AfriMTEB and AfriE5: Benchmarking and Adapting Text Embedding Models for African Languages

Kosei Uemura[1*] **Miaoran Zhang**[3] **David Ifeoluwa Adelani**[2,4]

[1]University of Toronto [2]Mila - Quebec AI Institute, McGill University
[3]Saarland University, Saarland Informatic Campus [4]Canada CIFAR AI Chair
k.uemura@mail.utoronto.ca
mzhang@lsv.uni-saarland.de
david.adelani@mila.quebec

Abstract

Text embeddings are essential building components of many NLP tasks such as retrieval and clustering. Despite the recent release of the Massive Multilingual Text Embedding Benchmark (MMTEB), African languages remain significantly under-represented. In this work, we introduce AfriMTEB, a regional extension of MMTEB covering 59 languages, 14 tasks, and 38 datasets. Unlike MMTEB, where many tasks include few or no African languages, AfriMTEB substantially expands coverage, with tasks spanning between 2 and 56 African languages. To address uneven task–language coverage and enable fair evaluation, we further introduce AfriMTEB-Lite, a balanced subset that uniformly covers nine African languages across all tasks. Complementing the benchmark, we present AfriE5, an adaptation of the instruction-tuned mE5 model to African languages through cross-lingual contrastive learning. Our experimental results show that AfriE5 achieves the strongest overall macro-average among open-weight embedding models on AfriMTEB, with statistically significant gains on several task families, and is competitive with proprietary models such as Gemini Embedding-001.[1]

1 Introduction

Text embeddings are core building blocks for NLP systems in information retrieval, clustering, semantic similarity, and classification (Gao et al., 2022; Feng et al., 2022). However, evaluations on diverse tasks are often limited to a few high resource languages such as English (Muennighoff et al., 2023) or Chinese (Xiao et al., 2024). Many under-represented languages are excluded due to a lack of datasets or non-discoverability of community-created benchmarks (Ojo et al., 2025).

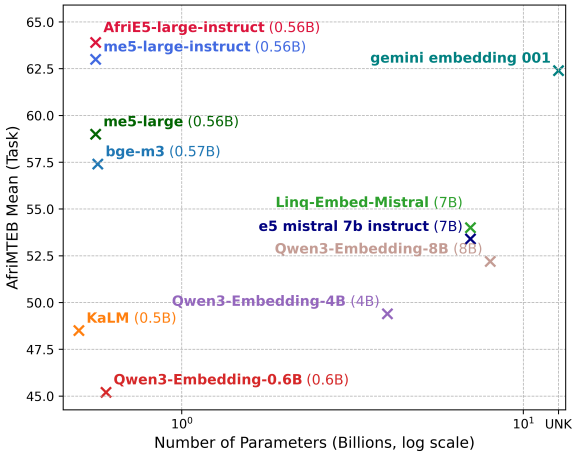

Figure 1: **Model size vs. mean performance on AfriMTEB.** Parameter counts (billions, log scale) are shown on the x-axis, and mean scores across AfriMTEB-Full tasks on the y-axis. *AfriE5-large-instruct* achieves the best overall performance (64.6) despite having far fewer parameters than many models.

While recent text embedding benchmarks have improved language coverage in recent years, such as MMTEB (Enevoldsen et al., 2025), African languages remain under-represented, where many of the tasks covered are either based on massive evaluation of translation datasets (NLLB-Team et al., 2022; Federmann et al., 2022), and the tasks derived or repurposed from translation benchmarks such as Belebele (Bandarkar et al., 2024a) and SIB-200 (Adelani et al., 2024). As a result, the quality of text embeddings for languages in the African region remains unknown, as the region has few standardized tools for comparing models across tasks and languages (Alabi et al., 2025).

In this paper, we introduce **AfriMTEB**, a standardized benchmark designed to evaluate text embeddings for African languages across diverse tasks and application settings. AfriMTEB addresses the lack of task-diverse and systematically comparable evaluations for this region by consolidat-

[*]Work done during internship at Mila
[1]The code is publicly available at https://github.com/LLMforLRL/FlagEmbedding-AfriE5.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3697–3717

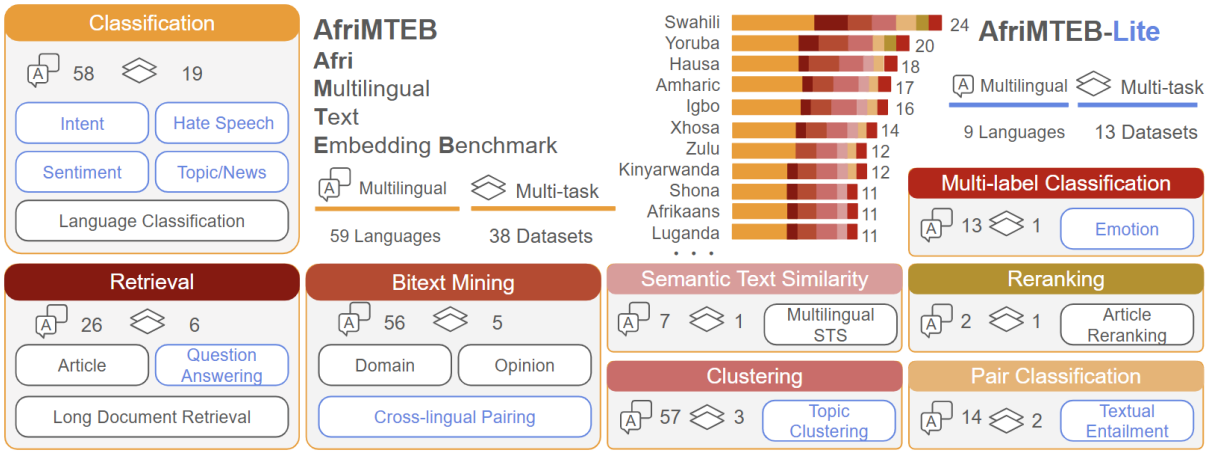

Figure 2: Overview of AfriMTEB and AfriMTEB-Lite. The Full suite spans 59 languages and 38 datasets across 7 families; the Lite suite provides uniform coverage for 9 languages and 13 datasets.

ing and extending existing resources into a unified benchmark. As shown in Figure 2, the *Full* suite (i.e., AfriMTEB-Full) covers 59 languages and 38 datasets spanning bitext mining, classification (single-label, pair, and multi-label), semantic text similarity, retrieval, clustering, and reranking.

To enable controlled, fair, and compute-efficient evaluation, we additionally introduce **AfriMTEB-Lite**, a compact suite over nine geographically and typologically diverse African languages (Amharic, Oromo, Igbo, Yoruba, Hausa, Swahili, Kinyarwanda, Xhosa, and Zulu). AfriMTEB-Lite is constructed by selecting, within each task family, datasets with maximal language overlap, ensuring that almost every task includes all nine languages. This design mitigates the uneven task–language coverage present in large multilingual benchmarks and supports reliable comparison, ablation, and statistical analysis. Crucially, AfriMTEB-Lite enables fine-grained analysis across languages and tasks, allowing us to examine where models succeed or fail by language, resource level, and task family—analyses that are difficult to conduct reliably under the uneven coverage of existing multilingual benchmarks.

In addition to benchmarking, we develop ***AfriE5-Large-Instruct***, adapting strong embedding models to African languages. Starting from the instruction-tuned *mE5-Large-Instruct* (Wang et al., 2024a,c), we leverage cross-lingual contrastive learning with knowledge distillation . We construct a contrastive learning dataset by translating MNLI and SNLI datasets (Williams et al., 2018; Bowman et al., 2015) in African languages with NLLB-200 (3.3B) (NLLB-Team et al., 2022), followed by automatic filtering by SSA-COMET (Li et al., 2025), an African COMET-based quality estimation (QE) metric. Each example is expanded into multiple sources or target directions to encourage cross-lingual alignment. Additionally, we use BGE Reranker v2 m3 (Chen et al., 2024) to extract teacher scores for knowledge distillation.

We evaluate popular text embedding models, such as BGE-M3 (Chen et al., 2024), mE5 (Wang et al., 2024c), Qwen embeddings (Zhang et al., 2025), Gemini embedding-001 (Lee et al., 2025), and Embedding Gemma (Vera et al., 2025). On AfriMTEB-Lite, our adapted model *AfriE5-Large-Instruct*, trained only on nine African languages, achieves an average score of 63.7, surpassing mE5-large-instruct (62.0) and Gemini embedding-001 (63.1). Remarkably, despite being tuned with data and languages aligned to the AfriMTEB-Lite, our model leveraged cross-lingual transfer to generalize to the full AfriMTEB benchmark covering 59 languages and 38 datasets, where it also delivers the best performance with an average score of 62.4, ahead of mE5-large-instruct (61.3) and Gemini embedding-001 (60.6) as shown in Figure 1. These results highlight that targeted cross-lingual adaptation on a carefully selected subset of languages can transfer effectively to a much larger set, yielding consistent improvements across task families while preserving the backbone's general utility.

2 AfriMTEB Benchmark

AfriMTEB is a region-specific benchmark designed to evaluate text embeddings for African languages across a broad range of tasks. While existing mul-

Task	Task Family	Datasets	Number of Languages	
			MMTEB	AfriMTEB
Bitext Mining	Btxt	*Flores, NTREX, BibleNLP, NollySenti, Tatoeba*	59	59
NLI	Pr Clf	*XNLI,* ***AfriXNLI***	1	15
General Topic class.	Clf	*SIB200Classification,* ***SIB200_14Classes***	0	56
News Topic class.	Clf	*MasakhaNEWS, TswanaNews, SiswatiNews, SwahiliNews IsiZuluNews,* ***KinNews***	16	17
Sentiment	Clf	*NaijaSenti, AfriSenti, MultilingualSentiment*	11	12
Hate speech	Clf	***AfriHate***	0	14
LID	Clf	*LanguageClassification, SouthAfricanLangClassification*	0	18
	Clf	*AfriSentiLangClassification*		
Intent	Clf	*MassiveIntent,* ***InjongoIntent***	3	16
Scenario	Clf	*MassiveScenario*	0	3
Emotion	Multi. Clf	***EmotionAnalysisPlus***	0	14
Semantic Relatedness	STS	*SemRel24STS*	7	7
Retrieval	Rtrvl	*Belebele, MIRACL, MIRACLRetrievalHardNegatives, MrTidy, XQuAD, XM3600T2I*	31	31
Clustering	Clust	*SIB200ClusteringFast, MasakhaNEWSClusteringP2P MasakhaNEWSClusteringS2S*	58	58
Reranking	Rrnk	*MIRACLReranking*	0	2

Table 1: **Tasks, task families and datasets included in AfriMTEB-Full**. The task families are Bitext Mining (Btxt), Pair Classification (Pr Clf), Classification (Clf), Semantic Text Similarity (STS), Multi-label Classification (Multi. Clf), Retrieval (Rtrvl), Clustering (Clust) and Reranking (Rrnk). We introduce additional the six datasets highlighted in bold. Each dataset is introduced in Appendix A.2.

tilingual benchmarks such as MMTEB provide valuable global coverage, their African-language representation is sparse and highly uneven. Several task families central to real-world embedding use—such as hate speech detection, intent classification, and multi-label emotion analysis—include *zero* African languages in MMTEB, while core classification benchmarks often cover only one to three African languages (e.g., XNLI). In contrast, AfriMTEB consolidates and extends existing resources into a standardized evaluation suite covering 59 languages, 14 tasks, and 38 datasets, including six newly introduced datasets that fill critical gaps in task and language coverage.

