Optimized Text Embedding Models and Benchmarks for Amharic Passage Retrieval

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

Yosef Worku Alemneh* Independent Researcher yosefwalemneh@gmail.com

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

Abstract

Neural retrieval methods using transformer-based pre-trained language models have advanced multilingual and cross-lingual retrieval. However, their effectiveness for low-resource, morphologically rich languages such as Amharic remains underexplored due to data scarcity and suboptimal tokenization. We address this gap by introducing Amharic-specific dense retrieval models based on pre-trained Amharic BERT and RoBERTa backbones. Our proposed *RoBERTa-Base-Amharic-Embed* model (110M parameters) achieves a 17.6% relative improvement in MRR@10 and a 9.86% gain in Recall@10 over the strongest multilingual baseline, *Arctic Embed 2.0* (568M parameters). More compact variants, such as *RoBERTa-Medium-Amharic-Embed* (42M), remain competitive while being over 13× smaller. Additionally, we train a *ColBERT-based late interaction retrieval model* that achieves the highest MRR@10 score (0.843) among all evaluated models. We benchmark our proposed models against both sparse and dense retrieval baselines to systematically assess retrieval effectiveness in Amharic. Our analysis highlights key challenges in low-resource settings and underscores the importance of language-specific adaptation. To foster future research in low-resource IR, we publicly release our dataset, codebase, and trained models.[1]

1 Introduction

As a foundational task in natural language processing (NLP), document retrieval plays a crucial role in applications such as open-domain question answering (Chen et al., 2017) and fact-checking (Thorne et al., 2018). Traditional retrieval systems such as TF-IDF and BM25 (Robertson and Walker, 1997; Robertson and Zaragoza, 2009) match queries to documents based on lexical overlap. While efficient, they struggle with vocabulary mismatch and semantic ambiguity, limiting their generalizability to synonyms and paraphrases. These challenges are particularly pronounced in morphologically rich languages, where high inflectional variability and complex morphology complicate exact-match retrieval. Suboptimal tokenization in multilingual models further exacerbates these issues, leading to over-segmentation and inefficient subword representations (Rust et al., 2021). As a result, word-based indexing methods fail to capture non-concatenative morphology, affixation, and orthographic variations, degrading retrieval effectiveness. To address these limitations, retrieval models must move beyond lexical overlap and incorporate robust semantic representations.

Neural retrieval models. Recent work has introduced several families of neural retrieval methods that leverage transformer-based pre-trained language models to improve retrieval effectiveness, particularly in monolingual English settings. These methods have significantly advanced document ranking, achieving state-of-the-art performance in benchmarks such as MS MARCO (Campos et al., 2016) and Natural Questions (Kwiatkowski et al., 2019). Broadly, they fall into three main categories (Yates et al., 2021): (i) learned sparse retrieval (e.g., SPLADE, Formal et al., 2021a), which enhances queries and documents with context-aware term expansions; (ii) dense retrieval (e.g., DPR, Karpukhin et al., 2020), which maps text into dense vector spaces for efficient retrieval, employing a dual-encoder architecture that encodes queries and documents separately, a design that limits their effectiveness for fine-grained relevance modeling; and (iii) cross-encoders (e.g., Nogueira and Cho, 2019; Nogueira et al., 2019), which address this limitation by jointly encoding query-document pairs, capturing richer contextual interactions, with a computational overhead that

*Equal contribution.

[1] https://github.com/kidist-amde/amharic-ir-benchmarks

restricts their use to re-ranking candidate documents (Humeau et al., 2020). As an alternative, late-interaction models (e.g., ColBERT, Khattab and Zaharia, 2020), introduce token-level interactions and strike a balance between the efficiency of dense retrieval and the expressiveness of cross-encoders.

A newer paradigm, generative information retrieval (Metzler et al., 2021; Tay et al., 2022; Chen et al., 2023), uses pre-trained encoder-decoder models to consolidate indexing, retrieval, and ranking into a single generative framework. While promising, GenIR lags behind dense retrieval in handling large-scale datasets and accommodating dynamic corpora, requiring further study of its scalability and adaptability (Pradeep et al., 2023).

Research gap. Despite these advances, neural retrieval remains understudied for morphologically complex, low-resource languages like Amharic. Most retrieval models are optimized for high-resource languages, and prior work has largely focused on cross-lingual transfer from these languages (Zeng et al., 2023). Despite advancements in multilingual embedding models (Wang et al., 2024; Yu et al., 2024), these approaches remain inadequate for morphologically rich languages due to suboptimal tokenization, poor subword segmentation, and weak cross-lingual transfer (Üstün et al., 2019). Section 2 further explores the importance of addressing this gap in information retrieval research.

