Improving Text Representations with Large Language Models

Hieu Man
Dept. of Computer Science, University of Oregon, OR, USA
hieum@uoregon.edu

Abstract

Recent advances in large language models (LLMs) have significantly improved the quality of text representations, enabling breakthroughs in dense retrieval, semantic search, and a range of downstream natural language processing tasks. However, leveraging LLMs for effective text embeddings faces persistent challenges, including architectural constraints such as causal attention, misalignment between pre-training and embedding objectives, and limited support for multilingual scenarios. This research addresses these challenges through two complementary contributions. First, we introduce ULLME, a unified framework that enables bidirectional attention and supports diverse fine-tuning strategies-including our novel Generation-augmented Representation Learning (GRL), which aligns embedding and generation objectives to produce richer text embeddings. ULLME consistently outperforms previous methods across a wide range of benchmarks and LLM architectures. Second, we present LUSIFER, a zero-shot multilingual adaptation framework that integrates a multilingual encoder with an LLM-based embedding model via a lightweight connector. Without requiring multilingual supervision, LUSIFER achieves strong multilingual and cross-lingual performance, especially in medium and low-resource languages, as demonstrated on a comprehensive benchmark covering 123 datasets in 14 languages. Together, these contributions advance the state of the art in text representation learning with LLMs by providing both a flexible, high-performance embedding framework and a practical solution for multilingual and cross-lingual embedding tasks.

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.

Despite these advances, significant challenges remain. LLMs, particularly those with decoder-only architectures, are primarily optimized for generative tasks and often employ causal attention mechanisms that limit their capacity for effective bidirectional context modeling-an essential property for high-quality text embeddings (Grattafiori et al., 2024; Jiang et al., 2023a). Moreover, the objectives used during LLM pre-training are typically misaligned with the requirements of dense retrieval and semantic similarity tasks, resulting in suboptimal performance when these models are directly applied to embedding-centric applications. Exist-

ing frameworks for LLM-based embeddings have further been limited by their narrow support for a small set of model architectures and fine-tuning strategies, restricting their practical utility and the pace of innovation in the field (Wang et al., 2024b; Muennighoff et al., 2024a; BehnamGhader et al., 2024a; Lee et al., 2024)

Another critical gap is the lack of robust multilingual embedding capabilities. As 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., 2023b; Wang et al., 2024d; 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.

This report addresses these challenges through two key contributions:

ULLME: A Unified Framework for Large Language Model Embeddings with Generation-Augmented Learning (Man et al., 2024): ULLME introduces a flexible, plug-and-play framework that enables bidirectional attention across a wide range of LLM architectures and supports multiple fine-tuning strategies, including contrastive learning, supervised fine-tuning, and direct preference optimization. Central to ULLME is the novel Generation-augmented Representation Learning (GRL) method, which enforces consistency between representation-based and generation-based relevance scores, leveraging the generative strengths of LLMs to produce richer and more effective embeddings. This unified approach not only addresses the architectural limitations of LLMs in embedding tasks but also streamlines the process of adapting and evaluating models across diverse retrieval and semantic applications.

LUSIFER: Language Universal Space Integration for Enhanced Multilingual Embeddings with Large Language Models (Man et al., 2025): LUSIFER tackles the multilingual gap by introducing a zero-shot framework that adapts English-centric LLM embedding models for multilingual tasks without requiring explicit multilingual supervision. The architecture integrates a robust multilingual encoder with a target LLM through a lightweight connector, enabling the transfer of language-agnostic semantic knowledge. LUSIFER is trained exclusively on English data yet achieves strong zero-shot transfer across 14 languages and 123 datasets, demonstrating significant gains especially for medium and low-resource languages. This approach lowers the barrier for extending advanced LLM embeddings to a truly global audience.

Together, these contributions establish a comprehensive toolkit and methodology for improving text representations with LLMs. By addressing both the technical limitations of current embedding frameworks and the persistent disparities in multilingual representation, this work sets new benchmarks for the field and provides practical solutions for researchers and practitioners seeking to deploy LLM-based embeddings in diverse, real-world scenarios

2 Related Work

2.1 LLMs for Dense Retrieval.

Recent advancements in this area have primarily addressed two key challenges: (i): Overcoming LLMs' Causal Attention Limitations by developing methods to enable bidirectional attention within LLMs (Muennighoff, 2022; Muennighoff et al., 2024b; BehnamGhader et al., 2024b; Lee et al., 2024), allowing models to consider both past and future context when computing embeddings, and (ii): Aligning LLM Pre-training with Text Ranking by fine-tuning LLMs via contrastive learning (Ma et al., 2023; Wang et al., 2024c; Lee et al., 2024). This process can also be augmented with additional objectives such as supervised fine-tuning (SFT) (Muennighoff et al., 2024b) or mask-filling tasks (BehnamGhader et al., 2024b). An alternative approach proposed by Springer et al. (2024) involves a prompting method where the input sequence is duplicated, enabling each token to attend to future tokens and mitigating the contextualization issues inherent in causal attention. While these methods have shown promise, they generally do not explicitly enforce consistency between the model's understanding of relevance in both the embedding and generation spaces. This limitation restricts