AfriMTEB follows the MTEB (Muennighoff et al., 2023) taxonomy and groups tasks into eight families. Table 1 lists the tasks, task families and the datasets included in the *Full* suite. Each dataset belongs to one of the eight task families: (1) Bitext Mining (Btxt), (2) Pair Classification (Pr Clf), (3) Semantic Text Similarity (STS), (4) Clustering (Clust), (5) Classification (Clf), (6) Multi-label Classification (Multi. Clf), (7) Retrieval (Rtrvl), (8) Reranking (Rrnk). The first four task families perform a task based on a pair of sentences, such as identifying translation pairs, recognizing textual entailment, or measuring semantic relatedness. The next two task families (5–6) focus on classifying a sentence or document into one or more categories. Finally, the last two task families (7–8) are information retrieval tasks, either retrieving relevant items based on a query or re-ranking retrieved results. We provide a full description in Appendix A.1.

2.1 Existing MMTEB Datasets

AfriMTEB builds on the MMTEB by selecting tasks that include African languages. From the original benchmark, we inherit datasets covering bitext mining (e.g., Flores, NTREX, Tatoeba), pair classification (XNLI), topic and sentiment classification (SIB-200, AfriSenti, MasakhaNEWS), semantic textual similarity (SemRel24STS), retrieval (MIRACL, XQuAD, XM3600), clustering (SIB-200, MasakhaNEWS), and reranking (MIRACL).

However, language coverage in these datasets is uneven: some tasks include only a few African languages, while others span broader multilingual settings. To ensure fairness, in the evaluation, we compute macro-averages over languages within each task, then average across tasks and families to obtain the overall AfriMTEB score. This prevents any single task or language from dominating the benchmark.

2.2 New Datasets in AfriMTEB

In AfriMTEB, we introduce six new datasets to broaden task coverage, increase difficulty, and im-

prove language coverage. Details of these additional datasets are provided below.

AfriXNLI An African extension of the XNLI benchmark that provides natural language inference data to 15 African languages (Adelani et al., 2025). By including AfriXNLI, we expand language coverage for the pair classification family beyond a single African language (Swahili), ensuring broader representation of African languages in entailment-style tasks.

EmotionAnalysisPlus A multi-label emotion data set that covers 32 languages, including 14 African languages (Belay et al., 2025; Muhammad et al., 2025b). Each sentence may be assigned multiple emotion labels such as JOY, ANGER, SADNESS, or FEAR. By including this dataset, AfriMTEB introduces the first *multi-label classification* task for African languages, thereby broadening the taxonomy beyond single-label settings.

AfriHate A multilingual hate-speech classification dataset covering 14 African languages (Muhammad et al., 2025a). Each instance is labeled as HATE, ABUSIVE, or NEUTRAL, providing a standardized benchmark for toxic content detection across diverse languages and registers. This dataset extends evaluation to socially relevant safety applications.

InjongoIntent A multilingual dataset for intent detection covering 16 African languages (Yu et al., 2025). It consists of short, conversational utterances annotated with 40 everyday intent categories, for example, requests such as "freeze account" or "play music." By focusing on dialogue-style classification, Injongo complements existing benchmarks like MASSIVE (FitzGerald et al., 2023), but offers broader African language coverage.

KinNews A Kinyarwanda news topic classification dataset with labels covering domains such as politics, business, and sports (Niyongabo et al., 2020). We include this dataset because *MasakhaNEWS* does not cover Kinyarwanda. Adding *KinNews* ensures we can cover Kinyarwanda in the AfriMTEB-Lite, since the Lite version requires maximally overlapping tasks, which is explained in the next section.

SIB200_14Classes A more challenging variant of the SIB-200 dataset, where labels are consolidated into 14 categories. This version still includes 56 African languages but is not limited to them (Adelani et al., 2024). By merging fine-grained topics into broader classes, intra-class diversity increases, which raises task difficulty. The inclusion of this dataset not only strengthens African language evaluation but also increases the difficulty for other covered languages in SIB-200, leading to more robust cross-lingual assessment.

2.3 AfriMTEB-Lite Construction

The *Lite* suite is introduced to address a fundamental limitation of both MMTEB and AfriMTEB-Full: *uneven task–language coverage*. In large multilingual benchmarks, different tasks often include vastly different sets of languages, making macro-averaged comparisons sensitive to missing language–task pairs and inflating variance. AfriMTEB-Lite explicitly resolves this issue by enforcing uniform coverage of the same nine African languages across all tasks, enabling controlled comparisons, tighter confidence intervals, and reproducible ablation studies. The Lite suite focuses on nine geographically and typologically diverse African languages, *Amharic, Oromo, Igbo, Yoruba, Hausa, Swahili, Kinyarwanda, Xhosa, and Zulu*. We retain only those datasets for which all nine languages are available, resulting in a compact yet representative benchmark of 13 datasets spanning classification, retrieval, bitext mining, clustering, pair classification, and multi-label classification. Specifically, it comprises AfriHate, AfriSenti, AfriXNLI, Belebele retrieval, EmotionAnalysisPlus, Flores bitext mining, NTREX bitext mining, InjongoIntent, MasakhaNEWS (for seven languages) and KinNews (for Kinyarwanda) for news classification[2], SIB200Classification, SIB200_14Classes, and SIB200ClusteringFast.

3 Adapting Embedding Models to African Languages

While AfriMTEB provides a standardized and task-diverse evaluation framework for African languages, it also exposes clear performance gaps in existing multilingual embedding models. To demonstrate how such gaps can be addressed in a data-efficient manner, we next present ***AfriE5-Large-Instruct***, an embedding model adapted using cross-lingual contrastive distillation and evaluated systematically under the AfriMTEB.

[2]Zulu is not covered in the news classification.

Model	Btxt	Clf	Clust	Multi. Clf	Pr Clf	Rrnk	Rtrvl	STS	Avg.
Small models (<1B)									
mmBERT-base	3.0	48.5	33.1	24.6	54.1	6.8	4.9	38.0	26.6
KaLM	49.9	36.3	46.8	25.9	60.0	49.3	52.8	53.4	46.8
Qwen3-Embedding 0.6B	33.6	35.9	37.6	25.2	58.0	52.9	52.0	54.1	43.7
bge-m3	70.0	40.0	47.3	26.8	66.4	**66.8**	70.2	58.7	55.8
mE5-large	79.7	43.3	45.3	27.7	64.3	65.2	69.2	62.5	57.2
mE5-large-instruct	**85.8**	49.8	61.9	28.6	63.8	61.9	74.1	64.8	61.3
AfriE5-large-instruct	85.5	49.7	**62.9**	29.8	67.9	64.0	75.4	63.7	**62.4**
Medium models (≈ 4B)									
Qwen3-Embedding-4B	42.5	42.9	39.6	25.8	57.4	60.6	61.5	54.9	48.2
Large models (≥7B)									
gte-Qwen2-7B-instruct	52.7	41.0	56.6	25.1	58.1	58.8	60.3	54.9	50.9
GritLM-7B	45.2	43.4	54.0	26.6	59.6	65.4	61.0	59.1	51.8
Linq-Embed-Mistral	45.2	43.2	55.4	27.1	59.4	65.1	59.2	62.5	52.1
SFR-Embedding-Mistral	46.3	42.5	58.0	26.2	58.9	63.8	58.1	62.0	52.0
E5 mistral 7b instruct	46.4	41.9	58.5	26.0	58.7	64.1	57.3	61.3	51.8
Qwen3-Embedding-8B	48.4	43.7	41.8	27.2	58.3	60.0	69.9	54.7	50.5
Undisclosed size									
gemini embedding 001	72.2	**50.0**	52.7	**32.7**	**71.6**	63.4	**77.5**	**65.0**	60.6

Table 2: **AfriMTEB-Full results.** Average performance of embedding models across languages grouped by task family. The final column reports the unweighted macro-average across task families. Best scores in each column are highlighted in bold. Differences between AfriE5 and mE5 that are statistically significant under paired bootstrap testing are reported in Appendix A.6.

3.1 Method

Our approach combines contrastive learning with knowledge distillation in a unified objective. Given a batch of B queries, each associated with a group of G passages, the total loss is:

$$\mathcal{L} = \mathcal{L}_{contrastive} + \mathcal{L}_{kd}. \quad (1)$$

Contrastive Learning The contrastive learning term is computed as follows:

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(s_{i,pos}/\tau)}{\sum_{j=1}^{B\cdot G}\exp(s_{i,j}/\tau)}. \quad (2)$$

Here $s_{i,j} = \cos(q_i, p_j)$ is the similarity between the query embedding $\mathbf{q}_i$ and passage embedding $\mathbf{p}_j$, and τ is a temperature hyperparameter. The numerator contains the similarity between query $\mathbf{q}_i$ and its corresponding positive passage $\mathbf{p}_{\text{pos}}$. The denominator aggregates similarities between $\mathbf{q}_i$ and all passages in the batch, covering both pre-mined hard negatives (derived from NLI contradiction examples and hard negative mining) and in-batch negatives (passages associated with other queries in the same training batch).

