Optimized Text Embedding Models and Benchmarks for Amharic Passage Retrieval

Kidist Amde Mekonnen*, **Yosef Worku Alemneh***, **Maarten de Rijke**
University of Amsterdam, The Netherlands
k.a.mekonnen@uva.nl, yosefwalemneh@gmail.com, m.derijke@uva.nl

ACL 2025 VIENNA JULY 27 - AUGUST 1

1. Introduction & Motivation

Document retrieval is essential in applications like QA and fact-checking.

Existing neural retrieval models underperform for low-resource, morphologically rich languages like Amharic.

Challenges include data scarcity, complex morphology, and poor tokenization in multilingual models.

We develop Amharic-specific dense retrieval models using BERT and RoBERTa backbones.

Our models outperform large multilingual baselines with significantly fewer parameters.

Why Amharic?

Ethiopia's official federal working language, spoken by over 60 million. One of the most widely spoken Semitic languages after Arabic.

Morphologically rich and written in Ge'ez script, standard tokenizers often over-segment words.

Root–pattern morphology and script complexity → frequent query–document mismatches.

Low-resource status ⇒ generic search often retrieves irrelevant or even harmful content (Nigatu and Raji, 2024).

2. Methodology

We train Amharic-specific **bi-encoders** and a **ColBERT late-interaction model** using RoBERTa and BERT backbones.

Bi-Encoders (mean-pool + L2 norm):

1. RoBERTa-Base-Amharic-Embed (12L, 768H, 110M)
2. RoBERTa-Medium-Amharic-Embed (8L, 512H, 42M)
3. BERT-Medium-Amharic-Embed (8L, 512H, 40M)

ColBERT: Late interaction with token-level max-sim pooling for fine-grained ranking.

Training Objective: Multiple Negatives Ranking Loss for the Bi-Encoders and Contrastive loss for ColBERT.

Loss (MNRL):

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{e^{f(q_i, p_i^+)}}{e^{f(q_i, p_i^+)} + \sum_{j \neq i} e^{f(q_i, p_j^-)}}$$

This encourages higher scores for relevant passages over in-batch negatives.

3. Main Results

RQ1. How well do Amharic-optimized embeddings improve ranking accuracy compared to general-purpose multilingual embedding models?

	Model	Params	MRR@10	NDCG@10	Recall @10	Recall @50	Recall @100
Multilingual models	gte-modernbert-base	149M	0.019	0.023	0.033	0.051	0.067
	gte-multilingual-base	305M	0.600	0.638	0.760	0.851	0.882
	multilingual-e5-large-instruct	560M	0.672	0.709	0.825	0.911	0.931
	snowflake-arctic-embed-l-v2.0	568M	0.659	0.701	0.831	0.922	0.942
Ours	BERT-Medium-Amharic-Embed	40M	0.682	0.720	0.843	0.931	0.954
	RoBERTa-Medium-Amharic-Embed	42M	0.735	0.771	0.884	0.955	0.971
	RoBERTa-Base-Amharic-Embed	110M	**0.775**†	**0.808**†	**0.913**†	**0.964**†	**0.979**†

† Our best model outperforms strong multilingual baselines across all metrics.