Framework	#Supported LLMs	Supported Fine-tuning Strategy		
		SFT	DPO	Contrastive
SentenceTrasformers (Reimers and Gurevych, 2019)	>10	✗	✗	✗
SGPT (Muennighoff, 2022)	1	✗	✗	✓
RepLLaMA (Ma et al., 2023)	1	✗	✗	✓
Echo-Embedding (Springer et al., 2024)	2	✗	✗	✗
GritLM (Muennighoff et al., 2024b)	2	✓	✗	✓
LLM2Vec (BehnamGhader et al., 2024b)	3	✗	✗	✓
NV-Emb (Lee et al., 2024)	1	✗	✗	✓
ULLME (**our**)	>10	✓	✓	✓

Table 1: Comparisions between ULLME and other LLM-Embedding frameworks. For ULLME, the module combination enables many possible models and 10 is the number of models we have tested for usability.

their ability to fully leverage the remarkable generative capabilities of LLMs for dense retrieval tasks. Our work, GRL, builds upon these foundations while addressing their limitations, introducing novel techniques to harmonize embedding-based and generation-based relevance scoring within a unified framework.

2.2 Frameworks of LLMs for Dense Retrieval.

Existing frameworks for LLMs in Dense Retrieval have been constrained by their limited support for LLM architectures and fine-tuning strategies. As shown in Table1, SentenceTransformers(Reimers and Gurevych, 2019) supports various types of LLMs but is primarily designed for inference without allowing fine-tuning, limiting its applicability in advancing state-of-the-art dense retrieval methods. Some recent works (Muennighoff, 2022; Ma et al., 2023; Lee et al., 2024), such as **Echo** (Wang et al., 2024c), **GritLM** (Muennighoff et al., 2024b), **LLM2Vec** (BehnamGhader et al., 2024b), and the models in the Hugging Face's MTEB leaderboard[1], have introduced implementations for LLM-based text embeddings. However, these approaches are often tailored to specific model architectures and training methods with hard-coded implementations, thus restricting their adaptability and use across different LLM architectures and fine-tuning strategies to meet diverse development and application demands. In contrast, our framework ULLME addresses these limitations by offering a flexible and extensible platform. ULLME can accommodate a diverse range of LLM backbones and supports various training approaches, making it highly versatile and broadly applicable.

[1] https://huggingface.co/spaces/mteb/leaderboard

2.3 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 complexity 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., 2024d; 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.

2.4 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 ULLME - Unified framework for Large Language Model Embedding

We present an overview of our ULLME framework in Section 3.1 while Section 3.2 details the key technical methods.

3.1 Overview

ULLME addresses the limitations of existing LLM-based dense retrieval frameworks by offering a flexible and comprehensive solution. The framework operates in three main stages. First, it enables bidirectional attention within LLMs by replacing the causal attention mask with a bidirectional one. This crucial modification extends the models' ability to consider both past and future context when generating embeddings, significantly enhancing its capacity for dense retrieval tasks. The transformed model is then returned as a PyTorch object, providing users with the flexibility to integrate it into various frameworks or pipelines.

```
from ullme.models import ULLME

model = ULLME(
    model_name_or_path="mistralai/Mistral-7B-v0.1",
    model_backbone_type="mistral",
    lora_name="ullme-mistral",
    loar_r=16,
    lora_alpha=32,
    )
input_sentence = "This a example sentence."
model_inputs = model.tokenizer(
    [input_sentence],
    return_tensors='pt'
    )
model_output = model(
    input_ids=model_inputs['input_ids'],
    attention_mask=model_inputs['attention_mask'],
    is_generate=False
    )
>> {'rep': (1, hidden_dim)}
```

Listing 1: Extending bidirectional attention for LLMs via ULLME.

We will elaborate on this process in Section 3.2.1. Second, ULLME supports a diverse array of fine-tuning strategies, including Contrastive Learning, Supervised Fine-tuning (SFT), Direct Preference Optimization (DPO), and our novel Generation-augmented Representation Learning (GRL). This versatility allows for tailored optimization across a wide spectrum of retrieval tasks and domains, as detailed in Section 3.2.2. Finally, the framework streamlines the evaluation process by incorporating direct support for model validation using the Massive Text Embedding Benchmark (MTEB) library (Section 3.3). This integration facilitates comprehensive assessment across numerous retrieval and embedding tasks. By seamlessly combining these elements, ULLME provides an extensive toolkit for leveraging LLMs in diverse dense retrieval tasks, encompassing everything from initial model adaptation to fine-tuning and evaluation. Our comprehensive approach aims to accelerate research and development for of LLM-based dense retrieval, offering researchers and practitioners a comprehensive platform for innovation and advancement.

