The Multilingual Curse at the Retrieval Layer: Evidence from Amharic

Yosef Worku Alemneh[*]
Independent Researcher
yosefwalemneh@gmail.com

Kidist Amde Mekonnen[*]
University of Amsterdam
k.a.mekonnen@uva.nl

Maarten de Rijke
University of Amsterdam
m.derijke@uva.nl

Abstract

Multilingual retrieval increasingly underpins cross-lingual question answering and retrieval-augmented generation. Strong zero-shot scores on multilingual benchmarks are often taken as evidence that current encoders transfer reliably across many languages. We argue that this assumption breaks down for underrepresented, morphologically rich languages, and use Amharic as a diagnostic case. Under a shared passage retrieval protocol covering dense, late-interaction, learned sparse, and cross-encoder paradigms, we compare zero-shot multilingual retrievers, Amharic-fine-tuned multilingual retrievers, and monolingual Amharic retrievers. The strongest zero-shot multilingual retriever underperforms the strongest monolingual Amharic first-stage retriever by 23% relative MRR@10. Fine-tuning two recent multilingual embedding models on the same Amharic supervision yields 32–60% relative MRR@10 gains over zero-shot, but the best Amharic-fine-tuned multilingual model remains below the strongest monolingual Amharic retriever. These findings indicate that zero-shot multilingual retrieval is not a sufficient proxy for equitable information access in the LLM era: for underrepresented languages, retrieval must be evaluated and adapted in-language rather than inferred from aggregate multilingual benchmarks. To foster future research, we publicly release the dataset, codebase, and trained models.[1]

1 Introduction

Multilingual retrieval as an access layer. Multilingual retrieval is becoming a central component of how language technologies access information across languages, particularly in semantic search, cross-lingual information access, and retrieval-augmented generation. Recent multilingual embedding models are increasingly positioned as off-the-shelf retrieval encoders for broad language coverage (Vera et al., 2025; Microsoft, 2026; Wang et al., 2024; Zhang et al., 2024; Yu et al., 2024). Strong benchmark scores may therefore make zero-shot multilingual retrieval appear sufficient across languages.

A retrieval-layer multilingual curse. We argue that this reading is premature. The gap is particularly visible, and consequential, for underrepresented, morphologically rich languages. When retrieval representations transfer poorly to such languages, downstream systems that rely on retrieved evidence inherit a retrieval-imposed quality ceiling. As a result, multilingual RAG systems and LLM-based question answering pipelines may degrade in ways that are obscured by aggregate multilingual evaluations.

Why Amharic. We make this concrete using Amharic. Amharic is a Semitic language with over 58 million first- and second-language speakers (Basha et al., 2023; Eberhard et al., 2024), written in the Ethiopic (Ge'ez) script. Its root-and-pattern morphology, extensive affixation, and script-specific orthography make it challenging for multilingual encoders, which often rely on subword segmentations and vocabularies that are poorly matched to underrepresented, morphologically rich languages (Rust et al., 2021; Mekonnen et al., 2025). Amharic is therefore a useful diagnostic case rather than an idiosyncratic one: it is widely used but persistently under-resourced in IR, and the factors that strain retrieval here, namely script, morphology, and limited language-specific supervision, recur across many of the world's languages.

Benchmark and model suite. To stress-test multilingual retrieval transfer, we extend the Amharic passage retrieval benchmark introduced by Mekonnen et al. (2025) and construct Dataset V2, with

[*]Equal contribution.

[1]https://github.com/rasyosef/amharic-neural-ir

Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models (MeLLM 2026), pages 201–210

68K query–passage pairs from AMNEWS (Azime and Mohammed, 2021), XL-SUM (Hasan et al., 2021), Amharic Wikipedia, and AmQA (Abedissa et al., 2023). We evaluate four retrieval paradigms, dense bi-encoders, late-interaction retrievers, learned sparse retrievers, and cross-encoder re-rankers under a shared protocol. We compare zero-shot multilingual retrievers, Amharic-fine-tuned multilingual retrievers, and monolingual Amharic retrievers, including those of Mekonnen et al. (2025), with BM25 as a sparse lexical reference. This extends prior Amharic neural IR with two additional retrieval families and a multi-source benchmark.