Knowledge Distillation The distillation term aligns the student's predicted distribution with the teacher's scores using cross-entropy:

$$\mathcal{L}_{kd} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{G} P_{teacher}^{(i,j)} \log P_{student}^{(i,j)}, \quad (3)$$

where $P_{teacher}^{(i,j)}$ and $P_{student}^{(i,j)}$ are normalized softmax scores for the query i and the j-th passage, produced by the teacher reranker and the student encoder, respectively.[3] Teacher's scores are obtained during training data construction, where each example is annotated with both positive/negative labels and teacher's score. We detail this data creation process in the next section.

3.2 Cross-lingual Training Data Construction

We constructed cross-lingual training data by leveraging large-scale natural language inference corpora, a supervision signal that has proven effective for learning sentence embeddings (Gao et al., 2022; Reimers and Gurevych, 2019). Specifically, we used MultiNLI and SNLI (Williams et al., 2018; Bowman et al., 2015) as source datasets in English. Each sentence pair was translated into the nine African target languages using NLLB-200 (3.3B) (NLLB-Team et al., 2022). We then estimated translation quality using SSA-COMET-MTL (Li et al., 2025), a COMET variant (Rei et al., 2020) that covers the target African languages, and filtered pairs

[3]We used embedding cosine similarity for the student encoder and logits of the concatenated query and passage for the teacher reranker.

Model	Hate	Senti	NLI	Retrvl	Emo	Btxt		Intent	News	SIB-200			Avg.
						Flores	NTREX			14Classes	Class	Clust	
Small models (< 1B)													
mmBERT-base	48.2	39.9	54.0	4.0	27.3	1.8	2.2	72.5	55.1	2.4	34.2	5.0	28.9
KaLM	48.5	44.3	61.0	51.0	28.9	53.2	58.6	63.4	74.6	7.3	48.9	16.8	46.4
Qwen3-Embedding 0.6B	48.6	41.5	58.1	46.3	28.1	33.0	38.1	68.5	70.5	3.5	40.0	12.0	40.7
EmbeddingGemma 300m	45.4	39.0	53.6	10.6	26.6	12.2	17.9	49.4	59.0	1.4	29.6	4.6	29.1
bge-m3	50.1	47.9	68.7	69.9	29.4	78.1	80.4	75.4	72.7	10.2	55.6	21.0	55.0
mE5-large	49.8	48.9	65.2	69.9	31.3	86.9	88.7	77.1	77.7	11.6	60.4	25.6	57.8
mE5-large-instruct	51.5	47.0	64.5	75.7	31.5	**91.4**	91.5	75.5	78.8	22.0	71.2	43.9	62.0
AfriE5-large-instruct	51.7	50.7	69.0	77.7	32.8	91.2	**92.0**	75.4	**79.5**	**26.2**	**72.0**	**45.7**	**63.7**
Medium models (≈ 4B)													
Qwen3-Embedding-4B	50.0	41.2	56.8	58.3	28.3	47.7	49.0	70.1	76.0	7.9	49.2	17.5	46.0
Large models (≥ 7B)													
gte-Qwen2-7B-instruct	46.0	43.5	58.6	55.6	27.8	58.4	60.0	57.1	79.8	10.4	50.8	22.4	47.5
GritLM-7B	51.4	45.6	59.8	55.0	29.6	46.5	52.0	70.8	75.3	8.8	50.8	22.2	47.3
Linq-Embed-Mistral	51.0	43.8	59.7	52.1	30.0	46.7	52.0	70.3	76.5	7.6	50.7	22.1	46.9
SFR-Embedding-Mistral	49.3	44.5	58.9	50.9	28.8	47.7	53.0	62.5	77.0	8.5	49.7	21.9	46.1
E5 mistral 7b instruct	49.0	45.7	58.7	50.2	28.6	47.9	53.3	61.8	76.0	7.2	49.2	21.8	45.8
Qwen3-Embedding 8B	50.8	46.2	58.8	68.3	29.1	58.3	57.3	67.7	77.6	8.0	53.5	20.9	49.7
Undisclosed size													
gemini embedding 001	**55.0**	**53.8**	**75.3**	**83.6**	**35.5**	88.1	84.2	**83.9**	76.8	17.7	69.2	34.6	63.1

Table 3: **AfriMTEB-Lite results.** Average performance of embedding models across nine African languages (AMH, GAZ, HAU, IBO, KIN, SWA, XHO, YOR, ZUL) on 12 tasks. The final column gives the unweighted macro average across tasks. Best scores per column are highlighted in bold. Differences between AfriE5 and mE5 that are statistically significant under paired bootstrap testing are reported in Appendix A.6.

below a threshold of 0.75 to ensure data quality. We select SSA-COMET because it currently shows the strongest correlation with human judgments for African language pairs among available MT quality estimation metrics.[4]

To encourage cross-lingual alignment, each example was expanded into multiple configurations: (i) premise in the target language and hypothesis in the source, (ii) premise in the source and hypothesis in the target, (iii) both in the target language, and (iv) both in the source language (i.e., English). For MultiNLI, we reformulated examples as query–positive/negative pairs (entailment as *pos*, contradiction as *neg*). For SNLI, we followed the same strategy but included all two-way annotations (positive and negative). We employ hard negative mining using *mE5-Large-Instruct* (Wang et al., 2024c) to identify challenging negative examples. For each query, we encode all corpus passages and perform FAISS-based nearest neighbor search to retrieve the top-k most similar passages. We sample 15 hard negatives from a rank window (ranks 2-200), excluding positives and the query itself. Hard negatives are semantically similar to the query but irrelevant, forcing the model to learn fine-grained discrimination.

[4]We analyze the effect of different quality thresholds on the resulting sample size in Appendix A.4.

3.3 Experimental Settings

We fine-tuned *mE5-Large-Instruct* (Wang et al., 2024c) on our curated cross-lingual training data using the open-sourced `FlagEmbedding`[5] training repository. We followed the default implementation of contrastive learning with cross-device negatives and enabled knowledge distillation with teacher scores provided by BGE Reranker v2 m3 (Chen et al., 2024). All samples in a batch were drawn from the same dataset to maintain consistent supervision. Training was performed on a single GPU and the key hyperparameters are summarized in Table 5. We trained the model for one epoch with logging every 100 steps and checkpoint saving every 100 steps. The resulting model is referred to as ***AfriE5-Large-Instruct***. We provide brief descriptions of the baseline models in Appendix A.3.

4 Results

4.1 AfriMTEB Results

Smaller-sized E5 variants have comparable performance to Gemini Embedding. As shown in Table 2, despite being small models (< 1B), the mE5 family matches or surpasses the proprietary model, gemini embedding 001, especially on bitext mining (i.e., Btxt). *AfriE5-Large-Instruct* attains

[5]https://github.com/FlagOpen/FlagEmbedding

the best macro average at 62.4, edging out *Gemini embedding* (60.6) and *mE5-Large-Instruct* (61.3). This indicates that strong multilingual coverage and targeted adaptation can outweigh model size or access to proprietary training data. Paired bootstrap tests in Appendix A.6 show that the overall macro improvement of *AfriE5-Large-Instruct* over *mE5-Large-Instruct* is statistically significant on AfriMTEB-Full.

AfriE5 and mE5 excel on bitext mining and clustering. On bitext mining, *mE5-Large-Instruct* and *AfriE5-Large-Instruct* have comparable performance of 85 points, both are far ahead of other opens and the API baseline (*Gemini embedding* at 72.2). For clustering, *AfriE5-Large-Instruct* leads with 62.9 and *mE5-Large-Instruct* is next (61.9), while larger models and Gemini embeddings are behind. These families reward models that learn language-agnostic semantic spaces with robust cross-lingual alignment, which appears to be a particular strength of the E5 lineage.

AfriE5 improves reranking with cross-lingual training dataset and knowledge distillation. Although the base model *mE5-Large-Instruct* records a reranking score of 61.9, *AfriE5-Large-Instruct* raises this to 64.0. It also surpasses Gemini embedding 001 (63.4) on the same task (Table 2). This improvement can be attributed to the training recipe, where *AfriE5-Large-Instruct* leverages knowledge distillation from the BGE-M3 cross-encoder in addition to contrastive supervision. The result suggests that incorporating reranker-derived soft labels enhances cross-lingual alignment for ranking-style objectives.

Gemini Embedding excels on classification tasks over E5 variants. While E5 variants lead overall, *Gemini embedding 001* is strongest on most classification-style families: single-label classification (50.0 vs. 49.8 for *mE5-Large-Instruct* and 49.7 for *AfriE5-Large-Instruct*), multi-label classification (32.7 vs. 28.6/29.8), and pair classification (71.6 vs. 63.8/67.9). It also tops retrieval (77.5) and semantic textual similarity (65.0). These strengths suggest Gemini's instruction and data mix particularly benefits discriminative judgment tasks and sentence-level similarity scoring.

Language coverage is more crucial than text embedding model sizes. Large 7B and 8B encoders do not translate into higher AfriMTEB scores. All 7B/8B model scores cluster in the low–mid 50s, for example, *gte-Qwen2-7B-instruct* at 50.9, *GritLM-7B* at 51.8, *Qwen3-Embedding-8B* at 50.5. They are clearly below the smaller E5 variants of around 61.3 point. This gap underscores that broad, balanced language coverage and task diversity matter more than parameter count alone.

AfriE5 generalizes to 59 languages despite training on only 9. Although *AfriE5-Large-Instruct* is adapted using supervision centered on only nine African languages, it achieves the highest macro average (62.4) across 59 languages. Relative to *mE5-Large-Instruct*, *AfriE5-Large-Instruct* shows consistent family-level gains on pair classification (+4.1), reranking (+2.1), retrieval (+1.3), clustering (+1.0), and multi-label classification (+1.2), with small trade-offs on bitext mining (–0.3), single-label classification (–0.1), and STS (–1.1). Importantly, paired bootstrap analysis over task–language cells shows that improvements in pair classification, reranking, retrieval, and multi-label classification are statistically significant, while changes in other families are not (Appendix A.6). This pattern indicates that targeted cross-lingual distillation yields transferable gains beyond the training languages, particularly for retrieval-style and alignment-sensitive tasks.

4.2 AfriMTEB-Lite Results

Here, we present the overall aggregated results by dataset and language in Table 10 and Figure 3. The comprehensive results are in Appendix A.5.