Our contribution. To address the gap identified above, we focus on Amharic and introduce optimized retrieval models and benchmarks, making the following key contributions: (i) *Amharic text embeddings*: we develop dense retrieval models for Amharic, leveraging Amharic BERT and RoBERTa as base models, improving passage ranking accuracy for morphologically complex text. (ii) *The first systematic benchmark for Amharic*: we evaluate both sparse and dense retrieval models on Amharic, establishing strong baselines for future research. (iii) *A language-specific vs. multilingual analysis*: we show that Amharic-optimized models consistently outperform multilingual embeddings, underscoring the value of language-specific adaptation. (iv) *A public benchmark dataset*: We repurpose the Amharic News Text Classification Dataset (AM-NEWS) by treating headlines as queries and corresponding articles as passages, creating MS MARCO-style query-passage pairs with heuristic relevance labels. This enables reproducible evaluation of passage ranking models for Amharic. We refer to this processed version as the *Amharic Passage Retrieval Dataset*. The dataset is publicly available on Hugging Face,[2] and all code and preprocessing scripts are released on GitHub.[3]

2 Motivation

Recent studies highlight systemic shortcomings in low-resource language technologies, leading to retrieval failures, biased outputs, and exposure to harmful or policy-violating content (Shen et al., 2024; Nigatu and Raji, 2024). For example, Nigatu and Raji (2024) find that Amharic-speaking YouTube users frequently encounter such content due to retrieval systems misinterpreting user intent behind benign queries. These errors stem from foundational limitations in information retrieval (IR) systems, which are optimized for high-resource languages like English and struggle with morphologically complex languages like Amharic. The consequences extend beyond search engines: Sewunetie et al. (2024) demonstrate that retrieval failures in machine translation propagate gender bias, defaulting Amharic occupational terms to male forms even when the context is gender-neutral. Such errors reflect broader research gaps in NLP, where systems disproportionately prioritize high-resource languages, thereby exacerbating inequities faced by underrepresented linguistic communities (Shen et al., 2024).

Amharic, the working language of Ethiopia's federal government and one of the most widely spoken Semitic languages (Gezmu et al., 2018), presents unique challenges for IR. Its root-based templatic morphology allows a single root to generate numerous derived forms through affixation and vowel pattern changes. These morphological variations, combined with the Ge'ez script, an Abugida writing system with 33 base characters and over 230 syllabic forms, make Amharic structurally and morphologically distinct from Indo-European and other high-resource languages. As a result, conventional retrieval models tend to underperform without language-specific adaptation. Addressing these challenges requires Amharic-specific embedding models tailored for passage retrieval. While recent efforts (Belay et al., 2021; Azime et al., 2024b) have advanced Amharic NLP, their primary focus

[2]https://huggingface.co/datasets/rasyosef/amharic-news-retrieval-dataset

[3]https://github.com/kidist-amde/amharic-ir-benchmarks/tree/main/data

has not been on optimizing retrieval performance.

Our work fills this gap by developing and benchmarking retrieval methods specifically adapted to Amharic's linguistic characteristics, laying a foundation for more equitable and semantically accurate information access in low-resource language settings.

3 Related Work

Retrieval systems commonly adopt a two-stage pipeline to optimize efficiency and effectiveness: (i) First-stage retrieval efficiently retrieves candidate documents using lightweight methods such as sparse or dense retrieval. (ii) Re-ranking refines the results using computationally more intensive models, such as cross-encoders.

Sparse retrieval. Sparse retrieval is fundamental in IR, with BM25 known for its efficiency, interpretability, and cross-domain robustness (Robertson and Zaragoza, 2009). However, it struggles with vocabulary mismatch and morphological variability, challenges that are particularly acute in morphologically rich languages like Amharic. Learned sparse retrieval (LSR) methods (Formal et al., 2021b,a) attempt to mitigate these issues by dynamically weighting and expanding terms, thereby enhancing relevance while maintaining interpretability (Dai and Callan, 2020). However, LSR faces limitations in low-resource settings due to the scarcity of annotated data, dialectal diversity, and morphological complexity (e.g., Amharic's templatic morphology), which necessitate subword-aware tokenization or morphological analyzers that are often unavailable.