Main findings. The headline finding is clear: the strongest zero-shot multilingual retriever underperforms the strongest monolingual Amharic first-stage retriever by 23% relative MRR@10. Fine-tuning two recent multilingual embedding models on the same Amharic supervision narrows the gap substantially, yielding 32–60% relative MRR@10 gains over their zero-shot counterparts. However, the best Amharic-fine-tuned multilingual model remains below the strongest monolingual Amharic retriever despite having roughly 2.5× more parameters. Cross-encoder re-ranking on top of an Amharic dense bi-encoder reaches MRR@10 of 0.830, the highest score in our evaluation. These results suggest that zero-shot multilingual retrieval is not yet a sufficient proxy for equitable information access in the LLM era. One consequence is that multilingual RAG systems built on zero-shot retrievers may inherit a measurable retrieval-layer quality ceiling for languages like Amharic unless retrieval is evaluated and adapted in-language.

Contributions. (i) We use Amharic to test whether strong zero-shot multilingual retrieval transfers to an underrepresented, morphologically rich language. (ii) We construct an expanded Amharic passage retrieval benchmark with 68K query–passage pairs and evaluate four retrieval paradigms under a shared protocol. (iii) We quantify a persistent zero-shot transfer gap and show that Amharic fine-tuning substantially narrows but does not eliminate it. (iv) We release the benchmark, code, evaluation scripts, and model checkpoints to support reproducible research on Amharic retrieval and information access for low-resource languages.[2]

[2]Benchmark; code and evaluation scripts; monolingual Amharic models; Amharic-fine-tuned multilingual models.

2 Related Work

Multilingual retrieval and its evaluation. Dense multilingual encoders have become a standard basis for multilingual retrieval, with models such as multilingual E5, Arctic Embed, mGTE, EmbeddingGemma, and harrier-oss-v1 introduced as general-purpose retrieval encoders across languages (Wang et al., 2024; Yu et al., 2024; Zhang et al., 2024; Vera et al., 2025; Microsoft, 2026). Recent multilingual learned sparse retrievers further extend this line to sparse lexical representations by mapping multilingual inputs into a shared lexical space (Nguyen et al., 2026). This trend is increasingly consequential for retrieval-augmented generation: multilingual RAG systems rely on retrieval as the evidence-selection layer, and recent work has begun to evaluate retrieval quality in multilingual generation settings (Chirkova et al., 2024; Thakur et al., 2024).

Multilingual retrieval benchmarks have also expanded: MIRACL provides native-annotated ad hoc retrieval data over Wikipedia across 18 languages (Zhang et al., 2023b), mMARCO extends MS MARCO via machine translation (Bonifacio et al., 2021), and XOR-TyDi frames open-retrieval QA as cross-lingual retrieval (Asai et al., 2021). These resources make evaluation more systematic, but coverage remains uneven: underrepresented languages with complex morphology and divergent scripts are less consistently represented, and aggregate scores can obscure language-level degradation. Tokenizer-level analyses further show that multilingual tokenizers often over-segment such languages, producing fragmented representations that can affect downstream modeling (Rust et al., 2021). For retrieval, where representation quality directly shapes ranking, this motivates shared-protocol evaluation at the language level rather than reliance on aggregate multilingual scores.

A common response to weak zero-shot transfer is in-language fine-tuning: adapting multilingual retrievers using language-specific query–passage supervision (Bonifacio et al., 2021; Zhang et al., 2023a). For underrepresented languages, however, such supervision is often weakly labeled, domain-restricted, or limited in scale. For Amharic, prior work has shown that monolingual Amharic retrievers outperform strong multilingual baselines, and that Amharic fine-tuning can substantially improve multilingual retrieval (Mekonnen et al., 2025). We build on this evidence by testing whether the

same pattern holds under a broader setup: a more diverse Amharic benchmark, additional retrieval paradigms, and a unified comparison of zero-shot multilingual retrievers, Amharic-fine-tuned multilingual retrievers, and monolingual Amharic retrievers. This allows us to quantify the gap that remains after zero-shot transfer and after in-language fine-tuning.

Amharic neural information retrieval. Amharic IR has been constrained by the scarcity of relevance-judged collections. The earliest dedicated resource, 2AIRTC (Yeshambel et al., 2020), is a TREC-style ad hoc retrieval collection, but its limited query set and incomplete judgments make it difficult to use as the sole basis for evaluating neural retrievers. Recent work on Amharic neural IR has therefore relied on weakly supervised passage retrieval benchmarks constructed from public Amharic resources.

The work most directly related to ours is (Mekonnen et al., 2025), which introduced monolingual Amharic dense retrievers based on Amharic BERT and RoBERTa backbones, evaluated multilingual dense baselines and BM25, fine-tuned a multilingual retriever with Amharic supervision, and trained a ColBERT-style late-interaction retriever. Their benchmark was derived from AMNEWS (Azime and Mohammed, 2021), using headlines as queries and article bodies as relevant passages. They showed that monolingual Amharic retrievers outperform strong multilingual embedding baselines, and that Amharic fine-tuning substantially improves multilingual retrieval effectiveness.