AfriE5 achieves the strongest overall performance on AfriMTEB-Lite. On the Lite suite of nine African languages and twelve tasks, *AfriE5-Large-Instruct* attains the highest overall macro average with a score of 63.7, compared to 62.0 for *mE5-Large-Instruct* and 63.1 for *Gemini embedding* (Table 10). While performance differences vary across task families, paired bootstrap tests over languages show statistically significant improvements for *AfriE5-Large-Instruct* on 9 of the 12 tasks, including AfriXNLI, SIB200 variants, clustering, and retrieval, as well as for the overall macro- and micro-averages (Appendix A.6). In addition, multi-seed experiments confirm that these gains are stable, with standard deviation within ± 0.1 on the Lite macro score (Appendix A.7). These results demonstrate that cross-lingual contrastive distillation improves effectiveness for African languages even when training is restricted to a compact set of nine languages.

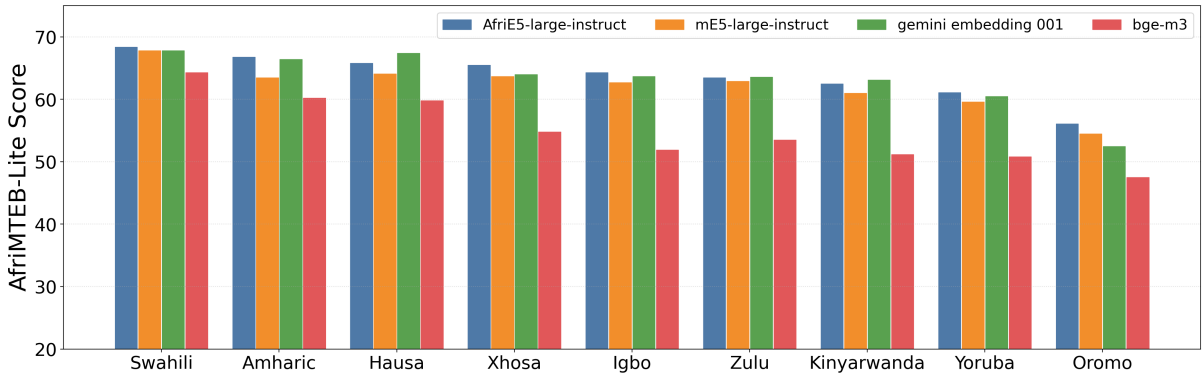

Figure 3: Performance on **AfriMTEB-Lite** across nine target languages. Bars show average scores for four representative embedding models. AfriE5-large-instruct consistently achieves the highest or near-highest scores.

AfriE5 achieves the highest macro averages on 6 languages out of 9. Figure 3 summarizes language-level macro averages across all Lite tasks, with exact values reported in Table 10. AfriE5 achieves the highest average performance on six of the nine languages—Swahili, Amharic, Xhosa, Igbo, Yoruba, and Oromo; consistently outperforming both *mE5-Large-Instruct* and *Gemini embedding 001*. The gains are particularly pronounced for lower-resource languages such as Oromo and Xhosa, where AfriE5 shows clear margins over all baselines.

Our adaptation boosts classification tasks, especially SIB200_14Classes. Relative to *mE5-Large-Instruct*, *AfriE5-Large-Instruct* shows clear improvements on several classification benchmarks. It achieves higher scores on AfriSenti (+3.7), AfriXNLI (+4.5), and SIB200clustering (+1.8). The largest gain appears on SIB200_14Classes (+4.4; 26.2 vs. 21.8), a setting that requires grouping semantically diverse texts into broad topical categories. These improvements indicate that the cross-lingual, NLI-style contrastive learning used in AfriE5 strengthens semantic alignment across languages, leading to more effective cross-lingual transfer for classification and clustering tasks.

5 Ablation

We conduct controlled experiments where all training settings are held fixed (base architecture, loss, knowledge distillation from BGE reranker, batch/group sizes, number of steps, and sampling) while varying a single factor at a time. The results are reported in Table 4.

Impact of cross-lingual dataset expansion As described in Section §3.2, we compare training setups that differ only in whether cross-lingual dataset expansion is applied. Without cross-lingual dataset expansion (×), each NLI example is used within a single language at a time; only target–target sentence pairs (e.g., Swahili–Swahili) are constructed. With cross-lingual dataset expansion (✓), each example is expanded into four configurations: target–target, English–English, target–English, and English–target. Comparing row 3 (×, 0.75) and row 5 (✓, 0.75), the average increases from 62.3 to 63.2. The cross-lingual expansion of the data set exposes the model to richer cross-lingual contrasts: First, higher scores on bitext mining tasks show that the expansion improves the model's ability to align semantically equivalent sentences between languages, producing a more coherent multilingual embedding space. Second, classification tasks such as SIB200-14 benefit from this stronger alignment, as the model learns to abstract over diverse topical domains without relying on language-specific cues. [6]

Machine translation quality threshold (QE) With expansion enabled (rows 1–4), the best overall score is achieved after filtering using a COMET variant, SSA-COMET-MTL QE=0.75 (63.2), compared to 62.5 at 0.67 and a sharp drop to 58.3 at 0.80. A low threshold (0.67) retains large noisy translations (around 433K training samples), which slightly hurts reasoning-heavy tasks. In contrast, a high threshold (0.80) filters out too much data (remaining only 7500 samples), reducing linguistic and semantic diversity and leading to weaker transfer, especially in classification. The middle ground (0.75) balances translation quality with coverage

[6] Belebele retrieval was added to AfriMTEB-Lite after the ablation study had been finalized; therefore, results on Belebele are not reported in Table 4.

						Btxt				SIB-200			
Cross-lingual Expansion	**QE Thres.**	**Hate**	**Senti**	**NLI**	**Emo**	**Flores**	**NTREX**	**Intent**	**News**	**14Classes**	**Class**	**Clust**	**Avg.**
✓	0.67	**52.2**	50.9	66.3	32.3	**91.6**	**94.1**	**77.6**	**82.7**	23.8	71.7	44.8	62.5
✓	0.70	51.3	52.0	69.0	32.6	91.3	92.2	75.2	79.5	26.9	73.1	45.5	62.6
✓	0.75	51.4	**53.3**	**69.1**	**32.8**	91.2	93.8	77.1	82.6	**26.2**	**72.0**	**45.7**	**63.2**
✓	0.80	51.5	49.9	68.2	29.8	88.2	91.0	85.9	79.9	11.2	61.8	24.2	58.3
×	0.75	52.3	52.8	68.2	32.0	88.5	91.2	81.2	82.0	21.8	70.6	45.2	62.3

Table 4: **Ablation study of AfriE5-large-instruct on AfriMTEB-Lite.** We vary the use of cross-lingual dataset expansion and the translation quality threshold during filtering. Best scores per column are highlighted in bold. Note that Belebele retrieval is not covered in this ablation study.

with around 60K samples, yielding the most reliable improvements across tasks. We provide the number of samples retained by SSA-COMET-MTL in Appendix A.4.

6 Related Work

Multilingual text embedding benchmarks The Massive Text Embedding Benchmark (MTEB) introduced a common taxonomy and leaderboard for sentence embeddings (Muennighoff et al., 2023). Since then, several language- or region-specific variants have emerged, including MMTEB (with Europe and Indic tracks), SEA-BED for Southeast Asia, PL-MTEB for Polish, and MTEB-French (Ponwitayarat et al., 2025; Poświata et al., 2024; Ciancone et al., 2024). Additional efforts target specific language families or regions: C-MTEB (Chinese), German-focused suites, JMTEB (Japanese), Korean-focused suites, ruMTEB (Russian), FaMTEB (Persian/Farsi), and VN-MTEB (Vietnamese) (Xiao et al., 2024; Wehrli et al., 2024; Li et al., 2024; Snegirev et al., 2025; Zinvandi et al., 2025; Pham et al., 2025). Our work follows this trend by building an Africa-focused extension with broad task coverage.

Multilingual text embedding models Commercial/API models widely used in practice include OpenAI's text-embedding-3 series, Google's gemini-embedding-001, and Cohere's Embed v3 (Neelakantan et al., 2022; Lee et al., 2025). Open-weight models are an active research area: me5, e5, and e5-mistral-7b-instruct provide strong general-purpose and instruction-tuned baselines (Wang et al., 2024c,a); Qwen3-Embedding and Embedding Gemma offer lightweight multilingual options (Zhang et al., 2025; Vera et al., 2025); GTE proposes efficient, general-purpose embeddings (Li et al., 2023a); and classic multilingual encoders LaBSE remain strong references (Feng et al., 2022). The recent BGE-M3 model integrates multilingual, multi-function training and is a competitive open baseline (Chen et al., 2024). Despite progress, coverage and performance on many African languages remain uneven, motivating region-specific evaluation and targeted adaptation.

7 Conclusion

We presented AfriMTEB, a large-scale benchmark for African languages spanning 59 languages and 14 tasks, and AfriE5, an adaptation of mE5-large-instruct via cross-lingual contrastive distillation. AfriE5 achieves the strongest overall macro-average among open-weight embedding models on both AfriMTEB-Full and AfriMTEB-Lite, with statistically significant gains on several task families, while remaining competitive with proprietary baselines such as Gemini Embedding-001. Our ablations show that cross-lingual dataset expansion and balanced translation filtering are crucial for these gains. Together, AfriMTEB and AfriE5 provide a standardized evaluation framework and a practical reference point for advancing text embedding research for African languages.

Limitations

AfriMTEB expands coverage to 59 languages, yet many African languages, dialects, orthographies, and code-switched registers remain underrepresented. Several datasets inherit noise or heterogeneity from crowd labels, repurposed tasks, and preprocessing. Moreover, our adaptation data relies on machine translation via NLLB-200 and automatic quality estimation using SSA-COMET, which vary in reliability across language pairs and domains. Distillation from a single teacher (BGE Reranker v2 m3) can also transfer its biases, potentially advantaging languages and styles that align with the MT/QE pipeline and the teacher's preferences.

Our evaluation is limited to text-only sentence

and paragraph embeddings and reports macro-averages across tasks and languages; alternative weightings (e.g., by population or application criticality) and additional metrics (calibration, robustness, fairness) could yield different conclusions. End-to-end RAG quality, multimodal retrieval, and very long-context document embeddings are out of scope, and domain coverage skews toward formal/news text over colloquial or specialized domains. True parity with closed-weight baselines (e.g., data mixtures, architectures, inference settings) is infeasible; repeated API evaluations may be affected by nondeterminism, and some open models may be sensitive to prompt formats or poolers we did not exhaustively tune.