3.2 Key Features

3.2.1 Enabling Bidirectional Attention

To enable bidirectional attention in LLMs, ULLME requires only minimal code modifications, as illustrated in Listing 1. The framework's user-friendly design allows for easy initialization with various LLM backbones by simply specifying the "`model_name_or_path`" and

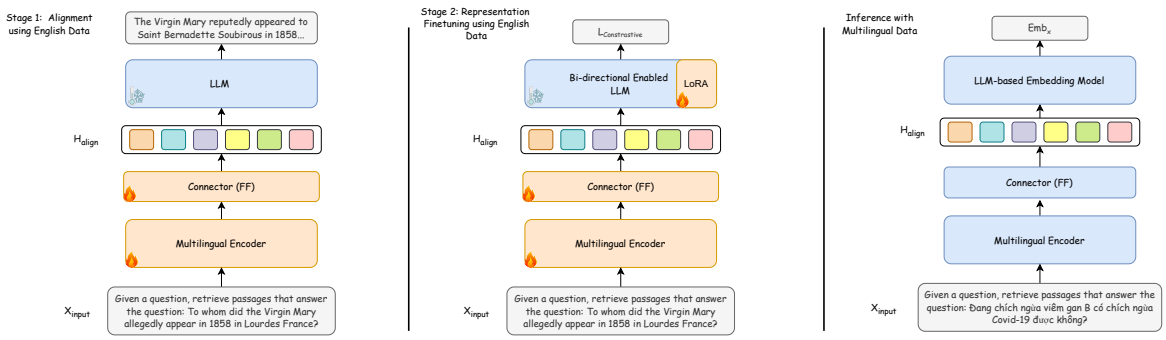


Figure 1: Overview of LUSIFER. **Left**: Align a multilingual encoder with the target English-centric LLM only using English data and a minimal set of trainable parameter. **Center**: End-to-end representation finetune through contrastive learning on English text-embedding tasks using LoRA. **Right**: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

"model_backbone_type" parameters. ULLME seamlessly integrates with Hugging Face Transformers, loading pre-trained LLMs directly from their repository. Additionally, our framework supports parameter-efficient fine-tuning through Low-Rank Adaptation (LoRA) (Hu et al., 2022), offering flexibility in model adaptation. Once initialized, the model can be used to compute sequence representations. The "is_generate" parameter plays a crucial role in controlling the attention mechanism: when set to "False", the model employs bidirectional attention, optimizing it for dense retrieval tasks, while "True" reverts the model to causal attention, mimicking the standard Hugging Face Transformer model output. This dual functionality allows ULLME to serve both as an advanced specialized embedding model and as a language model when needed, providing developers with a flexible tool that can conveniently transition between bidirectional and causal attention modes. ULLME provides various methods for extracting text embeddings from LLMs, such as using representations from the first token, last token, mean, or weighted mean pooling. However, it defaults to averaging the representation vectors from the final layers (mean) for better performance on our datasets.

3.2.2 Fine-tuning Strategies

Our ULLME framework supports multiple fine-tuning strategies, as illustrated in Listing 2.

Contrastive Learning. ULLME's Contrastive Learning objective utilizes in-batch negatives (Chen et al., 2020; Gao et al., 2021b). The contrastive loss is formally defined as: $\mathcal{L}_{CL} = -\log \frac{\exp\left(s_{rt}(q,p^+)\right)}{\exp\left(s_{rt}(q,p^+)\right)+\sum_{p^-\in B}\exp\left(s_{rt}(q,p^-)\right)}$.

Here, B represents a mini-batch, q is the input query, p^+ denotes the positive (relevant) passage, and p^- represents negative (non-relevant) passages sampled from the current training mini-batch. The function $s_{rt}(q,p)$ computes the relevance score between a query and a passage using cosine similarity of the induced representations for q and p. To enhance the effectiveness of Contrastive Learning, especially under limited GPU memory constraints, ULLME incorporates advanced techniques such as GradCache (Gao et al., 2021a) and cross-device contrastive loss computation. These optimizations allow for efficient training with larger batch sizes and more diverse negative samples, which are crucial for learning high-quality representations.

```
from ullme.trainer import GradCacheTrainer

trainer = GradCacheTrainer(
    con_loss_type='NTXentLoss',
    gen_loss_type='dpo', # 'sft'
    use_kl_loss=True
)
trainer.fit_epoch(
    model=model,
    train_loader=train_dataloader,
)
```

Listing 2: Finetuning LLMs for text embedings via ULLME.

Supervised Fine-tuning (SFT). In addition to contrastive learning, ULLME supports SFT, a strategy that enhances LLMs' ability to generate high-quality passages in response to queries. ULLME implements SFT using a next-word prediction objective: $\mathcal{L}_{SFT} = -\frac{1}{N}\sum_{i=1}^{N}\log \pi_\theta(w_i|w_{<i},q)$. Here, N is the length of the positive passage p^+, w_i is the i-th token in p^+, and $\pi_\theta(w|x)$ is the conditional likelihood of w given x, computed by the

LLM θ. Importantly, during SFT loss computation, ULLME reverts to using causal attention, mirroring standard LLM behavior.