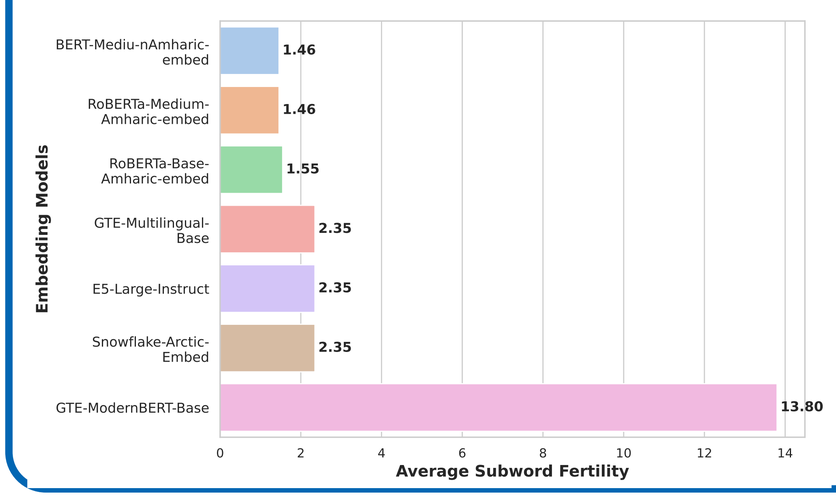
4. Amharic Retrieval Benchmarks

RQ2. How do different retrieval paradigms compare in effectiveness, establishing a benchmark for Amharic passage retrieval?

Type	Model	MRR@10	NDCG@10	Recall @10	Recall @50	Recall @100
Sparse retrieval	BM25-AM	0.657	0.682	0.774	0.847	0.871
Dense retrieval	RoBERTa-Base-Amharic-Embed	0.775	0.808	0.913	0.964	0.979
Dense retrieval	ColBERT-RoBERTa-Base-Amharic	**0.843**†	**0.866**†	**0.939**†	**0.972**†	0.979

† ColBERT achieves best performance across all metrics.

5. Tokenization Quality ↔ Retrieval Performance

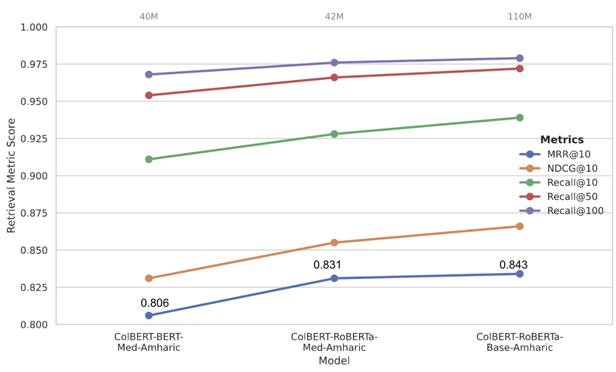


RQ3. How does tokenization quality, particularly subword segmentation, impact retrieval effectiveness in morphologically rich, low-resource languages like Amharic?

Tokenizer	Tokenized Output	Subword Count	Subword Fertility
RoBERTa-Base-Amharic-embed	['__የ', 'ተዷጋገመ', 'ው', '__የመረሉት', '__መጎቀጥቀጥ', 'ና', '__የእ', 'ሳተ', '__ገሞራ', '__ምልክት', '__በአፋር', '__ክልል']	12	1.50
snowflake-arctic-embed-l-v2.0	['__የተ', 'ዷጋ', 'ገ', 'መው', '__የመ', 'ረሉት', '__መን', 'ቀጥ', 'ቀጥ', 'ና', '__የእ', 'ሳተ', '__ገ', 'ሞ', 'ራ', '__ምልክት', '__በአ', 'ፋር', '__ክልል']	19	2.38

6. Model Size vs. Performance

RQ4. How does base model size affect retrieval in late interaction models?



This demonstrates that well-tuned medium-size encoders strike the best balance between accuracy and efficiency.

7. Fine-Tuning Multilingual Models

Model	MRR@10	NDCG@10	Recall @10	Recall @50	Recall @100
snowflake-arctic-embed-l-v2.0	0.659	0.701	0.831	0.922	0.942
snowflake-arctic-embed-l-v2.0-AM	**0.827**†	**0.855**†	**0.942**†	**0.977**†	**0.985**†

Insight: Off-the-shelf multilingual models remain suboptimal; in-language, retrieval-specific supervision yields large improvements

Conclusion

Key Takeaways, Limitations & Future Work

- Amharic-specific models outperform large multilingual ones.
- Tokenization quality is crucial in morphologically rich languages.
- Evaluation on heuristic headline–article pairs lacks gold labels.
- Future: create human-annotated Amharic retrieval dataset.
- Extend benchmarks beyond news (e.g., QA, social media).
- Develop morphology-aware tokenization.

Scan me