While AfriE5 trained on nine languages generalizes well to 59 in our tests, transfer may degrade for typologically distant or extremely low-resource languages; broader community datasets and multi-teacher/multi-signal training are promising avenues to mitigate these limitations.

Acknowledgements

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada. David Adelani acknowledges the funding of IVADO and the Canada First Research Excellence Fund. We acknowledge the use of Gen AI tools for grammar checking, and no scientific content was generated by the tools.

References

ExploreAI Academy and Joanne M. 2022. South african language identification. https://kaggle.com/competitions/south-african-language-identification. Kaggle.

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.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, and et. al. 2023. Masakhanews: News topic classification for african languages. *Preprint*, arXiv:2304.09972.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwuneke, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, and 8 others. 2025. IrokoBench: A new benchmark for African languages in the age of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2732–2757, Albuquerque, New Mexico. Association for Computational Linguistics.

Jesujoba O Alabi, Michael A Hedderich, David Ifeoluwa Adelani, and Dietrich Klakow. 2025. Charting the landscape of african nlp: Mapping progress and shaping the road ahead. *arXiv preprint arXiv:2505.21315*.

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*, page 4623–4637. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024a. 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)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024b. 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.

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025. Evaluating the capabilities of large language models for multi-label emotion understanding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3523–3540, Abu Dhabi, UAE. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *Preprint*, arXiv:1508.05326.

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.

Mathieu Ciancone, Imene Kerboua, Marion Schaeffer, and Wissam Siblini. 2024. Mteb-french: Resources for french sentence embedding evaluation and analysis. *Preprint*, arXiv:2405.20468.

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. *Preprint*, arXiv:1911.02116.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *Preprint*, arXiv:1809.05053.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Veysel Çağatan, and 63 others. 2025. MMTEB: Massive multilingual text embedding benchmark. In *The Thirteenth International Conference on Learning Representations*.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. *Preprint*, arXiv:2007.01852.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2022. Simcse: Simple contrastive learning of sentence embeddings. *Preprint*, arXiv:2104.08821.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Xinshuo Hu, Zifei Shan, Xinping Zhao, Zetian Sun, Zhenyu Liu, Dongfang Li, Shaolin Ye, Xinyuan Wei, Qian Chen, Baotian Hu, and 1 others. 2025. Kalm-embedding: Superior training data brings a stronger embedding model. *arXiv preprint arXiv:2501.01028*.

Amanda Kann. 2024. Massively multilingual token-based typology using the parallel Bible corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11070–11079, Torino, Italia. ELRA and ICCL.

Junseong Kim, Seolhwa Lee, Jihoon Kwon, Sangmo Gu, Yejin Kim, Minkyung Cho, Jy yong Sohn, and Chanyeol Choi. 2024. Linq-embed-mistral:elevating text retrieval with improved gpt data through task-specific control and quality refinement. Linq AI Research Blog.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, Xiaoqi Ren, Shanfeng Zhang, Daniel Salz, Michael Boratko, Jay Han, Blair Chen, Shuo Huang, Vikram Rao, Paul Suganthan, and 28 others. 2025. Gemini embedding: Generalizable embeddings from gemini. *Preprint*, arXiv:2503.07891.

Senyu Li, Jiayi Wang, Felermino D. M. A. Ali, Colin Cherry, Daniel Deutsch, Eleftheria Briakou, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. Ssa-comet: Do llms outperform learned metrics in evaluating mt for under-resourced african languages? *Preprint*, arXiv:2506.04557.

Shengzhe Li, Masaya Ohagi, and Ryokan Ri. 2024. JMTEB: Japanese Massive Text Embedding Benchmark. https://huggingface.co/datasets/sbintuitions/JMTEB.

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

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

Andani Madodonga, Vukosi Marivate, and Matthew Adendorff. 2023. Izindaba-tindzaba: Machine learning news categorisation for long and short text for isizulu and siswati. *Preprint*, arXiv:2306.07426.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning. *Preprint*, arXiv:2509.06888.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfr-embedding-mistral:enhance text retrieval with transfer learning. Salesforce AI Research Blog.

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

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

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Saminu Mohammad Aliyu, Nelson Odhiambo Onyango, Lilian D. A. Wanzare, Samuel Rutunda, Lukman Jibril Aliyu, Esubalew Alemneh, Oumaima Hourrane, Hagos Tesfahun Gebremichael, Elyas Abdi Ismail, Meriem Beloucif, Ebrahim Chekol Jibril, Andiswa Bukula, Rooweither Mabuya, Salomey Osei, and 8 others. 2025a. Afrihate: A multilingual collection of hate speech and abusive language datasets for african languages. *Preprint*, arXiv:2501.08284.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, and 7 others. 2023. Afrisenti: A twitter sentiment analysis benchmark for african languages. *Preprint*, arXiv:2302.08956.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alipio Jeorge, and Pavel Brazdil. 2022. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis. *Preprint*, arXiv:2201.08277.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, and 29 others. 2025b. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *Preprint*, arXiv:2502.11926.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, and 2 others. 2025c. Semeval-2025 task 11: Bridging the gap in text-based emotion detection. *Preprint*, arXiv:2503.07269.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, and 6 others. 2022. Text and code embeddings by contrastive pre-training. *Preprint*, arXiv:2201.10005.

Rubungo Andre Niyongabo, Hong Qu, Julia Kreutzer, and Li Huang. 2020. Kinnews and kirnews: Benchmarking cross-lingual text classification for kinyarwanda and kirundi. *Preprint*, arXiv:2010.12174.

NLLB-Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. Afrobench: How good are large language models on african languages? *Preprint*, arXiv:2311.07978.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Abinew Ali Ayele, Pavan Baswani, Meriem Beloucif, Chris Biemann, Sofia Bourhim, Christine De Kock, Genet Shanko Dekebo, Oumaima Hourrane, Gopichand Kanumolu, Lokesh Madasu, Samuel Rutunda, and 8 others. 2024. Semrel2024: A collection of semantic textual relatedness datasets for 13 languages. *Preprint*, arXiv:2402.08638.

Loc Pham, Tung Luu, Thu Vo, Minh Nguyen, and Viet Hoang. 2025. Vn-mteb: Vietnamese massive text embedding benchmark. *Preprint*, arXiv:2507.21500.

Wuttikorn Ponwitayarat, Raymond Ng, Jann Railey Montalan, Thura Aung, Jian Gang Ngui, Yosephine Susanto, William Tjhi, Panuthep Tasawong, Erik Cambria, Ekapol Chuangsuwanich, Sarana Nutanong, and Peerat Limkonchotiwat. 2025. Seabed: Southeast asia embedding benchmark. *Preprint*, arXiv:2508.12243.

Rafał Poświata, Sławomir Dadas, and Michał Perełkiewicz. 2024. Pl-mteb: Polish massive text embedding benchmark. *Preprint*, arXiv:2405.10138.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference*

on Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

Iyanuoluwa Shode, David Ifeoluwa Adelani, Jing Peng, and Anna Feldman. 2023. Nollysenti: Leveraging transfer learning and machine translation for nigerian movie sentiment classification. *Preprint*, arXiv:2305.10971.

Artem Snegirev, Maria Tikhonova, Anna Maksimova, Alena Fenogenova, and Alexander Abramov. 2025. The russian-focused embedders' exploration: rumteb benchmark and russian embedding model design. *Preprint*, arXiv:2408.12503.

Ashish V. Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. *Preprint*, arXiv:2205.12522.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual mt. *Preprint*, arXiv:2010.06354.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, and et. al. 2025. Embeddinggemma: Powerful and lightweight text representations. *Preprint*, arXiv:2509.20354.

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. *Preprint*, arXiv:2401.00368.

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

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Silvan Wehrli, Bert Arnrich, and Christopher Irrgang. 2024. German text embedding clustering benchmark. *Preprint*, arXiv:2401.02709.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. *Preprint*, arXiv:1704.05426.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. *Preprint*, arXiv:2309.07597.

Hao Yu, Jesujoba O. Alabi, Andiswa Bukula, Jian Yun Zhuang, En-Shiun Annie Lee, Tadesse Kebede Guge, Israel Abebe Azime, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson K. Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, and 3 others. 2025. Injongo: A multicultural intent detection and slot-filling dataset for 16 african languages. *Preprint*, arXiv:2502.09814.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. tydi: A multi-lingual benchmark for dense retrieval. *Preprint*, arXiv:2108.08787.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

Erfan Zinvandi, Morteza Alikhani, Mehran Sarmadi, Zahra Pourbahman, Sepehr Arvin, Reza Kazemi, and Arash Amini. 2025. Famteb: Massive text embedding benchmark in persian language. *Preprint*, arXiv:2502.11571.

A Appendix

A.1 Task Categories

AfriMTEB follows the MTEB taxonomy and groups tasks into eight families. Table 1 lists the families and the datasets included in the *Full* suite.

Bitext Mining. Given two sentence sets from different languages, the task is to identify translation pairs. Embeddings are used to compute similarities and find the best match for each sentence.

Pair Classification. This task involves a pair of input sentences with a binary or categorical relationship label (e.g., entailment vs. contradiction). Predictions are based on embedding similarity.

Classification. Single-text classification where each input is mapped to one label among several categories (e.g., topic, sentiment, hate speech, or language ID). A linear classifier is trained on top of embeddings.

Multi-label Classification. Texts may be assigned multiple labels simultaneously (e.g., emotions). A multi-label classifier is applied on top of embeddings to handle overlapping categories.

Clustering. Given a collection of texts, the task is to group them into clusters that correspond to gold categories. Embeddings are clustered with algorithms such as k-means.

Semantic Text Similarity. This task measures the degree of semantic similarity between sentence pairs, either within or across languages, based on embedding similarity.

Retrieval. Given a query, the task is to retrieve relevant documents from a large corpus. Both queries and documents are embedded, and similarity scores determine ranking.

Reranking. Given a query and a candidate set of documents, the goal is to rank the candidates by relevance using embeddings. This task focuses on fine-grained ranking quality.

A.2 Dataset Descriptions

A.2.1 Bitext Mining datasets

Flores (Goyal et al., 2022). FLORES is a widely used multilingual parallel corpus designed to support evaluation of cross-lingual transfer. It consists of sentence-aligned translations across a large number of languages, including many African languages. In AfriMTEB, FLORES is cast as a bitext mining task: given a sentence in one language, the model must retrieve its correct translation from a pool of candidate sentences in another language using embedding similarity. This task evaluates the ability of embeddings to align semantically equivalent content across languages.