Direct Preference Optimization (DPO). ULLME incorporates Direct Preference Optimization (DPO) (Rafailov et al., 2023) as an advanced fine-tuning strategy, offering an alternative to traditional Supervised Fine-tuning (SFT). DPO has demonstrated superior effectiveness in LLM fine-tuning. Moreover, the DPO approach inherently accounts for both preferred and rejected outputs, making it intuitively more suitable for aligning models with text-ranking objectives compared to SFT. In ULLME's implementation, the ground-truth relevant passage p^+ for a query q is treated as the preferred output, while negative and irrelevant passages p^- are considered dispreferred. The DPO loss function is designed to encourage the model to assign higher generation probabilities to p^+ compared to any p^-: $\mathcal{L}_{DPO} = -\log\sigma\left(\beta\log\frac{\pi_\theta(p^+|q)}{\pi_{ref}(p^+|q)} - \beta\log\frac{\pi_\theta(p^-|q)}{\pi_{ref}(p^-|q)}\right)$. In this formulation, σ represents the sigmoid function, β is a scaling factor, and $\pi_{ref}(p|q)$ denotes the conditional likelihood computed by the original pre-trained LLM (the reference model).

In addition to the standard DPO formulation, ULLME includes implementations of advanced variants such as Kahneman-Tversky Optimization (KTO) (Ethayarajh et al., 2024) and Contrastive Preference Optimization (CPO) (Xu et al., 2024). The modular architecture of ULLME facilitates the seamless integration of new preference optimization techniques as they emerge, ensuring that the framework remains at the forefront of LLM fine-tuning advancements. Finally, to maintain consistency with the model's pre-training paradigm, ULLME employs causal attention when computing the DPO loss, similar to the approach used in SFT.

Generation-augmented Representation Learning (GRL). ULLME further introduces a novel fine-tuning strategy GRL that explicitly aligns the LLMs' understanding of passage-query text relevance in embedding and generation spaces to boost representation learning. As such, GRL first computes a generation-based relevance score $s_{gen}(q,p)$ utilizing the conditional generation likelihood of a passage candidate p given input query q from LLMs: $s_{gen}(q,p) = \frac{1}{t}\sum_{i=1}^{t}\log\pi_\theta(w_i|w_{<i},q)$, where t is the length of p and w_i is the i-th token in p.

Next, we seek to recognize the consistency of the query-passage relevance scores obtained from the representations (i.e., $s_{rt}(q,p)$) and the generation likelihood (i.e., $s_{gen}(q,p)$). Particularly, let U be the set of m candidate passages for q. For each candidate passage $p_i \in U$, we compute $s_{rt}(q,p_i)$ and $s_{gen}(q,p_i)$, then normalize these scores to obtain the representation and generation relevance distributions over U: $P_{rt}(q,p_i) = \frac{\exp(s_{rt}(q,p_i))}{\sum_{p'\in U}\exp(s_{rt}(q,p'))}$ and $P_{gen}(q,p_i) = \frac{\exp(s_{gen}(q,p_i))}{\sum_{p'\in U}\exp(s_{gen}(q,p'))}$.

Afterward, we minimize the KL divergence between their distributions: $\mathcal{L}_{KL} = \sum_{p\in U}P_{rt}(q,p)\log\frac{P_{rt}(q,p)}{P_{gen}(q,p)}$, serving as a training signal to enrich representation learning for LLMs.

Finally, the overall training loss for GRL combines the contrastive loss $\mathcal{L}_{CL}$, the direct preference optimization loss $\mathcal{L}_{DPO}$, and the KL-divergence loss $\mathcal{L}_{KL}$: $\mathcal{L}_{GRL} = \lambda_{CL}\mathcal{L}_{CL} + \lambda_{DPO}\mathcal{L}_{DPO} + \lambda_{KL}\mathcal{L}_{KL}$, where λ_{CL}, λ_{DPO}, and λ_{KL} are weighting hyperparameters.

```
from ullme.models import WrappedULLME
from ullme.eval import eval_mteb_dataset

model = WrappedULLME(
    model_name_or_path="mistralai/Mistral-7B-v0.1",
    model_backbone_type="mistral",
    lora_name="ullme-mistral",
    loar_r=16,
    lora_alpha=32,
    model_checkpoint="path/to/your/checkpoint"
    )
eval_result = eval_mteb_dataset(
    model=model,
    dataset_name='MSMARCO',
    langs=['eng'],
    )
>> {'eng': 35.8}
```

Listing 3: Evaluation on MTEB dataset via ULLME.