We build on this foundation rather than reintroducing Amharic dense retrieval. Our contribution is to extend the prior setup in three directions. First, we evaluate on a multi-source V2 benchmark derived from AMNEWS (Azime and Mohammed, 2021), XL-SUM (Hasan et al., 2021), Amharic Wikipedia, and AmQA (Abedissa et al., 2023), moving beyond a single news-derived source. Second, we broaden the retrieval stack by adding learned sparse retrieval and cross-encoder re-ranking to the dense and late-interaction models studied previously. Third, we provide a shared-protocol comparison of zero-shot multilingual retrievers, Amharic-fine-tuned multilingual retrievers, and monolingual Amharic retrievers. This lets us ask not only whether monolingual Amharic retrieval helps, but also whether recent multilingual encoders close the gap once fine-tuned with the same Amharic supervision.

3 Experimental Setting

We evaluate Amharic passage retrieval using a shared benchmark and protocol that covers monolingual Amharic retrievers, zero-shot multilingual retrievers, Amharic-fine-tuned multilingual retrievers, first-stage retrieval, and second-stage re-ranking. Because the evaluated paradigms differ in supervision format, scoring function, and retrieval backend, we treat the results as benchmark-level comparisons of practical retrieval approaches rather than controlled architectural ablations.

3.1 Benchmark and evaluation

Dataset. We use the Amharic Passage Retrieval Dataset V2, a refined multi-source extension of the benchmark introduced by Mekonnen et al. (2025). The dataset is derived from AMNEWS (Azime and Mohammed, 2021), XL-SUM (Hasan et al., 2021), Amharic Wikipedia, and AmQA (Abedissa et al., 2023). Following weakly supervised retrieval practice (Hermann et al., 2015; Wu et al., 2020; Mekonnen et al., 2025), headlines serve as queries and article bodies as relevant passages for news-style sources, while AmQA questions serve as queries and answer-containing paragraphs as relevant passages. After MD5-based deduplication, the benchmark contains 68,000 query–passage pairs with a fixed 90/10 train–test split.

Relevance and metrics. Each query has a single labeled positive passage derived from source alignment. Relevance judgments are therefore binary and likely incomplete: unlabeled passages may still be relevant, but only the source-aligned passage is treated as positive. We report Recall@5, Recall@10, MRR@10, and NDCG@10 for first-stage retrieval, and the same metrics at cutoff 10 after re-ranking the top-50 first-stage candidates. Under single-positive supervision, Recall@k should be read as a hit-style measure, while MRR@10 and NDCG@10 reflect rank sensitivity with respect to the aligned positive rather than exhaustive relevance.

3.2 Retrieval models

Monolingual Amharic retrievers. We use *monolingual Amharic retrievers* to refer to models initialized from Amharic-only backbones and trained for Amharic retrieval, distinguishing them from multilingual encoders evaluated zero-shot or fine-tuned on Amharic supervision. We train monolingual Amharic retrievers across four paradigms: dense

bi-encoders (Reimers and Gurevych, 2019), late-interaction retrievers (Khattab and Zaharia, 2020), learned sparse retrievers (Formal et al., 2021), and cross-encoder re-rankers (Nogueira and Cho, 2019). For each paradigm, we train Medium and Base variants initialized from `roberta-medium-amharic` and `roberta-base-amharic` (Alemneh, 2024), respectively. Table 1 summarizes architectures, training data, and objectives for all eight checkpoints.

(i) Dense bi-encoders encode queries and passages independently, mean-pool final hidden states, and ℓ_2-normalize the resulting embeddings. They are trained with MultipleNegativesRankingLoss (Henderson et al., 2017) and Matryoshka representations (Kusupati et al., 2022). (ii) Late-interaction retrievers use token-level MaxSim scoring and are implemented in PyLate (Chaffin and Sourty, 2025); they project contextualized token embeddings to 128 dimensions, truncate queries to 32 tokens and passages to 256 tokens, and train with one positive and up to four pre-mined negatives per query. (iii) Learned sparse retrievers use SPLADE-style pooling over the tokenizer vocabulary and combine a contrastive ranking objective with FLOPs-based sparsity regularization (Formal et al., 2021, 2022). (iv) Cross-encoder re-rankers jointly encode query–passage pairs and predict a scalar relevance score using weighted Binary Cross-Entropy with `pos_weight=7`.