NTREX (Federmann et al., 2022). NTREX is a multilingual parallel dataset derived from professionally translated content, covering a broad set of languages. Similar to FLORES, it is framed as a bitext mining task in AfriMTEB, where embeddings are used to retrieve aligned translations. NTREX complements FLORES by providing different domains and translation characteristics, allowing evaluation of robustness across translation styles.

BibleNLP (Kann, 2024). BibleNLP consists of verse-aligned Bible translations across many languages, including low-resource African languages. Although the domain is religious text, the strict verse alignment makes it well suited for evaluating cross-lingual semantic alignment. In AfriMTEB, the task is to retrieve the correct verse translation given a source verse, testing whether embeddings capture meaning consistently across languages despite domain specificity.

NollySenti (Shode et al., 2023). NollySenti originates as a sentiment analysis dataset for Nigerian languages but is additionally repurposed in AfriMTEB as a bitext-style mining task by leveraging aligned or parallel content. This dataset evaluates whether embeddings trained primarily for semantic similarity can also recover aligned text across languages in a more informal, social-media-driven domain.

Tatoeba (Tiedemann, 2020). The Tatoeba dataset contains sentence-aligned translations created by a community of contributors and covers a wide range of languages. In AfriMTEB, Tatoeba is used for bitext mining, where models must identify translation pairs based on embedding similarity. The dataset is noisier than professionally curated corpora.

A.2.2 Pair Classification datasets

XNLI (Conneau et al., 2018). XNLI is a cross-lingual natural language inference benchmark derived from MultiNLI. Each example consists of a premise–hypothesis sentence pair labeled as entailment, contradiction, or neutral. In AfriMTEB, XNLI is treated as a pair classification task, where embeddings of the two sentences are combined to predict the relation label, evaluating whether embeddings preserve fine-grained semantic relations.

AfriXNLI (Adelani et al., 2025). AfriXNLI extends the XNLI framework to additional African languages. It follows the same premise–hypothesis structure and label space as XNLI, but focuses on languages that are largely absent from existing NLI benchmarks. This dataset allows AfriMTEB to evaluate whether multilingual embeddings generalize NLI-style reasoning to African languages.

A.2.3 Classification datasets

SIB200Classification (Adelani et al., 2024). SIB200Classification is derived from the SIB-200 benchmark and formulates topic classification with a reduced set of coarse-grained categories. Each document is assigned exactly one topic label. In AfriMTEB, it is used to evaluate general topic classification across a large number of languages with balanced label distributions.

SIB200_14Classes (Adelani et al., 2024). SIB200_14Classes is a more fine-grained variant of topic classification from SIB-200, where texts are assigned to one of fourteen topic categories. This dataset is more challenging due to the larger label space and semantic overlap between topics. It is included in AfriMTEB to assess how well embeddings separate closely related topical categories across languages.

MasakhaNEWS (Adelani et al., 2023). MasakhaNEWS is a multilingual African news classification benchmark covering several African languages. Each news article is labeled with a topic such as politics, sports, or business. In AfriMTEB, it serves as a core news topic classification dataset and reflects realistic, domain-specific text encountered in African media.

TswanaNews, SiswatiNews (Madodonga et al., 2023), SwahiliNews, IsiZuluNews (Madodonga et al., 2023), KinNews (Niyongabo et al., 2020). These datasets are language-specific news classification benchmarks for Setswana, SiSwati, Swahili, isiZulu, and Kinyarwanda respectively. Each dataset follows the same formulation as MasakhaNEWS, with articles labeled by topic. KinNews is newly incorporated in AfriMTEB to extend news classification coverage to Kinyarwanda, ensuring broader geographic representation.

NaijaSenti (Muhammad et al., 2022). NaijaSenti is a sentiment analysis dataset for Nigerian languages, primarily focused on informal and social-media-style text. Each example is labeled with sentiment polarity (e.g., positive, negative, neutral). In AfriMTEB, it evaluates whether embeddings capture affective meaning in informal language varieties.

AfriSenti (Muhammad et al., 2023). AfriSenti is a multilingual African sentiment dataset covering multiple languages and domains. Compared to NaijaSenti, it includes a broader linguistic scope and more diverse text sources. It is used to assess cross-lingual sentiment classification performance.

MultilingualSentiment (Muennighoff et al., 2023). This dataset is a general multilingual sentiment benchmark included to provide additional coverage beyond African-specific datasets. It allows comparison between African-language performance and more widely studied multilingual sentiment settings.

AfriHate (Muhammad et al., 2025a). AfriHate is a hate speech and offensive language classification dataset designed specifically for African languages. Each text is labeled according to whether it contains hate or abusive content. Its inclusion addresses a gap in prior benchmarks, where African languages were largely absent from hate speech evaluation.

LanguageClassification, SouthAfricanLangClassification (Academy and M, 2022), AfriSentiLangClassification (Muhammad et al., 2023). These datasets are used for language identification (LID). The task is to predict the language of a given text. SouthAfricanLangClassification focuses on closely related South African languages, while AfriSentiLangClassification is derived from sentiment data and tests LID under informal text conditions.

MassiveIntent (Ousidhoum et al., 2024). MassiveIntent comes from the MASSIVE dataset and consists of short user utterances labeled with intent categories (e.g., request, command). It evaluates whether embeddings encode intent-level semantics across languages.

InjongoIntent (Yu et al., 2025). InjongoIntent is an intent classification dataset targeting African languages. It mirrors the structure of MassiveIntent but focuses on underrepresented languages and locally relevant intents, providing a more realistic evaluation for African conversational systems.

MassiveScenario (Ousidhoum et al., 2024). MassiveScenario is another subset of the MASSIVE dataset, where utterances are labeled by scenario or domain (e.g., travel, finance). In AfriMTEB, it is treated as a standard single-label classification task.

A.2.4 Multi-label Classification datasets

EmotionAnalysisPlus (Muhammad et al., 2025c). EmotionAnalysisPlus is a multi-label emotion clas-

sification dataset where each text may express multiple emotions simultaneously (e.g., joy, anger, sadness). Unlike sentiment analysis, this task requires modeling overlapping affective states. It is included in AfriMTEB to evaluate multi-label prediction capabilities in African languages.

A.2.5 Semantic Text Similarity datasets

SemRel24STS (Ousidhoum et al., 2024). SemRel24STS is a semantic textual similarity dataset where sentence pairs are annotated with graded similarity scores. Models are evaluated by correlating embedding-based similarity with human judgments. This dataset tests fine-grained semantic sensitivity rather than categorical decisions.

A.2.6 Retrieval datasets

Belebele (Bandarkar et al., 2024b). Belebele is a multilingual retrieval benchmark derived from reading comprehension data. Given a query (often a question), the task is to retrieve the most relevant passage from a set of candidates. It evaluates cross-lingual retrieval performance under realistic QA-style conditions.

MIRACL and MIRACLRetrievalHardNegatives (Zhang et al., 2023). MIRACL is a large-scale multilingual information retrieval benchmark covering many languages. Queries are paired with large document collections, and models must retrieve relevant passages. The Hard Negatives variant augments the task with challenging non-relevant passages, making ranking more difficult and discriminative.

MrTidy (Zhang et al., 2021), XQuAD (Artetxe et al., 2020), XM3600T2I (Thapliyal et al., 2022). These datasets extend retrieval evaluation to different domains and modalities. MrTidy and XQuAD focus on text-based retrieval, while XM3600T2I evaluates cross-lingual text-to-image retrieval. Together, they broaden the retrieval evaluation beyond standard document search.

A.2.7 Clustering datasets

SIB200ClusteringFast (Adelani et al., 2024). This dataset is derived from SIB-200 and evaluates topic clustering. Texts must be grouped into clusters corresponding to gold topic labels, using only embedding similarity and clustering algorithms such as k-means.

MasakhaNEWSClusteringP2P and MasakhaNEWSClusteringS2S (Adelani et al., 2023). These datasets formulate clustering tasks from MasakhaNEWS articles using different clustering protocols. They evaluate whether embeddings induce meaningful topical structure in African news text.

A.2.8 Reranking datasets

MIRACLReranking (Zhang et al., 2023). MIRACLReranking is a fine-grained ranking task built on MIRACL. Given a query and a small set of candidate passages, the goal is to rerank them by relevance. This task focuses on precise ordering rather than coarse retrieval, testing the discriminative power of embeddings.

Parameter	Value
Training epochs	1
Batch size (per device)	8
Group size	8
Learning rate	1e-5
Warmup ratio	0.1
Max query length	512
Max passage length	512
Padding multiple	8
Knowledge distillation	True (KL divergence)
Negatives cross-device	Enabled
Embedding normalization	True (L2)
Sentence pooling	Mean pooling
Temperature	0.02
Precision	FP16

Table 5: Key training configurations used in fine-tuning.

A.3 Baseline Model Descriptions

mmBERT-base (Marone et al., 2025) is a modern multilingual encoder trained on more than 3T tokens covering >1,800 languages, extending the ModernBERT (Warner et al., 2025) (fast encoder with long context) to the multilingual regime. The training recipe introduces curriculum-style annealed language learning, inverse masking, and temperature-based sampling to emphasize low-resource languages while retaining strong performance on high-resource ones. Reported results show that mmBERT-base surpasses prior multilingual encoders such as XLM-R on standard NLU and retrieval benchmarks, approaching the English-only ModernBERT on GLUE despite being trained predominantly on non-English data. Inference follows standard encoder usage (mean pooling over last hidden states and L2 normalization).

KaLM-Embedding (Hu et al., 2025) is a multilingual embedding family that prioritizes training-data quality over sheer scale, combining (i) persona-based synthetic examples distilled from LLMs,

(ii) ranking-consistency filtering, and (iii) semi-homogeneous task batching for efficient contrastive learning. Many public checkpoints are built on compact Qwen2 backbones (e.g., ∼0.5B) and instruction-tuned variants for downstream retrieval and semantic similarity. Technical reports describe v1.5/v2 updates with improved data curation and training strategy, yielding strong MTEB performance for their size.