3.3 Evaluation Process

ULLME streamlines the evaluation process by integrating direct support for evaluating LLM-based text embedding models over MTEB[2], a widely-used Massive Text Embedding Benchmark with diverse tasks and datasets. This integration facilitates comprehensive model development with different methods and extensive assessment across numerous retrieval and embedding tasks in a single framework. ULLME wraps a fine-tuned model into a "`WrappedULLME`" instance, ensuring compatibility with MTEB's requirements for direct eval-

[2] https://github.com/embeddings-benchmark/mteb

uation. In addition to supporting ULLME's fine-tuned models, our evaluation function is designed to perform seamlessly with most LLM models available in the Hugging Face ecosystem, including the latest LLM-Embedding models in the MTEB leaderboard. Users can easily specify the desired model through the "`model_name_or_path`" parameter, enabling effortless evaluation of various LLMs without the need for extensive configuration. ULLME allows users to select specific datasets and language subsets for evaluation. The evaluation results are reported using MTEB's predefined main scores of the corresponding dataset, ensuring standardized and comparable metrics across different models, as demonstrated in Listing 3.

	Phi 1.5	Mistral-2-7B	LlaMa-3-8B
Echo*	36.00	50.26	51.11
LLM2Vec*	54.47	57.47	58.04
Base	31.15	42.31	42.33
Causal + CL	51.83	54.03	54.68
Bi + CL	52.70	55.41	55.86
Bi + CL + SFT	53.88	57.01	56.83
GRL$_{SFT}$	55.01	58.37	57.50
GRL (ours)	**55.76**	**59.50**	**59.27**

Table 2: Model performances on MTEB datasets using MSMARCO for training data. The numbers are averaged over 56 datasets of MTEB, covering diverse tasks such as Retrieval, Reranking, Clustering, Pair Classification, Classification, Semantic Textual Similarity, and Summarization. The best results are in bold and * indicates our implementation/reproduced results using the same training data.

4 ULLME's Experiments

Our ULLME framework supports various LLM architectures and fine-tuning strategies for text embeddings with convenient interface. To highlight the framework's flexibility, we demonstrate the operations of ULLME with three different base LLMs ranging from 1.5B to 8B parameters: Phi-1.5B (Li et al., 2023a), Mistral-7B-Instruct-v0.2 (Jiang et al., 2023b), and Meta-LLama3-8B-Instruct (AI@Meta, 2024). For each LLM, we evaluate ULLME's performance for different combinations of attention and fine-tuning approaches, including: **Base**: Original causal model, **Causal + CL**: Causal model fine-tuned with Contrastive Learning, **Bi + CL**: Bidirectional-enabled model fine-tuned with Contrastive Learning, and **Bi + CL + SFT**: Bidirectional-enabled model fine-tuned with Contrastive Learning and SFT. In addition, we report the performance of our Generation-augmented Representation Learning (GRL) method for fine-tuning LLMs in ULLME, featuring the full model GRL and **GRL**$_{SFT}$, a variant of GRL that replaces DPO with SFT for tuning. Finally, we compare the performance of ULLME's models with recent state-of-the-art methods for LLM-based text embeddings, including **Echo** (Wang et al., 2024c) and **LLM2Vec** (BehnamGhader et al., 2024b).

Settings. Following prior work (Qu et al., 2021; Ren et al., 2021; Ma et al., 2023), we use a curated subset of the MSMARCO dataset (Bajaj et al., 2018a) for model training. MTEB datasets are employed for evaluation. To train the models, we utilize LoRA (Hu et al., 2022) with $r = 16$ and $\alpha = 32$, and enable various optimization techniques, i.e., GradCache, gradient checkpointing, mixed precision training, and FSDP (Zhao et al., 2023a), to minimize GPU memory requirements. We utilize the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 2e-4 and a batch size of 512 with the number of hard negative passages per example was set to 8. We train the models for one epoch on MSMARCO. The weights for the GRL loss components include $\lambda_{CL} = \lambda_{KL} = 1$ and $\lambda_{DPO} = 0.5$. The scaling factor β in the DPO loss was set to 0.1.

Results. Table 2 showcases the performance of various models on the MTEB datasets. Compared to previous methods Echo and LLM2Vec, it is clear that our ULLME framework can be used to train diverse and competitive LLM-based embedding models for different base LLMs and tasks in MTEB. Among various architectures in ULLME, we observe that the combination of contrastive learning and SFT leads to better performance than the individual techniques, demonstrating their complementary benefits for LLM-based embeddings. Notably, our proposed Generation-augmented Representation Learning (GRL) method in ULLME consistently outperforms the best baseline, LLM2Vec, across different base models ranging from 1.5B to 8B parameters. This highlights the effectiveness of using generation probabilities to guide representation learning in GRL. Finally, we note that the inference time of the fine-tuned models with ULLME is comparable to the original LLMs, processing 16K, 12K, and 12.8K tokens per second for Phi-1.5B, Mistral-7B-Instruct-v0.2, and Meta-LLama3-8B-Instruct, respectively.