Multilingual baselines. We benchmark five multilingual dense embedding models released for retrieval applications: multilingual-e5-large-instruct (Wang et al., 2024), snowflake-arctic-embed-l-v2.0 (Yu et al., 2024), gte-multilingual-base (Zhang et al., 2024), embeddinggemma-300m (Vera et al., 2025), and harrier-oss-v1-270m (Microsoft, 2026). All five are evaluated as zero-shot multilingual retrievers using each model's released interface: native tokenizer, pooling strategy, encoding procedure, and model-specific query or passage prompts when recommended by the model authors. For harrier-oss-v1-270m, we follow the model card and apply the recommended query-side instruction prompt during zero-shot evaluation, while passages are encoded without an instruction.

Of these multilingual baselines, two recent multilingual embedding models, embeddinggemma-300m and harrier-oss-v1-270m, are also evaluated as Amharic-fine-tuned multilingual retrievers. We fine-tune them on the Dataset V2 training split using the recipe described in Section 3.3. Holding the training split and adaptation recipe fixed across these baselines reduces confounding from supervision differences and focuses the comparison on multilingual initialization and transfer. We additionally include BM25 as a sparse lexical reference and the monolingual Amharic retrievers from Mekonnen et al. (2025) as prior in-language reference points.

3.3 Training and indexing

All monolingual Amharic models are trained on NVIDIA A100 40GB GPUs with AdamW, mixed precision (FP16), and a fixed random seed of 42. Medium variants are initialized from *roberta-medium-amharic*, and Base variants from *roberta-base-amharic*. Dense bi-encoders and cross-encoders use early stopping on validation NDCG@10; late-interaction and learned sparse models are trained for a fixed number of epochs. Table 2 summarizes the per-paradigm training configurations.

For Amharic-fine-tuned multilingual retrievers, we fine-tune embeddinggemma-300m and harrier-oss-v1-270m on a single NVIDIA H100 GPU using SentenceTransformers with MultipleNegativesRankingLoss wrapped in Matryoshka Loss. Both models are trained for 6 epochs with batch size 32 and 4 gradient accumulation steps, giving an effective batch size of 128. We use a cosine learning-rate schedule with peak learning rate 4×10^{-5}, warmup ratio 0.025, and BF16 mixed precision. Training triples pair each query with its aligned positive passage and up to four pre-mined negatives provided in the dataset. For each query, we select the two most similar and two least similar passages from the available negative candidates, using the dataset-provided cosine similarity scores. Fine-tuned multilingual models are evaluated without prompts, matching the prompt-free fine-tuning format.

Tokenization can materially affect retrieval in morphologically rich languages (Rust et al., 2021; Mekonnen et al., 2025); we therefore use the native tokenizer of each backbone and apply a fixed text normalization pipeline across all experiments. Dense bi-encoders use FAISS-based nearest-neighbor retrieval; late-interaction models use the Voyager HNSW index via PyLate; learned sparse retrievers use inverted-index retrieval over sparse term-weight vectors. Retrieval depth and paradigm-specific search settings are held fixed within each comparison.

Checkpoint	Family	Backbone	Params	Dim.	Train N	Data	Objective	MaxLen
`Embed-Med`	Dense	RoBERTa-Med	42M	512	122,938	Triples	MNR+MRL	510
`Embed-Base`	Dense	RoBERTa-Base	110M	768	245,876	Triples	MNR+MRL	510
`ColBERT-Med`	Late-inter.	RoBERTa-Med	42M	128	118,938	Triples	ColBERT-IB	256
`ColBERT-Base`	Late-inter.	RoBERTa-Base	110M	128	118,938	Triples	ColBERT-IB	256
`SPLADE-Med`	Sparse	RoBERTa-Med	42M	$\|V\|$	245,876	Triples	SPLADE-SR	510
`SPLADE-Base`	Sparse	RoBERTa-Base	110M	$\|V\|$	245,876	Triples	SPLADE-SR	510
`Re-rank-Med`	Cross-enc.	RoBERTa-Med	42M	1	491,752	Pairs	BCE-w7	510
`Re-rank-Base`	Cross-enc.	RoBERTa-Base	110M	1	491,752	Pairs	BCE-w7	510

Table 1: Monolingual Amharic retrievers evaluated in this paper. We include Medium and Base variants across dense, late-interaction, learned sparse, and cross-encoder paradigms; checkpoint names are abbreviated, with full identifiers in the released model collections. Dim. denotes output dimensionality ($|V|$ for sparse vocabulary size), and MaxLen denotes the training truncation length. Objectives: MNR+MRL = MultipleNegativesRankingLoss with Matryoshka representations; ColBERT-IB = in-batch ColBERT training; SPLADE-SR = sparsity-regularized SPLADE loss; BCE-w7 = binary cross-entropy with `pos_weight=7`.