Qwen3-Embedding (0.6B / 4B / 8B) (Zhang et al., 2025) is a purpose-built series of dense encoders for text embeddings and reranking, offered in 0.6B/4B/8B sizes with a 32k token context window and coverage of 100+ languages. The models are instruction-aware and support flexible output dimensionalities (via MRL-style prefix truncation), with typical maximum embedding sizes of 1,024 (0.6B), 2,560 (4B), and 4,096 (8B); matching reranker models are available at each size. Training leverages Qwen3 LLMs both as backbones and as data synthesizers across domains and languages, improving robustness for retrieval and reranking workloads. Inference uses mean pooling and L2 normalization; common deployment stacks expose a user-selectable output dimensionality for storage/latency trade-offs.

BGE-M3 (Chen et al., 2024) is a versatile embedding model unifying three capabilities in a single encoder: Multi-Functionality (dense, multi-vector, and sparse retrieval), Multi-Linguality (100+ languages), and Multi-Granularity (robust from short queries to long documents, up to ∼8,192 tokens). The training pipeline uses self-knowledge distillation across retrieval functions to align representations and enables hybrid retrieval without switching models. It is widely used as a strong multilingual baseline for retrieval, clustering, and classification.

mE5-Large and **mE5-Large-Instruct** (Wang et al., 2024c) are multilingual members of the E5 family built on XLM-RoBERTa-large (Conneau et al., 2020), trained with a two-stage recipe that first performs weakly supervised contrastive pre-training on roughly one billion multilingual text pairs and then supervised fine-tuning on curated embedding tasks; the instruction variant further formats supervision with concise task instructions to specialize representations for retrieval and related tasks. Both use a 24-layer encoder that produces 1,024-dimensional vectors and inherit broad (100 language) coverage from the XLM-RoBERTa backbone. At inference time they follow the E5 prompt conventions (e.g., query / passage style inputs), apply mean pooling over the last hidden states, and L2-normalize the output, yielding strong performance on retrieval, semantic similarity, clustering, and classification benchmarks in both monolingual and cross-lingual settings.

gte-Qwen2-7B-instruct (Li et al., 2023b) is a 7B-parameter instruction-tuned General Text Embedding model built on Qwen2-7B, targeting high-quality multilingual embeddings with a 32k context window. At release, it reported leading scores on MTEB English/Chinese subsets, reflecting a training mixture that combines instruction-style contrastive objectives and curated negatives. It is used as a strong large-model baseline for retrieval, reranking, and semantic similarity.

GritLM-7B (Muennighoff et al., 2024) unifies generation and embeddings in one model via Generative Representational Instruction Tuning (GRIT) . A custom modeling component adds bidirectional attention paths so that the same backbone can function as a strong encoder for embeddings without sacrificing generative performance (Muennighoff et al., 2024). The paper reports SOTA-level MTEB results for 7B-class open models alongside strong generative benchmarks, demonstrating that a single model can excel at both modalities (Muennighoff et al., 2024).

Linq-Embed-Mistral (Kim et al., 2024) is a Mistral-7B–based embedding model developed with task-tailored data crafting, filtering, and hard-negative mining to improve retrieval quality. The technical report and model card document leading MTEB retrieval scores at release (e.g., average ∼68.2 on 56 datasets; retrieval score ∼60.2), achieved through homogeneous task ordering and mixed-task fine-tuning strategies. It is frequently used as a competitive 7B encoder baseline for dense retrieval.

SFR-Embedding-Mistral (Meng et al., 2024) applies transfer learning on top of E5-mistral-7b-instruct and Mistral-7B-v0.1, with additional multi-task training and optimized negative sampling aimed at retrieval tasks. Public materials position it as a top-performing 7B embedding model for search, clustering, and classification workloads.

E5 Mistral 7B Instruct (Wang et al., 2024b) initializes the E5 instruction-tuning recipe from Mistral-

7B-v0.1, producing a 7B encoder specialized for instruction-aware embeddings . As with other E5 variants, inference benefits from task instructions plus "`query:`/`passage:`" prefixes, and mean pooling with L2 normalization is used for final vectors. It serves both as a competitive baseline and as a foundation for further transfer learning (e.g., SFR-Embedding-Mistral).

Gemini Embedding-001 (Lee et al., 2025) is Google's multilingual text-embedding model available via the Gemini API and Vertex AI, trained with Matryoshka Representation Learning (MRL) so that leading vector prefixes remain useful at smaller dimensions . The default output is 3,072 dimensions, but APIs allow setting output dimensionality (e.g., 1,536 or 768) with minimal quality loss, enabling flexible storage/latency trade-offs. The model supports 100+ languages and has been a strong performer on multilingual MTEB since its early releases.

A.4 Detailed Statistics of Machine Translation Quality

To assess the quality of the machine translation data used in AfriMTEB, we rely on automatic evaluation with **SSA-COMET** (Li et al., 2025). SSA-COMET is a recently released metric trained on SSA-MTE, a large-scale human-annotated evaluation dataset covering 13 African language pairs with over 63,000 sentence-level judgments. Compared to earlier African-focused metrics such as AfriCOMET, SSA-COMET provides stronger correlation with human ratings and better robustness in low-resource settings.

Language	0.67	0.75	0.80
Amharic (amh_Ethi)	44,166	4,617	281
Oromo (gaz_Latn)	73,846	14,598	2,818
Hausa (hau_Latn)	31,799	5,851	1,056
Igbo (ibo_Latn)	16,778	1,279	136
Kinyarwanda (kin_Latn)	81,003	4,867	337
Swahili (swh_Latn)	116,338	21,996	1,849
Xhosa (xho_Latn)	18,143	2,007	351
Yoruba (yor_Latn)	31,377	3,316	411
Zulu (zul_Latn)	20,229	1,535	224
Total	**433,629**	**60,066**	**7,463**

Table 6: Number of translation pairs retained after filtering with SSA-COMET at three thresholds. Lower thresholds yield more data, while stricter thresholds retain fewer but higher-quality examples.

From Table 6, we observe a clear trade-off between dataset size and quality. At a relaxed threshold of 0.67, over 430k sentence pairs are retained, ensuring broad coverage across all nine target languages. Increasing the threshold to 0.75 reduces the pool to around 60k examples, striking a balance between filtering noise and preserving sufficient training data. At the strictest cutoff of 0.80, only 7.5k pairs remain, indicating that high-quality translations are relatively scarce. Language-level differences are also evident: Swahili, Oromo, and Kinyarwanda consistently contribute the largest number of pairs, while Igbo and Amharic see the sharpest reductions under stricter filtering, reflecting variation in MT quality across languages.

Language	0.67	0.75	0.80
Amharic (amh_Ethi)	44,166	4,617	281
Oromo (gaz_Latn)	73,846	14,598	2,818
Hausa (hau_Latn)	31,799	5,851	1,056
Igbo (ibo_Latn)	16,778	1,279	136
Kinyarwanda (kin_Latn)	81,003	4,867	337
Swahili (swh_Latn)	116,338	21,996	1,849
Xhosa (xho_Latn)	18,143	2,007	351
Yoruba (yor_Latn)	31,377	3,316	411
Zulu (zul_Latn)	20,229	1,535	224
Kongo (kon_Latn)	–	18,030	–
Kabuverdiano (kea_Latn)	–	810	–
Somali (som_Latn)	–	4,203	–
Twi (twi_Latn)	–	6,090	–
Sotho (sot_Latn)	–	20,190	–
Tsonga (tso_Latn)	–	43,371	–
Swati (ssw_Latn)	–	4,062	–
Lingala (lin_Latn)	–	20,091	–
Shona (sna_Latn)	–	23,025	–
Northern Sotho (nso_Latn)	–	33,912	–
Plateau Malagasy (plt_Latn)	–	11,736	–
Tswana (tsn_Latn)	–	35,481	–
Afrikaans (afr_Latn)	–	19,872	–
Nyanja (nya_Latn)	–	53,367	–
Egyptian Arabic (arz_Arab)	–	6,960	–
Total	**433,629**	**405631**	**7,463**

Table 7: Updated counts of translation pairs retained after filtering with SSA-COMET at three thresholds.

A.5 Detailed AfriMTEB-Lite Results

Table 11 shows the comprehensive results across all evaluated datasets and languages in AfriMTEB-Lite.

A.6 Statistical Significance

To quantify the robustness of the observed improvements of AfriE5 over mE5, we conduct a paired bootstrap significance analysis on both AfriMTEB-Lite and AfriMTEB-Full.

For AfriMTEB-Lite, we treat each language within a task as one observation and compute paired differences Δ = AfriE5 − mE5 per task–

Table 8: Statistical Significance on AfriMTEB-Lite Tasks. Mean performance difference Δ = AfriE5 − mE5 (in %), 95% confidence intervals (CI), and one-sided p-values ($H_0 : \Delta \leq 0$) estimated via paired bootstrap over languages.

Task	Δ	95% CI	p-value	n_{langs}
AfriHate Classification	+0.17	$[-1.17, +1.38]$	0.390	9
AfriSenti Classification	+3.74	$[+2.08, +5.40]$	$\mathbf{< 0.001}$	7
News Classification	+0.64	$[-0.06, +1.30]$	**0.035**	8
AfriXNLI	+4.50	$[+3.82, +5.16]$	$\mathbf{< 0.001}$	9
Emotion Analysis	+1.28	$[+0.30, +2.25]$	**0.004**	9
Flores Bitext Mining	-0.17	$[-0.26, -0.06]$	0.998	9
Injongo Intent	-0.04	$[-1.41, +1.45]$	0.540	9
NTREX Bitext Mining	+0.53	$[-0.14, +1.54]$	0.109	9
SIB-200 (14 Classes)	+4.16	$[+2.58, +5.75]$	$\mathbf{< 0.001}$	9
SIB-200 Classification	+0.84	$[+0.17, +1.59]$	**0.004**	9
SIB-200 Clustering	+1.76	$[+0.77, +2.91]$	$\mathbf{< 0.001}$	9
Belebele Retrieval	+2.00	$[+1.59, +2.41]$	$\mathbf{< 0.001}$	9
Overall (Macro)	+1.62	$[+0.79, +2.57]$	$\mathbf{< 0.001}$	12
Overall (Micro)	+1.59	$[+1.17, +2.02]$	$\mathbf{< 0.001}$	105

language cell. We then perform 10,000 paired bootstrap resamples over languages and report (i) the mean difference Δ (in absolute points), (ii) 95% confidence intervals obtained via the percentile method, and (iii) one-sided p-values for the null hypothesis $H_0 : \Delta \leq 0$. The detailed results are shown in Table 8.