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

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.

5.1 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 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.

5.2 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

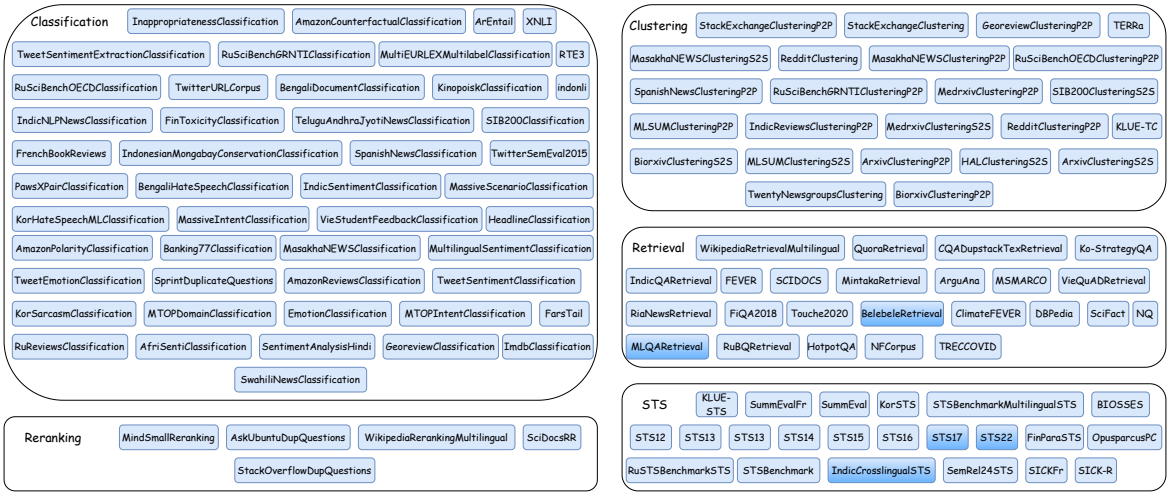


Figure 2: Overview of tasks and datasets in our benchmark. Crosslingual datasets are marked with a blue shade.

a contrastive learning process, effectively teaching the model to distinguish between positive and 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., 2024a; BehnamGhader et al., 2024a; 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.

6 LUSIFER's Experiment

In this section, we first introduce the benchmark datasets and evaluation metrics in Section 6.1. Then, we describe the experimental setup, including the model implementation, training data, and training details in Section 6.2. Afterward, we present the main results in Section 6.3, and analyze the effectiveness of LUSIFER's components in Section 6.6. Finally, we visualize the LUSIFER's representations in multilingual space to obtain insights into its lingual-agnostic capabilities in Section 6.7.

6.1 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., 2024e)	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., 2023b)	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 3: 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., 2023b)	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**

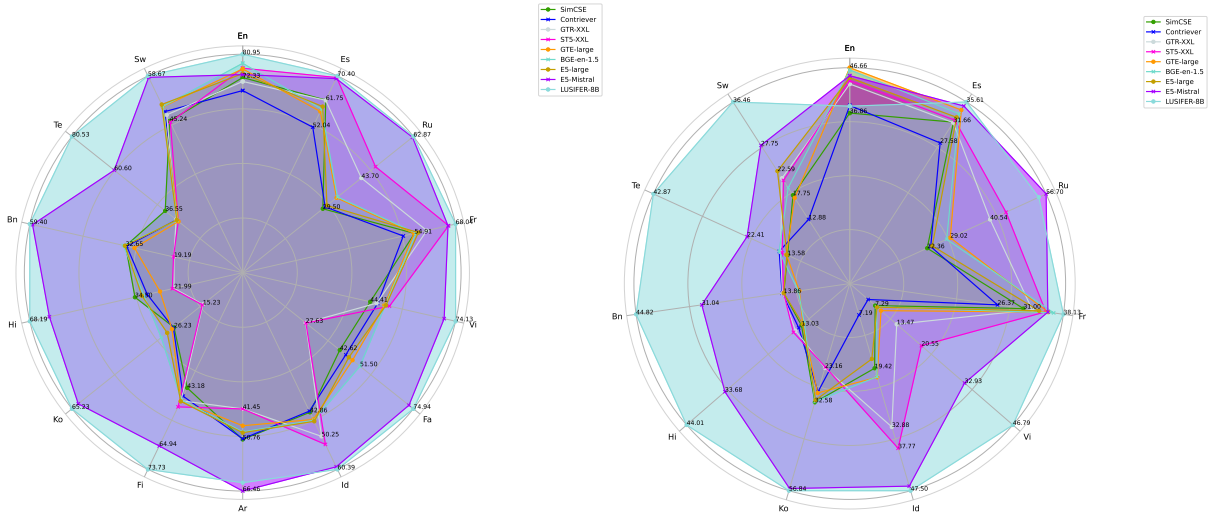
Table 4: Cross-lingual evaluation results. The table presents average metrics for each model over all languages of the datasets, with the highest score for each language emphasized in bold.