Paradigm	LR	Schedule	WU	BS	Epochs	ES
Dense	6×10^{-5}	cosine	2.5%	64–128	6	Yes
Late-inter.	1×10^{-5}	linear	0%	32	4	–
Sparse	6×10^{-5}	cosine	2.5–5%	32–48	4–6	–
Cross-enc.	4×10^{-5}	cosine	5%	64	4	Yes

Table 2: Training hyperparameters for monolingual Amharic retrievers. LR = learning rate; WU = warmup ratio; BS = batch size; ES = early stopping on validation NDCG@10.

4 Experimental Results

We report first-stage retrieval, multilingual fine-tuning, and two-stage re-ranking results on the Amharic Passage Retrieval Dataset V2. Comparisons follow the shared protocol described in Section 3. Because each query is associated with a single source-aligned positive, Recall@k should be interpreted as a hit-style measure, while MRR@10 and NDCG@10 reflect rank sensitivity with respect to the labeled positive.

4.1 First-stage retrieval: a persistent zero-shot gap

Table 3 reports first-stage retrieval effectiveness for BM25, zero-shot multilingual dense retrievers, Amharic-fine-tuned multilingual dense retrievers, monolingual Amharic retrievers from prior work, and the monolingual Amharic retrievers introduced in this work. Three patterns are visible.

Lexical retrieval remains competitive. BM25 reaches MRR@10 of 0.612, outperforming four of the five zero-shot multilingual dense retrievers. Lexical matching is therefore not a weak baseline in this setting: several recent multilingual encoders with hundreds of millions of parameters do not uniformly improve over sparse lexical retrieval on Amharic.

Zero-shot multilingual retrieval remains below monolingual Amharic retrieval. Among zero-shot multilingual dense retrievers, *snowflake-arctic-embed-l-v2.0* performs best, scoring 0.653/0.701 on MRR@10/NDCG@10. The strongest monolingual Amharic first-stage retriever, *ColBERT-Base-Amharic*, reaches 0.803/0.835 with 110M parameters. This corresponds to a 23.0% relative MRR@10 gap and a 19.1% relative NDCG@10 gap. The gap is not explained by parameter count: *snowflake-arctic-embed-l-v2.0* has 568M parameters, over five times larger than *ColBERT-Base-Amharic*.

The gap persists across monolingual Amharic paradigms. The strongest monolingual Amharic dense bi-encoder, learned sparse retriever, and late-interaction retriever all outperform the strongest zero-shot multilingual retriever. *Embed-Base-Amharic* reaches MRR@10 of 0.774, *SPLADE-Base-Amharic* reaches 0.754, and *ColBERT-Base-Amharic* reaches 0.803. Late interaction yields the best first-stage effectiveness, while learned sparse retrieval provides an inverted-index alternative that remains competitive with dense retrieval. Within each monolingual Amharic family, Base variants outperform their Medium counterparts, indicating that additional capacity remains useful even in this under-resourced setting.

4.2 Multilingual fine-tuning narrows but does not close the gap

The first-stage results show that zero-shot multilingual retrievers underperform monolingual Amharic retrievers. We next ask whether multilingual encoders recover this gap when given the same

Model	Params (M)	R@5	R@10	MRR@10	NDCG@10
Sparse lexical retrieval					
BM25	–	0.734	0.789	0.612	0.655
Monolingual Amharic retrievers from prior work					
roberta-amharic-text-embedding-medium	42	0.750	0.807	0.616	0.662
roberta-amharic-text-embedding-base	110	0.790	0.844	0.657	0.703
colbert-roberta-amharic-base	110	0.860	0.899	0.736	0.776
Zero-shot multilingual dense retrievers					
embeddinggemma-300m	300	0.558	0.621	0.448	0.489
gte-multilingual-base	305	0.690	0.755	0.557	0.605
harrier-oss-v1-270m	270	0.697	0.753	0.576	0.619
multilingual-e5-large-instruct	560	0.736	0.791	0.603	0.648
snowflake-arctic-embed-l-v2.0	568	0.795	0.848	0.653	0.701
Amharic-fine-tuned multilingual dense retrievers					
embeddinggemma-300m + FT	300	0.813	0.862	0.718	0.753
harrier-oss-v1-270m + FT	270	0.860	0.903	0.760	0.795
Monolingual Amharic retrievers introduced in this work					
SPLADE-Medium-Amharic	42	0.858	0.896	0.728	0.769
SPLADE-Base-Amharic	110	0.871	0.906	0.754	0.792
Embed-Medium-Amharic	42	0.843	0.888	0.744	0.779
Embed-Base-Amharic	110	0.870	0.907	0.774	0.807
ColBERT-Medium-Amharic	42	<u>0.882</u>	<u>0.913</u>	<u>0.778</u>	<u>0.811</u>
ColBERT-Base-Amharic	110	**0.902**†	**0.930**†	**0.803**†	**0.835**†