Dense retrieval. Dense retrieval encodes queries and documents into a shared semantic space using neural network encoders, enabling efficient retrieval via approximate nearest neighbor (ANN) search based on embedding similarity (Johnson et al., 2019; Karpukhin et al., 2020; Xiong et al., 2021). While it helps mitigate lexical mismatch, its effectiveness in low-resource languages is hindered by the need for large-scale labeled training data. Multilingual models such as mBERT (Pires et al., 2019), XLM-R (Conneau et al., 2020), and African language-specific models like SERENGETI (Adebara et al., 2023) and AfriBERTa (Ogueji et al., 2021) partially address data scarcity through cross-lingual pretraining. However, their effectiveness in morphologically complex languages like Amharic has not been thoroughly investigated.

Recent advances in unsupervised contrastive learning, such as Contriever (Izacard et al., 2022), have demonstrated strong zero-shot and multilingual retrieval performance, especially in cross-lingual transfer scenarios. Nonetheless, their effectiveness in morphologically complex languages like Amharic remains unexplored, as current evaluations do not account for challenges arising from root-based and templatic morphologies.

Beyond data scarcity, retrieval performance is further constrained by morphological complexity and tokenization challenges. Amharic's templatic morphology often causes standard subword tokenizers to over-segment words into non-morphemic units, leading to fragmented representations that obscure semantic relationships. Broader research on multilingual tokenization quality (Rust et al., 2021) shows that excessive segmentation in morphologically rich languages introduces noise into subword representations, degrading performance in downstream tasks.

Despite recent advances in multilingual dense retrieval, state-of-the-art models such as Arctic Embed 2.0 (Yu et al., 2024) and Multilingual E5 (Wang et al., 2024), which topped the *MTEB Embedding Leaderboard*[4] at the time of our study, continue to struggle with highly inflected languages. These models often produce suboptimal tokenizations, fragmented subword representations, and inefficient embeddings, ultimately limiting their retrieval effectiveness. Our empirical findings in Section 6.3 illustrate the extent to which tokenization errors impair retrieval performance in Amharic.

Bridging the gap in Amharic IR. Retrieval systems are primarily optimized for high-resource languages, exacerbating performance disparities in low-resource settings like Amharic (Nigatu and Raji, 2024). Prior research in Amharic IR has explored pre-trained embeddings (Word2Vec, fastText, AmRoBERTa, Belay et al., 2021), morphological tools (e.g., annotation frameworks, WordNet-based query expansion, Yeshambel et al., 2021), and cross-lingual transfer via multilingual models (AfriBERTa, Azime et al., 2024a). However, systematic evaluations of sparse and dense retrieval architectures remain absent, making principled comparisons difficult and leaving the effectiveness of different paradigms in Amharic IR largely unexamined.

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

Yeshambel et al. (2020) introduce 2AIRTC, a TREC-style test collection for standardized Amharic IR evaluation, but it lacks baseline retrieval benchmarks and complete relevance judgments, making recall-based assessments unreliable. To ensure robust evaluation, we conduct our main experiments on the Amharic Passage Retrieval Dataset, which we derive by preprocessing the Amharic News Text Classification Dataset (AM-NEWS) (Azime and Mohammed, 2021) into MS MARCO-style query-passage pairs (see Section 5). A detailed analysis of 2AIRTC, its limitations, and our supplementary evaluations on this dataset is provided in Appendix A.

To address these gaps, our work introduces Amharic-specific retrieval models that incorporate both strong and compact encoder backbones (Section 4.2), optimized using contrastive training to better handle Amharic's morphological complexity. We also develop and evaluate a late-interaction ColBERT model tailored for Amharic, and benchmark both sparse and dense retrieval architectures. This enables rigorous, reproducible comparisons across retrieval paradigms.

4 Methodology

In this section, we outline our approach to Amharic dense retrieval. We begin by reviewing dense retrieval and ColBERT architectures, which underpin our framework. We then introduce our Amharic embedding models, describing their architecture, training setup, and optimization strategy.

4.1 Preliminaries

Dense retrieval models

Dense retrieval maps queries and passages into a shared vector space using transformer-based encoders (Karpukhin et al., 2020). Given a query q and a set of candidate passages $P = \{p_1, p_2, ..., p_N\}$, a dense retrieval model maps each input to a fixed-length vector representation via a transformer encoder $\text{Enc}(\cdot)$:

$$q_{\text{enc}} = \text{Enc}_Q(q), \quad p_{\text{enc}} = \text{Enc}_P(p) \tag{1}$$

The relevance score between a query q and a passage p is computed using a similarity function $f(q, p) = \text{sim}(q_{\text{enc}}, p_{\text{enc}})$, where $\text{sim}(\cdot, \cdot)$ typically denotes the dot product or cosine similarity.