6.2 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., 2023a) 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., 2023b) 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).

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., 2018b) for the masked reconstruction and autoregressive completion tasks, respectively. For representation finetuning, we adopt the retrieval datasets as follows: MS MARCO (Bajaj et al., 2018b), 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.

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., 2023b), BGE-en-1.5 (Xiao et al., 2023), E5-large (Wang et al., 2024a); and dense retrieval models with Large Language Model (LLM) backbone: GTR-XXL (Ni et al., 2021b), ST5-XXL (Ni et al., 2021c), E5-Mistral (Wang et al., 2024b). Moreover, we include the follow-



(a) Classification tasks

(b) Clustering tasks

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

ing 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., 2024e), and UDEVER-Bloom-7B (Zhang et al., 2023).

6.3 Main Results

Table 3 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 underrepresented 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., 2024e). These results further demonstrate the benefits of LUCIFER for multilingual representation learning while avoiding expensive multilingual data for text embeddings.

6.4 Cross-Lingual Evaluation

Table 4 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 multiple languages effectively.

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.

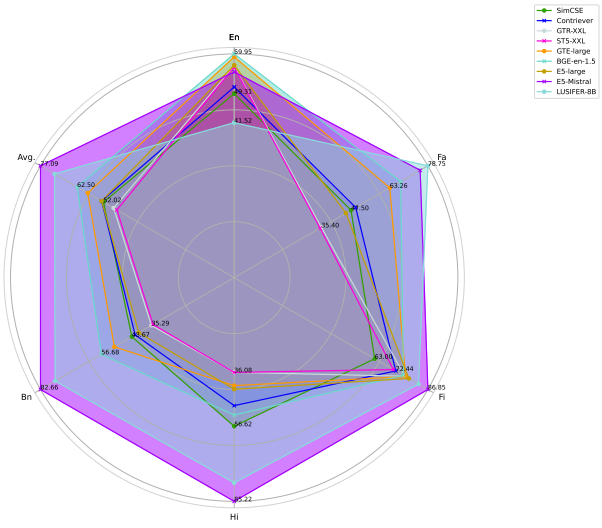


Figure 4: Performance comparison of LUSIFER and baseline models on Reranking tasks.

6.5 Task-Specific Performance

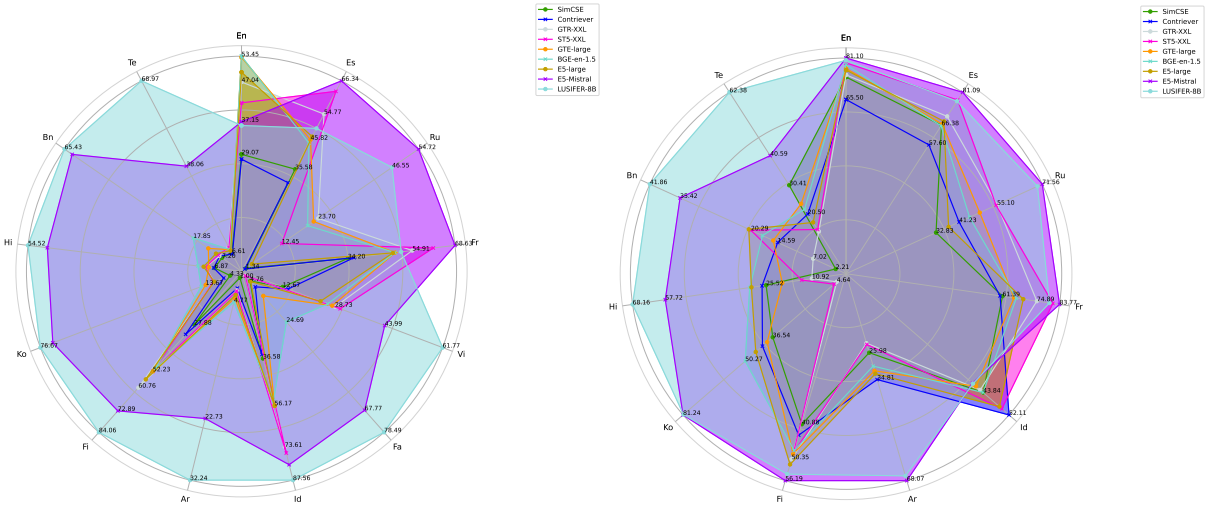
Figure 3, 4, 5 present the performance comparison of LUSIFER and baseline models on Classification, Clustering, Reranking, Retrieval, and STS tasks. LUSIFER consistently outperforms the baseline models across 4 out of 5 tasks, with the largest improvements observed in Clustering and Retrieval tasks, 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.

6.6 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 5 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.