Table 3: First-stage retrieval results on the Amharic Passage Retrieval Dataset V2. Best values are bolded and second-best values are underlined. Among zero-shot multilingual dense retrievers, *snowflake-arctic-embed-l-v2.0* performs best. Amharic fine-tuning yields large gains, but the strongest monolingual Amharic retriever achieves the best results across all metrics. For zero-shot *harrier-oss-v1-270m*, we use the recommended query-side instruction prompt. † indicates significant improvement over the strongest Amharic-fine-tuned multilingual dense retriever, *harrier-oss-v1-270m + FT*, using paired t-tests over per-query scores with $p < 0.05$.

Amharic supervision. We fine-tune two recent multilingual embedding models, *embeddinggemma-300m* and *harrier-oss-v1-270m*, on the Dataset V2 training split using the dense bi-encoder fine-tuning recipe described in Section 3. Both fine-tuned models are evaluated without prompts, matching the prompt-free fine-tuning format. For Harrier zero-shot evaluation, we use the recommended query-side instruction prompt; this improves zero-shot NDCG@10 from 0.545 to 0.619, confirming that the prompt is part of the intended retrieval interface.

Fine-tuning produces large gains. Fine-tuning substantially improves both multilingual encoders. Gemma improves from MRR@10 0.448 to 0.718, a 60.3% relative gain. Harrier improves from prompted zero-shot MRR@10 0.576 to 0.760, a 32.0% relative gain. The same pattern holds for NDCG@10: Gemma improves from 0.489 to 0.753, and Harrier from 0.619 to 0.795. These gains show that multilingual encoders can learn effective Amharic retrieval behavior when given in-language supervision.

Fine-tuning does not reach the strongest monolingual Amharic retriever. Despite these gains, the best Amharic-fine-tuned multilingual model remains below the strongest monolingual Amharic retriever. Harrier fine-tuned reaches 0.760/0.795 on MRR@10/NDCG@10, below *ColBERT-Base-Amharic* at 0.803/0.835. The residual gap is 5.4% relative MRR@10 and 4.8% relative NDCG@10. Gemma fine-tuned remains further behind, with MRR@10/NDCG@10 of 0.718/0.753. Thus, in-language fine-tuning narrows the zero-shot gap substantially, but it does not eliminate it.

Parameter scale is not sufficient. The Amharic-fine-tuned multilingual models have 270M–300M parameters, while the strongest monolingual Amharic first-stage retriever has 110M parameters. Harrier fine-tuned is competitive with sev-

Model	MRR@10	NDCG@10
Embed-Base-Amharic	0.774	0.807
+ Re-rank-Medium-Amharic	0.805	0.835
+ Re-rank-Base-Amharic	**0.830**	**0.856**

Table 4: Two-stage re-ranking results on the Amharic Passage Retrieval Dataset V2. The first row reports the first-stage Embed-Base-Amharic retriever; cross-encoders re-rank its top-50 candidates, with final rank-sensitive metrics computed at cutoff 10. The Base cross-encoder improves over the first-stage retriever by 7.2% relative MRR@10 and 6.1% relative NDCG@10.

eral monolingual Amharic first-stage models, but it does not match the strongest monolingual Amharic retriever. Gemma fine-tuned also remains below the 42M-parameter *Embed-Medium-Amharic* in MRR@10. These results suggest that multilingual pretraining and parameter scale are not substitutes for language-specific retrieval modeling and supervision.

4.3 Two-stage re-ranking

We re-rank the top-50 candidates retrieved by *Embed-Base-Amharic* using the two monolingual Amharic cross-encoders. We use the dense bi-encoder as the candidate generator because it provides a standard single-vector first-stage retrieval setup and makes the re-ranking gains directly interpretable. Table 4 shows that re-ranking improves over the dense first-stage retriever in both metrics. *Re-rank-Base-Amharic* raises MRR@10 from 0.774 to 0.830, a 7.2% relative gain, and NDCG@10 from 0.807 to 0.856, a 6.1% relative gain. This is the highest score in our evaluation, indicating that joint query–passage scoring helps resolve hard candidate distinctions that are not fully captured by independent bi-encoder representations.

