The Multilingual Curse at the Retrieval Layer

Evidence from Amharic

Yosef Worku Alemneh*, Kidist Amde Mekonnen*, Maarten de Rijke

University of Amsterdam, The Netherlands *Equal contribution

yosefwalemneh@gmail.com, k.a.mekonnen@uva.nl, m.derijke@uva.nl

ACL 2026 SAN DIEGO JULY 2-7

1. Introduction & Motivation

Core question: Do strong multilingual retrievers transfer reliably to **Amharic**, or does retrieval quality break for this underrepresented, morphologically rich language before downstream RAG/QA begins?

Motivation

- Multilingual retrieval is a key access layer for cross-lingual QA and retrieval-augmented generation.
- Strong zero-shot benchmark scores are often treated as evidence of reliable multilingual transfer.
- We stress-test this assumption across dense, late-interaction, learned sparse, and cross-encoder retrieval.

Why Amharic?

- Amharic has over 58M speakers and is written in the Ethiopic (Ge'ez) script.
- Root-and-pattern morphology and script-specific orthography challenge multilingual tokenization.
- These factors recur across many underrepresented languages, making Amharic a diagnostic case.

Main finding: The strongest zero-shot multilingual retriever trails the strongest monolingual Amharic first-stage retriever by **23% relative MRR@10**, exposing a retrieval-layer multilingual gap.

2. Methodology

Setup. We train **monolingual Amharic retrievers** with RoBERTa-Amharic backbones and fine-tune two multilingual encoders using the same Amharic supervision.

Backbones:

- **RoBERTa-Base-Amharic:** 12 layers, 768H, 110M params.
- **RoBERTa-Medium-Amharic:** 8 layers, 512H, 42M params.

Retrieval paradigms and objectives:

- **Dense bi-encoder:** mean pooling + ℓ_2 norm; MNR + Matryoshka loss.
- **ColBERT late interaction:** token-level MaxSim; contrastive loss.
- **SPLADE sparse retriever:** vocabulary-space pooling; sparse ranking loss + FLOPs regularization.
- **Cross-encoder re-ranker:** joint query–passage scoring; weighted BCE loss.

Dataset. Amharic Passage Retrieval Dataset: 68K query-passage pair Amharic benchmark drawn from news, Wikipedia, and QA sources, fixed 90/10 train–test split with weakly supervised positives and mined hard negatives.

Multilingual fine-tuning. We fine-tune EmbeddingGemma-300M and Harrier-OSS-v1-270M with MNR wrapped in Matryoshka loss.

3. Main Results

RQ1. Do zero-shot multilingual retrievers match monolingual Amharic retrievers?

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	0.882	0.913	0.778	0.811
ColBERT-Base-Amharic	110	**0.902**†	**0.930**†	**0.803**†	**0.835**†

Finding 1. BM25 outperforms four of five zero-shot multilingual dense retrievers.
Finding 2. The strongest zero-shot multilingual retriever reaches 0.653 MRR@10, while the strongest monolingual Amharic retriever reaches 0.803.
Finding 3. Amharic fine-tuning narrows the gap, but the strongest monolingual Amharic retriever still performs best across all first-stage metrics.

4. Multilingual Fine-tuning

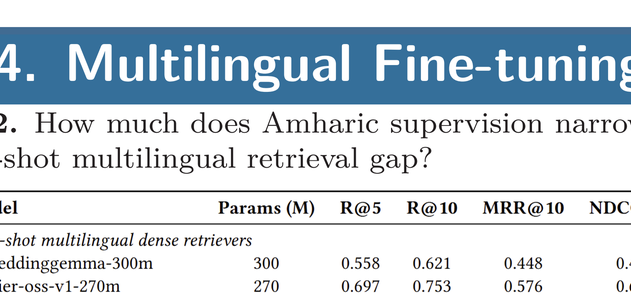
RQ2. How much does Amharic supervision narrow the zero-shot multilingual retrieval gap?

Model	Params (M)	R@5	R@10	MRR@10	NDCG@10
Zero-shot multilingual dense retrievers					
embeddinggemma-300m	300	0.558	0.621	0.448	0.489
harrier-oss-v1-270m	270	0.697	0.753	0.576	0.619
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

Finding. Fine-tuning yields large MRR@10 gains: EmbeddingGemma improves from 0.448 to 0.718, and Harrier improves from 0.576 to 0.760.
But the gap remains: the best Amharic-fine-tuned multilingual model still trails the strongest monolingual Amharic retriever.

5. Model Size vs. Performance

RQ4. Does increasing the Amharic backbone size improve retrieval effectiveness?



Finding. 110M Base models consistently improve MRR@10 over 42M Medium models.
Tradeoff. Medium models remain a strong choice when efficiency is a priority.

6. Two-stage Re-ranking

RQ3. Does cross-encoder re-ranking improve over first-stage retrieval?

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**

Finding. Re-ranking the top-50 candidates from Embed-Base-Amharic improves MRR@10 from 0.774 to **0.830** and NDCG@10 from 0.807 to **0.856**.
Takeaway. Joint query–passage scoring gives the highest overall score in our evaluation.

Conclusion

Key Takeaways, Limitations & Future Work

- Zero-shot multilingual retrieval is not a reliable proxy for Amharic retrieval quality.
- Monolingual Amharic retrievers outperform all zero-shot and fine-tuned multilingual baselines.
- Amharic fine-tuning yields large gains, but lags behind the strongest monolingual retriever.
- Two-stage re-ranking achieves the highest overall retrieval effectiveness.
- Limitations: weakly supervised labels and evidence from Amharic only.
- Future work: human-judged benchmarks and end-to-end Amharic RAG.

Scan me