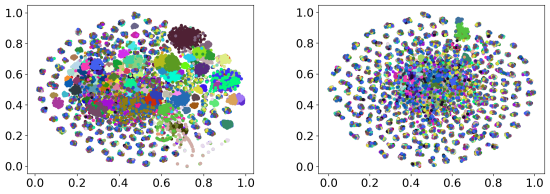
6.7 Model Representation Visualization

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 effectively enable zero-shot multilingual transfer for



(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.

LLM-based embedding methods.

7 Conclusion

This report presents two major contributions toward advancing the quality and versatility of text representations using large language models (LLMs): ULLME, a unified framework for LLM-based embeddings with generation-augmented learning, and LUSIFER, a novel approach for multilingual embedding adaptation without explicit multilingual supervision.

ULLME addresses fundamental challenges in leveraging LLMs for dense retrieval and representation learning, such as the limitations of causal attention and the misalignment between pre-training and retrieval objectives. By enabling bidirectional attention across a broad spectrum of LLM architectures and supporting diverse fine-tuning strategies-including contrastive learning, supervised fine-tuning, and direct preference optimization-ULLME provides a flexible, plug-and-play platform for embedding research and deployment. Its core innovation, Generation-augmented Representation Learning (GRL), enforces consistency between representation-based and generation-based relevance scores, effectively harnessing the generative strengths of LLMs for improved embedding quality. Extensive evaluations demonstrate that ULLME, particularly with GRL, consistently outperforms strong baselines and achieves state-of-the-art results on the Massive Text Embedding Benchmark (MTEB), validating its effectiveness and generality.

LUSIFER tackles the persistent gap in multilingual representation by introducing a zero-shot adaptation framework that aligns a robust multilingual encoder with an English-centric LLM embedding model through a lightweight connector. This design enables effective transfer of language-agnostic semantic knowledge, allowing the resulting model to perform strongly on multilingual and cross-lingual tasks without requiring any explicit multilingual training data. LUSIFER is comprehensively evaluated on a new benchmark covering five primary embedding tasks, 123 datasets, and 14 languages, and demonstrates substantial improvements over existing baselines, especially for medium and low-resource languages. In cross-lingual settings, LUSIFER achieves state-of-the-art performance, highlighting its ability to bridge the gap between high-resource and low-resource language applications efficiently and effectively.

Together, these contributions provide a comprehensive toolkit and methodology for improving both monolingual and multilingual text em-

beddings with LLMs. They lower the barrier for researchers and practitioners to deploy advanced embedding models in diverse real-world scenarios, from retrieval-augmented generation to semantic search and classification. The frameworks introduced here set new benchmarks for the field and open promising directions for future work, including further exploration of alignment strategies, extension to other modalities, and integration with emerging LLM architectures and training paradigms. By bridging architectural, training, and language gaps, this work significantly advances the state of text representation learning and paves the way for more inclusive and effective natural language understanding systems.

Acknowledgments

I would like to express my sincere gratitude to my co-authors, Hieu Man, Nghia Trung Ngo, Viet Dac Lai, Franck Dernoncourt, and my supervisor, Prof. Thien Huu Nguyen, for their invaluable contributions to the research presented in this report. Their expertise, insightful feedback, and dedication were instrumental in shaping both the ULLME and LUSIFER projects.

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.

AI@Meta. 2024. Llama 3 model card.

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. 2018a. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

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. 2018b. 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. 2024a. Llm2vec: Large language models are secretly powerful text encoders. *Preprint*, arXiv:2404.05961.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024b. 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.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

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.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *Preprint*, arXiv:2402.01306.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal

Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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. 2023a. Mistral 7b. *Preprint*, arXiv:2310.06825.

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. 2023b. Mistral 7b. *Preprint*, arXiv:2310.06825.

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.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023a. Textbooks are all you need ii: **phi-1.5** technical report. *arXiv preprint arXiv:2309.05463*.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023b. 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*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023. Fine-tuning llama for multi-stage text retrieval. *Preprint*, arXiv:2310.08319.

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.

Hieu Man, Nghia Trung Ngo, Viet Dac Lai, Ryan A. Rossi, Franck Dernoncourt, and Thien Huu Nguyen. 2025. Lusifer: Language universal space integration for enhanced multilingual embeddings with large language models. *Preprint*, arXiv:2501.00874.

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. 2022. Sgpt: Gpt sentence embeddings for semantic search. *Preprint*, arXiv:2202.08904.

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

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

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847, Online. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

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.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2825–2835, Online and Punta Cana, Dominican Republic. 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.

Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. 2024. Repetition improves language model embeddings. *Preprint*, arXiv:2402.15449.

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.

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. Improving text embeddings with large language models. *Preprint*, arXiv:2401.00368.

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

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024e. 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.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *Preprint*, arXiv:2401.08417.

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. 2023a. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Preprint*, arXiv:2304.11277.

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. 2023b. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Preprint*, arXiv:2304.11277.