5 Discussion

Zero-shot scores can hide language-level retrieval failures. The results expose a limitation in how multilingual retrieval is often evaluated: aggregate zero-shot performance can mask large language-specific deficits. Amharic is a useful diagnostic case because it combines properties that are common among underrepresented languages but often diluted in multilingual averages: non-Latin script, rich morphology, and limited language-specific retrieval supervision. The strongest zero-shot multilingual retriever reaches 0.653 MRR@10, while the strongest monolingual Amharic first-stage retriever reaches 0.803. This 23% relative gap occurs in the top-10 region, where retrieval output is typically consumed by users, re-rankers, or downstream generation systems. The consequence is not merely a lower aggregate score; it changes which evidence is made available to later stages of an information access pipeline. This is precisely the failure mode that aggregate multilingual benchmarks can obscure: a model may appear broadly reliable while still providing systematically weaker access for a particular language.

The gap is not reducible to retrieval architecture. The zero-shot gap is not explained by a single modeling choice. Monolingual Amharic dense, learned sparse, and late-interaction retrievers all outperform the strongest zero-shot multilingual baseline. This weakens a purely architectural interpretation: the issue is not simply that one scoring function is stronger than another, but that the multilingual representation space itself transfers imperfectly to Amharic. Architecture still matters, late interaction gives the strongest first-stage retrieval, and cross-encoder re-ranking reaches the highest overall score, but language-specific modeling changes the operating point before architecture-specific gains are considered. In practical terms, a stronger retrieval architecture cannot fully compensate for a representation space that is poorly aligned with the target language. This also helps explain why BM25 remains competitive: lexical overlap, despite its limitations, can be less brittle than a dense representation that fails to align Amharic queries and passages reliably.

Fine-tuning reveals capacity, but not parity. The fine-tuning results complicate a simple failure narrative. Multilingual encoders are not unable to learn Amharic retrieval: Gemma improves from 0.448 to 0.718 MRR@10, and Harrier improves from its prompted zero-shot score of 0.576 to 0.760. This shows that in-language supervision can unlock useful retrieval behavior that is not expressed in the zero-shot setting. However, fine-tuning does not establish parity with monolingual Amharic retrievers. Harrier remains below *ColBERT-Base-Amharic* despite having more than twice as many parameters, and Gemma remains further behind. The relevant comparison is therefore not only whether multilingual encoders improve after fine-tuning, but whether they match in-language alternatives at comparable effectiveness and deployment cost. In our setting, they become competitive, but they do not dominate smaller monolingual Amharic retrievers.

This distinction matters for how adaptation

should be interpreted. If fine-tuning only narrowed the gap from a weak zero-shot baseline, the result would support the value of Amharic supervision but would not show that multilingual pretraining is sufficient. The residual gap suggests that the starting representation, tokenizer, and pretraining distribution still matter after supervised adaptation. In-language fine-tuning is therefore necessary for competitiveness, but it is not a substitute for evaluating whether the resulting model matches monolingual alternatives.

Multilingual RAG evaluation should expose retrieval quality. These findings have direct implications for multilingual RAG and LLM-based question answering. Retrieval is the evidence-selection layer: if the relevant passage is not surfaced near the top of the ranked list, the generator receives incomplete or misleading context, and downstream reasoning cannot reliably recover evidence that was never retrieved. For underrepresented languages, RAG evaluation should therefore report retrieval-side quality per language rather than relying only on final generated answers or aggregate multilingual scores. More broadly, the evaluation question should move from whether a model nominally supports a language to whether it retrieves effectively for that language under realistic supervision, architecture, and deployment constraints.

This does not imply that every language requires a separate retrieval stack. Rather, zero-shot multilingual retrieval should not be treated as sufficient evidence of language coverage: for languages such as Amharic, reliable access requires direct retrieval-layer evaluation, fine-tuning tests, and comparison against strong monolingual baselines.

6 Conclusion

We use Amharic to test whether strong zero-shot multilingual retrieval transfers to an underrepresented, morphologically rich language. Under a shared passage retrieval protocol, zero-shot multilingual retrievers remain substantially below monolingual Amharic retrievers, with the strongest zero-shot model underperforming the strongest monolingual first-stage retriever by 23% relative MRR@10. Amharic fine-tuning yields large gains for multilingual encoders, but does not close the gap to the strongest monolingual Amharic retriever. These results show that multilingual retrieval quality cannot be inferred from aggregate zero-shot benchmarks alone: for languages such as Amharic, retrieval must be evaluated and adapted in-language before downstream information-access claims can be trusted.