ColBERT: Late interaction retrieval

ColBERT (Khattab and Zaharia, 2020) enhances retrieval by preserving token-level interactions between queries and passages. Rather than aggregating inputs into a single vector, it encodes:

$$q_{\text{enc}} = [\mathbf{h}_q^1, \mathbf{h}_q^2, \ldots, \mathbf{h}_q^m], \; p_{\text{enc}} = [\mathbf{h}_p^1, \mathbf{h}_p^2, \ldots, \mathbf{h}_p^n] \tag{2}$$

where $\mathbf{h}_q^i$ and $\mathbf{h}_p^j$ are contextualized token embeddings. Relevance is computed using maximum similarity pooling:

$$f(q, p) = \sum_{i=1}^{m} \max_{j \in \{1, \ldots, n\}} \text{sim}(h_q^i, h_p^j). \tag{3}$$

This allows fine-grained token-level matching while remaining efficient at inference time.

4.2 Amharic Text Embedding Models

We design three transformer-based dense retrieval models for Amharic, each with different parameter sizes. All models use a context length of `512` tokens to balance effectiveness and efficiency.

(1) **RoBERTa-Base-AM-Embed** (`110M` parameters): A 12-layer transformer with hidden size `768`, based on XLM-RoBERTa (Conneau et al., 2020). This model offers deep contextualized representations while remaining compatible with standard retrieval pipelines.

(2) **RoBERTa-Medium-AM-Embed** (`42M` parameters): A compact 8-layer transformer with hidden size `512`, optimized for retrieval latency and resource-constrained environments.

(3) **BERT-Medium-AM-Embed** (`40M` parameters): Based on the original BERT architecture (Devlin et al., 2019), with 8 layers and hidden size `512`. This model is suited for latency-sensitive applications.

Embedding Vector Generation: To obtain passage embeddings, we apply the following steps to the last hidden states of the pre-trained Amharic base models:

(i) **Mean pooling:** Aggregate token embeddings to form a fixed-length vector:

$$\mathbf{h}_{\text{pool}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{h}_t$$

where T is the sequence length.

(ii) **L2 normalization:** Normalize embeddings to unit length for cosine similarity:

$$\mathbf{h}_{\text{norm}} = \frac{\mathbf{h}_{\text{pool}}}{\|\mathbf{h}_{\text{pool}}\|_2}$$

Training setup. All models are initialized from Amharic pre-trained checkpoints (Amharic BERT and RoBERTa) and fine-tuned using contrastive learning with in-batch negatives on a corpus of 45K Amharic query-passage pairs. Models are trained for 4 epochs using the AdamW optimizer (lr = 5e-5) with cosine learning rate decay. We evaluate using MRR, NDCG, and Recall@K. Passages longer than 512 tokens are truncated. Additional implementation details are in Section 5.2.

Multiple negatives ranking loss (MNRL). Following (Reimers and Gurevych, 2019), we use in-batch negatives to train our models. For a batch of queries $\{\mathbf{q}_i\}_{i=1}^B$, their corresponding positives $\{\mathbf{p}_i^+\}_{i=1}^B$, and in-batch negatives $\mathcal{N}_i = \{\mathbf{p}_j\}_{j \neq i}$, the loss $\mathcal{L}$ is:

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(f(\mathbf{q}_i,\mathbf{p}_i^+))}{\exp(f(\mathbf{q}_i,\mathbf{p}_i^+)) + \sum_{\mathbf{p}_j^- \in \mathcal{N}_i} \exp(f(\mathbf{q}_i,\mathbf{p}_j^-))} \quad (4)$$

This loss encourages the model to assign higher similarity scores to the relevant passages $\mathbf{p}_i^+$ relative to the in-batch negatives $\mathcal{N}_i$, promoting discriminative representations in the shared embedding space.

