7.95	Ine tat	Cross 0.18 0.03	sents a trained	vera I on al
Cable 1: Comparative and for each model, with the nultilingual data.BaselinesSimCSE [16]Contriever [20]GTE-large [29]	nalysis ne high MLQ	of mod nest sco 2 7.41 9.75 16.99	lel perfor	forman each Bel	lebelel 18. 22. 31.	ss mul ge emp Retrie 35 94 82	val	anguag d in bo STS17 39.71 34.55 37.57	es and old. * d 37 41 53	tasks. lenotes S22 7.95 1.72 3.79	Ine tat	Cross 0.18 0.03 1.59	sents a trained	vera l on al
Cable 1: Comparative and for each model, with the nultilingual data.BaselinesSimCSE [16]Contriever [20]GTE-large [29]BGE-en-1.5 [68]	nalysis ne high MLQ	of mod nest sco 2AReti 7.41 9.75 16.99 16.64	lel perfor	forman each Bel	lebelel 18. 22. 31. 31.	oss mul ge emp Retriev 35 94 82 19	val	anguag d in bo STS17 39.71 34.55 37.57 40.40	es and old. * d 37 41 53 50	tasks. lenotes S22 7.95 1.72 3.79 0.77	Ine tat	cCross 0.18 0.03 1.59 1.11	sents a trained	vera l on al
Cable 1: Comparative and for each model, with the nultilingual data.BaselinesBaselinesSimCSE [16]Contriever [20]GTE-large [29]BGE-en-1.5 [68]E5-large [61]	nalysis ne high MLQ	of mod nest sco 2AReti 7.41 9.75 16.99 16.64 17.04	lel perfor	forman each l Bel	lebelel 18. 22. 31. 31. 31.	ss mul ge emp 35 94 82 19 12	val	anguag d in bo STS17 39.71 34.55 37.57 40.40 37.90	es and old. * d 37 41 53 50 54	tasks. lenotes S22 7.95 1.72 3.79 3.79 0.77 4.31	Ine tak	Cross 0.18 0.03 1.59 1.11 1.83	sents a trained	vera l on al
Cable 1: Comparative and for each model, with the nultilingual data.BaselinesBaselinesSimCSE [16]Contriever [20]GTE-large [29]BGE-en-1.5 [68]E5-large [61]ST5-XXL [45]	nalysis ne high MLQ	of mod hest sco 2AReti 7.41 9.75 16.99 16.64 17.04 20.82	lel perfor	forman each I Bel	lebelel 18. 22. 31. 31. 31. 41.	ss mul ge emp 35 94 82 19 12 68	val	anguag d in bo STS17 39.71 34.55 37.57 40.40 37.90 56.19	es and old. * d 37 41 53 50 54 59	tasks. lenotes S22 7.95 1.72 3.79 0.77 4.31 9.02	Ine tat	cCross 0.18 0.03 1.59 1.11 1.83 1.76	sents a trained	vera l on al
Cable 1: Comparative an for each model, with the nultilingual data.BaselinesBaselinesSimCSE [16] Contriever [20] GTE-large [29]BGE-en-1.5 [68] E5-large [61] ST5-XXL [45] GTR-XXL [44]	nalysis ne high MLQ	of mod nest sco 2AReti 7.41 9.75 16.99 16.64 17.04 20.82 20.19	lel perfor	forman each l Bel	lebelel 18. 22. 31. 31. 31. 31. 31. 33.	ss mul ge emp 35 94 82 19 12 68 02	val	anguag d in bo STS17 39.71 34.55 37.57 40.40 37.90 56.19 50.83	es and old. * d 37 41 53 50 54 59 60	tasks. lenotes S22 7.95 1.72 3.79 1.72 3.79 1.77 4.31 9.02 0.11	Ine tat	Cross 0.18 0.03 1.59 1.11 1.83 1.76 2.74	sents a trained	vera l on al
Cable 1: Comparative an for each model, with the nultilingual data.BaselinesSimCSE [16]Contriever [20]GTE-large [29]BGE-en-1.5 [68]E5-large [61]ST5-XXL [45]GTR-XXL [44]E5-Mistral [62]	nalysis ne high MLQ	of mod hest sco 2AReti 9.75 16.99 16.64 17.04 20.82 20.19 31.54	lel perfor	forman each l Bel	lebelel 18. 22. 31. 31. 31. 31. 31. 31. 31. 31. 31. 31	ss mul ge emp 35 94 82 19 12 68 02 75	val	sTS17 39.71 34.55 37.57 40.40 37.90 56.19 50.83 <b>81.12</b>	es and old. * d 37 41 53 50 54 59 60 71	tasks. lenotes S22 7.95 1.72 3.79 0.77 4.31 0.02 0.11 1.37	Ine tak	Cross 0.18 0.03 1.59 1.11 1.83 1.76 2.74 21.92	sents a trained	vera l on al

Hieu Man, Nghia Trung Ngo, Viet Dac Lai, Ryan A. Rossi, Franck Dernoncourt, Thien Huu Nguyen. LUSIFER: Language Universal Space Integration for Enhanced Multilingual Embeddings with Large Language Models @SIGIR2025