7 Limitations

Our claims are subject to four main limitations. First, the empirical evidence is limited to Amharic. We argue that the relevant conditions, script divergence, rich morphology, and limited multilingual pretraining coverage also occur in other underrepresented languages, but we do not show that the gap appears at comparable magnitudes elsewhere. Second, Dataset V2 uses weakly supervised, source-aligned positives: each query has a single labeled relevant passage, while other relevant passages may remain unlabeled. This makes absolute scores conservative and may affect the size of observed gaps relative to a fully judged collection. Third, the multilingual fine-tuning study covers two recent multilingual embedding models, Gemma and Harrier, with zero-shot results for five multilingual baselines; larger or differently instruction-tuned multilingual retrievers may behave differently. Finally, our RAG implications are inferred from retrieval metrics rather than measured in an end-to-end generation pipeline. Evaluating whether these retrieval gaps translate into answer-level degradation for Amharic RAG remains future work.

8 Ethical Considerations

This work targets Amharic, a widely spoken but under-served language in information access. Dataset V2 is derived from publicly released sources (AMNEWS, XL-SUM, Amharic Wikipedia, AmQA) under their original licenses; no new human-subject data is collected. Since query–passage labels are weakly derived from source structure, the benchmark is not a human-judged relevance collection. Improved retrieval can aid downstream RAG and QA, but may also propagate source biases if deployed without auditing. We release models and code for reproducible research, not as audited systems for high-stakes deployment.

Acknowledgements

This work was partially supported by NWO projects EINF-18163/L1, 024.004.022, NWA.-1389.20.183, and KICH3.LTP.20.006, and by the European Union's Horizon Europe program under grant agreement No. 101070212. The content reflects only the authors' views.

References

Tilahun Abedissa, Ricardo Usbeck, and Yaregal Assabie. 2023. AmQA: Amharic question answering dataset. *arXiv preprint arXiv:2303.03290*.

Yosef Worku Alemneh. 2024. Amharic bert and roberta models. Hugging Face model collection.

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics.

Israel Abebe Azime and Nebil Mohammed. 2021. An Amharic news text classification dataset. *arXiv preprint arXiv:2103.05639*.

Shaik Johny Basha, Duggineni Veeraiah, Boddu Venkat Charan, Wiltrud Sahithi Joyce Yeddu, and Devalla Ganesh Babu. 2023. Detection and comparative analysis of handwritten words of Amharic language to english using CNN-based frameworks. In *2023 International Conference on Inventive Computation Technologies (ICICT)*, pages 422–427. IEEE.

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2021. mMARCO: A multilingual version of the MS MARCO passage ranking dataset. *arXiv preprint arXiv:2108.13897*.

Antoine Chaffin and Raphaël Sourty. 2025. PyLate: Flexible training and retrieval for late interaction models. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM 2025, Seoul, Republic of Korea, November 10-14, 2025*, pages 6334–6339. ACM.

Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. 2024. Retrieval-augmented generation in multilingual settings. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 177–188, Bangkok, Thailand. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. *Ethnologue: Languages of the World*, 27th edition. SIL International, Dallas, Texas.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. From distillation to hard negative sampling: Making sparse neural IR models more effective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2353–2359, New York, NY, USA. Association for Computing Machinery.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2288–2292, New York, NY, USA. Association for Computing Machinery.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703. ACL.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1693–1701, Cambridge, MA, USA. MIT Press.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2022. Matryoshka representation learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Kidist Amde Mekonnen, Yosef Worku Alemneh, and Maarten de Rijke. 2025. Optimized text embedding models and benchmarks for Amharic passage retrieval. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10428–10445, Vienna, Austria. Association for Computational Linguistics.

Microsoft. 2026. Harrier-OSS-v1-270M. Hugging Face model card.

Thong Nguyen, Yibin Lei, Jia-Huei Ju, Eugene Yang, and Andrew Yates. 2026. Milco: Learned sparse retrieval across languages via a multilingual connector. In *International Conference on Learning Representations*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

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.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? On the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Nandan Thakur, Luiz Bonifacio, Crystina Zhang, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Boxing Chen, Mehdi Rezagholizadeh, and Jimmy Lin. 2024. "Knowing when you don't know": A multilingual relevance assessment dataset for robust retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12508–12526, Miami, Florida, USA. Association for Computational Linguistics.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, et al. 2025. EmbeddingGemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3597–3606, Online. Association for Computational Linguistics.

Tilahun Yeshambel, Josiane Mothe, and Yaregal Assabie. 2020. 2AIRTC: The Amharic adhoc information retrieval test collection. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings*, page 55–66, Berlin, Heidelberg. Springer-Verlag.

Puxuan Yu, Luke Merrick, Gaurav Nuti, and Daniel Campos. 2024. Arctic-Embed 2.0: Multilingual retrieval without compromise. *arXiv preprint arXiv:2412.04506*.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.

Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. 2023a. Toward best practices for training multilingual dense retrieval models. *ACM Trans. Inf. Syst.*, 42(2).

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