5 Experimental Setup

5.1 Training Data

We conduct our experiments using the Amharic Passage Retrieval Dataset, which we construct by preprocessing the Amharic News Text Classification Dataset (AMNEWS) (Azime and Mohammed, 2021). The original dataset contains 50,706 Amharic news articles categorized into six domains: Local News, Sports, Politics, International News, Business, and Entertainment. To simulate real-world retrieval scenarios, we treat article headlines as queries and the corresponding article bodies as passages. As the dataset lacks explicit relevance judgments, we adopt a heuristic supervision approach: each headline is assumed to be relevant to its associated article. To validate this assumption, we manually examined a random subset of query-passage pairs and confirmed high topical alignment between headlines and their articles. We also removed duplicates using MD5 hashing and reformatted the data into an MS MARCO-style passage retrieval format. This results in approximately 45K query-passage pairs. We split the dataset into training and test sets, reserving 10% for evaluation. The split is stratified by category to ensure balanced representation across all six news domains.

5.2 Implementation Details

Amharic embedding models. We trained our Amharic embedding models on a single A100 40GB GPU for 4 epochs using the Sentence Transformer Trainer from the `sentence-transformers` Python library.[5] Training was performed with a learning rate of 5e-5, batch size 128, cosine learning rate scheduler, and the multiple negatives ranking loss (MNRL) for optimization.

Sparse retrieval baselines. For BM25-based retrieval, we used the BM25Retriever from the `LlamaIndex` framework.[6]

Dense retrieval baseline. We implemented ColBERT using the `PyLate` library (Chaffin and Sourty, 2024),[7] adapting it for Amharic using the *RoBERTa-Medium-Amharic* encoder model. The model was trained with a learning rate of 1e-5 and batch size 32, using eight negative samples drawn from the top 150 passages retrieved by our *RoBERTa-Medium-Amharic-Embed model.*

Fine-tuning multilingual models. We fine-tuned the Snowflake-Arctic-Embed model on Amharic query–passage pairs for 4 epochs using the AdamW optimizer with a learning rate of 2e-5, batch size 128, and a linear warmup ratio of 0.1. We applied a weight decay of 0.01 and used a cosine scheduler with warmup.

Evaluation metrics. We evaluate retrieval effectiveness using standard ranking metrics in IR: (i) MRR@k (mean reciprocal rank): evaluates the average inverse rank of the first relevant passage. (ii) NDCG@k (normalized discounted cumulative gain): assesses ranking quality with graded relevance and logarithmic position discounting; in our case, it is computed using binary relevance labels. (iii) Recall@k: measures how often relevant passages appear within the top-k retrieved results.

6 Experimental Evaluation and Results

In this section we present our empirical evaluation, which is structured around the following research questions:

RQ1 How well do Amharic-optimized embeddings improve ranking accuracy compared to general-purpose multilingual embedding models? (Section 6.1)

[5] https://pypi.org/project/sentence-transformers/

[6] https://docs.llamaindex.ai/en/stable/examples/retrievers/bm25_retriever/

[7] https://github.com/lightonai/pylate

	Model	Params	MRR@10	NDCG@10	Recall		
					@10	@50	@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**†

Table 1: Performance comparison on the Amharic Passage Retrieval Dataset between our Amharic-optimized embedding models and state-of-the-art multilingual dense retrieval baselines, all based on a bi-encoder architecture. The multilingual models *snowflake-arctic-embed-l-v2.0* and *multilingual-e5-large-instruct* originate from Arctic Embed 2.0 (Yu et al., 2024) and Multilingual E5 Text Embeddings (Wang et al., 2024), respectively. Best results are shown in **bold**. Statistically significant improvements ($p < 0.05$) over the strongest multilingual baseline are marked with †, based on a paired t-test.

RQ2 How do different retrieval paradigms compare in effectiveness, establishing a benchmark for Amharic passage retrieval? (Section 6.2)

RQ3 How does tokenization quality, particularly subword segmentation, impact retrieval effectiveness in morphologically rich, low-resource languages like Amharic? (Section 6.3)

RQ4 To what extent does the base model size influence retrieval performance in late interaction models for low-resource settings like Amharic? (Section 6.4)

6.1 Evaluating Amharic Embeddings Against Multilingual Baselines

We investigate whether Amharic-optimized embedding models offer tangible advantages over general-purpose multilingual models in ranking Amharic passages. Table 1 compares three Amharic-specific models with four multilingual baselines using standard IR metrics. Across the board, Amharic-optimized models outperform multilingual counterparts, often with fewer parameters. The best-performing multilingual model, *Snowflake-Arctic-Embed* (568M parameters), achieves 0.659 MRR@10, whereas *RoBERTa-Base-Amharic-Embed* (110M parameters) reaches 0.775, reflecting a 17.6% relative gain. Similar improvements are observed in NDCG@10 (0.808 vs. 0.701) and Recall@10 (0.913 vs. 0.831), demonstrating consistent gains across top- and mid-rank positions. Notably, *RoBERTa-Medium-Amharic-Embed* (42M) outperforms all multilingual models in MRR@10 and Recall@10 despite being over 13× smaller than *Snowflake-Arctic-Embed*. This finding underscores that scaling multilingual models does not necessarily translate into better retrieval performance for low-resource languages.