#### Ablation Study

LUSIFER	(Full)
LUSIFER LUSIFER LUSIFER LUSIFER Table 5:	(Connector Only) (Frozen Multilingual Encoder) (Alignment Only) (Representation Finetuning Or Ablation study results o
score for	Should allow
	during trainir
٠	Two-stage tra
	importance o
Vis	ualization
	1.0
	0.8
	0.4
	0.2
	0.0 0.2 0
	(a) E5-Mis
] 1	Figure 6: t-SN the SIB200 da
•	Figure 6: t-SN the SIB200 da Present a mo
•	Figure 6: t-SN the SIB200 da Present a mo clusters acro
•	Figure 6: t-SN the SIB200 da Present a mo clusters acro Lingual-agno
• • spa	Figure 6: t-SN the SIB200 da Present a mo clusters acro Lingual-agno ces of differe
, ∙ spa	Figure 6: t-SN the SIB200 da Present a mo clusters acro Lingual-agno ces of differe
, ∙ spa	Figure 6: t-SN the SIB200 da Present a mo clusters acro Lingual-agno ces of differe
, ∙ spa	Figure 6: t-SN the SIB200 da Present a mo clusters acro Lingual-agno ces of differe
J • • Spa Lian Impr	Figure 6: t-SN the SIB200 da Present a mo clusters acro Lingual-agno ces of differe
∎ Spa Lian Impi Paris	Figure 6: t-SN the SIB200 da Present a mo clusters acro Lingual-agno ces of differe g Wang, Nan Yan oving text embeddingshad BehnamGhade
■ Spa Lian Impr Paris andS Prep	Figure 6: t-SN the SIB200 da Present a mo clusters acro Lingual-agno ces of differe g Wang, Nan Yan oving text embeddin shad BehnamGhade Siva Reddy. 2024.
■ Spa Lian Impr Paris andS Prep Edw	Figure 6: t-SN the SIB200 da Present a mo clusters acro Lingual-agno ces of differe g Wang, Nan Yan oving text embeddin shad BehnamGhade Siva Reddy. 2024. rint,arXiv:2404.059 ard J Hu, yelong sh
■ Spa Lian Impr Paris andS Prep Edw Weiz	Figure 6: t-SN the SIB200 da Present a mo clusters acro Lingual-agno ces of differe g Wang, Nan Yan coving text embeddin shad BehnamGhade Siva Reddy. 2024. rint,arXiv:2404.059 ard J Hu, yelong sh zhuChen. 2022. Lop
■ Spa Lian Impr Paris andS Prep Edw Weiz on L Luv	Figure 6: t-SN the SIB200 da Present a mo clusters acro Lingual-agno ces of differe g Wang, Nan Yan coving text embeddin shad BehnamGhade Siva Reddy. 2024. rint,arXiv:2404.059 ard J Hu, yelong sh chuChen. 2022. LoF earning Representat
■ Spa Spa Lian Impr Paris andS Prep Edw Weiz on L Luyu size	Figure 6: t-SN the SIB200 da Present a mo clusters acro Lingual-agno ces of differe g Wang, Nan Yan coving text embeddin shad BehnamGhade Siva Reddy. 2024. rint,arXiv:2404.059 ard J Hu, yelong sh zhuChen. 2022. LoF earning Representat





Clustering StackExchangeClusteringP2P StackExchangeClustering GeoreviewClusteringP2P teringS2S MLSUMClusteringS2S ArxivClusteringP2P HALClusteringS2S ArxivClusteringS wentyNewsgroupsClustering BiorxivClusteringP2P

Padova

2025

lebeleRetrieval ClimateFEVER DBPedia SciFact ieval RuBQRetrieval HotpotQA NFCorpus TRECCOVID

KLUE-STS SummEvalFr SummEval KorSTS STSBenchmarkMultilingualSTS BIOSSE RuSTSBenchmarkSTS STSBenchmark IndicCrosslingualSTS SemRel24STS SICKFr :

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

# EXPERIMENTS

	En	Es	Ru	Fr	Vi	Fa	Id	Ar	Fi	Ко	Hi	Bn	Te	Sw	Avg.
	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
	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
al 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
	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
etuning 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
results of LUSIFER's components. The table presents average metrics for each model, with the highest															
e emphasized in bold.															

allow the multilingual encoder and the LLM to be finetuned raining

ge training approach complement each other, especially the nce of the representation finetuning stage





t-SNE representation of 200 randomly samples from 00 dataset. The points are colored by the languages. a more mixed distribution of languages, with overlapping s across different languages

-agnostic capability, bridge the gaps between representation ifferent languages

## REFERENCES

an Yang, Xiaolong Huang, Linjun Yang,Rangan Majumder, and Furu Wei. 2024. mbeddings with large language models. Preprint, arXiv:2401.00368.

nGhader, Vaibhav Adlakha, MariusMosbach, Dzmitry Bahdanau, Nicolas Chapados, 2024. Llm2vec: Large language models are secretly powerful text encoders. 04.05961

long shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and 22. LoRA: Low-rank adaptation of large language models. In International Conference esentations.

Zhang, Jiawei Han, and Jamie Callan.2021a. Scaling deep contrastive learning batch bry limited setup. In Proceedings of the6th Workshop on Representation Learning for