On AfriMTEB-Lite, 9 out of 12 tasks show statistically significant improvements at $p < 0.05$, and both macro and micro averages are significant. The largest gains appear on AfriXNLI, SIB-200 (14 classes), SIB-200 clustering, AfriSenti, and Belebele retrieval, while Flores bitext mining shows a small regression.

For AfriMTEB-Full, we group task–language cells into the eight categories reported in the main paper (bitext mining, classification, clustering, multilabel classification, pair classification, reranking, retrieval, and STS). Within each category we again compute paired differences per cell and apply the same 10,000-resample paired bootstrap procedure. Table 9 reports the mean difference, 95% confidence intervals, one-sided p-values, and the number of cells per category.

AfriE5 significantly outperforms mE5 in four out of eight categories (multilabel classification, pair classification, reranking, retrieval), while bitext mining and STS show small, non-significant regressions. The overall macro average is significantly positive, whereas the micro average is not, largely because the large classification category shows only small, non-significant gains.

Table 9: Statistical Significance by Category (AfriMTEB-Full). Mean performance difference Δ = AfriE5 − mE5 averaged across all tasks/languages within each category. p-values are computed via paired bootstrap.

Category	Δ	95% CI	p-value	n_{cells}
Bitext Mining	-0.17	$[-0.36, +0.04]$	0.949	95
Classification	-0.16	$[-1.09, +0.67]$	0.622	194
Clustering	+1.02	$[-1.87, +3.76]$	0.235	26
Multilabel Classification	+1.22	$[+0.56, +1.88]$	$\mathbf{< 0.001}$	14
Pair Classification	+4.19	$[+3.47, +4.78]$	$\mathbf{< 0.001}$	17
Reranking	+2.13	$[+1.58, +2.66]$	$\mathbf{< 0.001}$	2
Retrieval	+1.37	$[+0.95, +1.75]$	$\mathbf{< 0.001}$	36
STS	-1.04	$[-3.02, +0.98]$	0.865	7
Overall (Macro)	+1.07	$[+0.08, +2.15]$	**0.015**	8
Overall (Micro)	+0.29	$[-0.22, +0.76]$	0.122	391

A.7 Multiple Seeds Training

To assess the stability of AfriE5 under our single-GPU training recipe, we additionally train three instances of AfriE5 with different random seeds (42, 123, 456), keeping all other hyperparameters and data fixed. Table 10 reports AfriMTEB-Lite results for mE5, the primary AfriE5 run reported in the main paper, and the three additional seed runs.

Across the three additional seeds, the overall AfriMTEB-Lite macro average varies only between 63.5 and 63.6 (vs. 63.7 for the main AfriE5 run). Per-task scores are also highly consistent, with differences typically well below one absolute point and no systematic reversals of the performance pattern relative to mE5. This confirms that our adaptation procedure is stable with respect to random initialization and data ordering, even under a modest compute budget (single A100, one epoch over the filtered training set).

						Btxt				SIB-200			
Model	**Hate**	**Senti**	**NLI**	**Retrvl**	**Emo**	**Flores**	**NTREX**	**Intent**	**News**	**14Classes**	**Class**	**Clust**	**Avg.**
mE5-large-instruct	51.5	47.0	64.5	75.7	31.5	**91.4**	91.5	75.5	78.8	22.0	71.2	43.9	62.0
AfriE5-large-instruct	**51.7**	50.7	69.0	77.7	**32.8**	91.2	**92.0**	75.4	79.5	**26.2**	72.0	45.7	**63.7**
seed42	51.5	50.4	68.5	**77.8**	32.7	91.3	91.9	75.1	**79.7**	26.1	**72.1**	**45.9**	63.6
seed123	51.4	**51.0**	**69.0**	77.6	32.7	90.8	91.5	75.4	79.5	26.0	71.9	45.7	63.5
seed456	51.5	50.4	68.6	77.7	32.7	**91.4**	92.0	**75.6**	79.5	25.5	71.9	45.6	63.5

Table 10: **AfriMTEB-Lite results.** Average performance of embedding models across nine African languages (AMH, GAZ, HAU, IBO, KIN, SWA, XHO, YOR, ZUL) on 12 tasks. Columns report task-level averages, and the final column gives the unweighted macro average across tasks. Best scores per column are highlighted in bold.

Dataset	amh	gaz	hau	ibo	kin	swa	xho	yor	zul	Avg.
AfriHateClassification										
bge-m3	52.52	52.19	47.32	56.03	51.71	64.81	36.85	50.77	38.49	50.08
gemini embedding 001	54.31	52.31	52.4	63.47	57.75	74.07	38.94	56.3	45.3	54.98
mE5-large-instruct	50.78	51.92	42.95	57.16	54.85	67.12	40.18	51.56	47.03	51.51
AfriE5-large-instruct	52.76	53.79	44.77	58.95	56.47	64.94	37.8	51.67	43.93	51.68
AfriSentiClassification										
bge-m3	48.45	34.8	68.83	54.43	46.12	44.52	–	38.21	–	47.91
gemini embedding 001	59.75	34.35	75.02	61.65	58.58	45.94	–	41.25	–	53.79
mE5-large-instruct	44.07	35.8	68.64	49.52	50.19	41.95	–	38.91	–	47.01
AfriE5-large-instruct	51.41	35.7	71.78	53.67	53.98	44.13	–	44.57	–	50.75
NewsClassification										
bge-m3	84.73	74.62	78.23	64.23	49.4	73.8	77.85	79.05	–	72.74
gemini embedding 001	84.41	82.09	77.14	68.9	58.58	71.95	87.31	84.21	–	76.82
mE5-large-instruct	87.82	79.54	80.71	77.59	57.27	79.43	85.29	83.07	–	78.84
AfriE5-large-instruct	89.52	81.75	81.52	78.03	58.31	78.15	85.59	82.99	–	79.48
AfriXNLI										
bge-m3	75.64	66.39	69.67	64.33	65.27	73.88	69.41	64.03	69.87	68.72
gemini embedding 001	81.32	66.8	78.57	78.05	73.79	78.34	75.42	71.72	73.38	75.27
mE5-large-instruct	66.85	62.96	63.51	65.05	62.01	68.09	66.66	61.89	63.7	64.52
AfriE5-large-instruct	72.28	68.57	67.89	70.36	64.68	72.1	72.41	65.43	67.52	69.03
EmotionAnalysisPlus										
bge-m3	20.99	28.54	24.92	23.49	33.9	37.14	21.52	40.58	33.86	29.44
gemini embedding 001	28.38	30.8	36.03	30.71	39.52	39.67	27.76	45.66	40.6	35.46
mE5-large-instruct	22.27	30.41	27.03	27.59	34.27	39.89	24.17	40.88	37.05	31.51
AfriE5-large-instruct	24.07	30.74	28.37	30.18	36.21	38.8	28.13	42.2	36.35	32.78
FloresBitextMining										
bge-m3	83.73	71.4	82.71	72.78	76.63	86.13	81.99	65.33	82.24	78.1
gemini embedding 001	91.3	77.8	90.01	87.9	90.13	91.43	90.37	83.99	90.28	88.13
mE5-large-instruct	92.58	88.51	91.5	91.65	92.05	93.22	92.58	87.88	92.72	91.41
AfriE5-large-instruct	92.37	88.67	91.32	91.32	91.83	92.91	92.29	87.92	92.51	91.24
InjongoIntent										
bge-m3	85.84	61.19	86.5	71.86	65.66	90.36	74.62	74.11	68.05	75.35
gemini embedding 001	89.5	65.23	95.25	83.14	77.72	93.64	87.38	84.28	79.22	83.93
mE5-large-instruct	80.86	64.89	85.05	74.45	65.95	80.17	78.11	79.38	70.25	75.46
AfriE5-large-instruct	82.84	62.31	85.95	72.3	65.2	84.42	79.67	77.73	68.36	75.42
NTREXBitextMining										
bge-m3	79.36	60.9	86.3	81.64	79.04	93.91	83.76	72.93	86.07	80.43
gemini embedding 001	87.2	65.69	88.16	87.03	81.02	94.49	86.33	77.45	90.35	84.19
mE5-large-instruct	92.07	73.8	94.81	94.57	92.65	97.45	93.82	89.26	95.15	91.51
AfriE5-large-instruct	92.87	78.04	94.67	94.63	91.83	97.43	93.88	89.67	95.37	92.04
SIB200-14Classes										
bge-m3	16.05	6.08	13.59	6.85	9.23	16.92	9.83	4.08	9.31	10.22
gemini embedding 001	18.27	8.13	18.58	16.24	23.23	20.03	18.46	13.64	23.07	17.74
mE5-large-instruct	21.24	12.61	22.11	21.49	23.8	32.44	21.63	17.23	25.51	22.01
AfriE5-large-instruct	30.21	13.28	27.48	25.91	29.02	33.05	27.06	20.71	28.82	26.17
SIB200Classification										
bge-m3	64.9	44.22	60.85	51.16	53.44	67.99	56.41	46.73	55.07	55.64
gemini embedding 001	71.72	57.15	69.65	69.84	73.69	73.4	71.01	65	71.3	69.2
mE5-large-instruct	72.87	57.09	71.27	73.7	74.68	76.63	74.1	66.05	74.14	71.17
AfriE5-large-instruct	75.47	59.82	72.57	73.69	75.16	76.46	73.8	67.1	74.04	72.01
SIB200ClusteringFast										
bge-m3	29.82	10.84	22.91	15.01	18.84	34.49	21.51	13.4	22.02	20.98
gemini embedding 001	39.18	20.92	40.94	37	37.29	38.3	34.57	27.63	35.68	34.61
mE5-large-instruct	48.96	30.47	44.47	46.11	47.42	50.32	46.65	36.03	45.04	43.94
AfriE5-large-instruct	54.35	31.61	44.87	46.56	48.65	49.95	49.53	38.31	47.52	45.71
BelebeleRetrieval										
BGE-m3	79.99	58.88	75.48	61.26	65.58	87.92	69.13	60.84	69.59	69.85
Gemini Embedding	91.07	68.23	86.46	80.26	86.01	92.65	86.33	75	86.61	83.62
mE5-Large-Instruct	81.08	66.34	77.54	73.72	76.93	86.97	77.63	62.97	77.97	75.68
AfriE5-Large-Instruct	83.34	69.42	78.71	75.71	78.34	87.99	80.16	65.25	80.26	77.69

Table 11: **Task-wise and per-language Performance.** Comparison of the selected models across all African languages for each task.