These findings emphasize three key insights: (i) Tokenization alignment matters: Amharic-optimized models better preserve word boundaries, reducing subword fragmentation and improving semantic matching (see Section 6.3). (ii) Parameter efficiency matters: Amharic-specific models achieve superior performance with significantly fewer parameters. (iii) Language-specific adaptation outperforms brute-force scaling: Fine-tuning on monolingual data yields greater benefit than applying large multilingual encoders out-of-the-box.

6.2 Benchmarking Sparse vs. Dense Retrieval for Amharic IR

We compare sparse and dense retrieval paradigms to establish strong baselines for Amharic passage retrieval. As shown in Table 2: (i) BM25 serves as a competitive sparse baseline, achieving 0.657 MRR@10 and 0.774 Recall@10, reaffirming its relevance in low-resource settings. (ii) Dense retrieval models outperform this baseline across all evaluation metrics. The bi-encoder model *RoBERTa-Base-Amharic-Embed* improves upon BM25 with 0.775 MRR@10 and 0.913 Recall@10, highlighting the benefits of Amharic-specific embeddings. Its Recall@100 score of 0.979 also indicates strong coverage across larger candidate sets. (iii) The best-performing system is *ColBERT-RoBERTa-Base-Amharic*, a late interaction model built on the same Amharic encoder. By incorporating token-level interactions, it significantly enhances precision, achieving 0.843 MRR@10 and 0.939 Recall@10, a 28.31% relative improvement in MRR over BM25. It also surpasses the bi-encoder at top and mid ranks (e.g., Recall@50: 0.972 vs. 0.964), while maintaining parity at Recall@100 (0.979). These results highlight the complementary strengths of Amharic-specific encoders and interaction-aware architectures. Overall, these findings demonstrate the effectiveness of dense retrieval methods, particularly late interaction models like ColBERT, when paired with language-specific pretraining. Both dense systems benefit from Amharic-optimized encoders, underscoring the importance of tailoring retrieval architectures to the linguistic characteristics of morphologically rich, low-resource languages.

6.3 Tokenization Quality and Retrieval Performance

This section investigates how tokenization quality, particularly subword segmentation, impacts retrieval effectiveness in morphologically rich, low-resource languages, using Amharic as a case study. We focus on subword fertility, defined as the average number of subword tokens per word (Pietra et al., 1997), as a key indicator of tokenization quality. Figure 1 presents fertility scores across various embedding models, based on a representative subset of `10k` Amharic passages.

Excessive subword segmentation (i.e., high fertility) increases computational overhead and fragments semantic representations, which degrades retrieval accuracy (Ali et al., 2024). For example: (i) *gte-modernbert-base* exhibits the highest fertility (13.80) and the weakest retrieval performance (MRR@10 = 0.019), demonstrating the detrimental effects of poor tokenization. In contrast, Amharic-optimized models such as *RoBERTa-Base-Amharic-Embed* achieve the lowest fertility (1.46) and the highest MRR@10 (0.775), indicating better alignment between tokenization and linguistic structure. (ii) Among multilingual models, *snowflake-arctic-embed-l-v2.0* demonstrates moderate fertility (2.35) and the best performance in its category (MRR@10 = 0.659), likely benefiting from its large parameter size (`568M`). However, it still underperforms relative to much smaller Amharic-specific models, suggesting that model size alone cannot compensate for tokenization inefficiencies.

These findings are consistent with prior work (Toraman et al., 2023; Ali et al., 2024), reinforcing the critical role of tokenizer alignment, particularly in morphologically complex languages, in improving computational efficiency and downstream retrieval performance. To further illustrate this issue, Figure 2 presents a qualitative comparison of subword tokenization for a representative Amharic sentence. We contrast the segmentation behavior of the best-performing Amharic-specific model (*RoBERTa-Base-Amharic-Embed*) with that of the strongest multilingual model (*snowflake-arctic-embed-l-v2.0*). The Amharic-specific model generates fewer and more linguistically coherent tokens, which likely contributes to its superior retrieval performance.

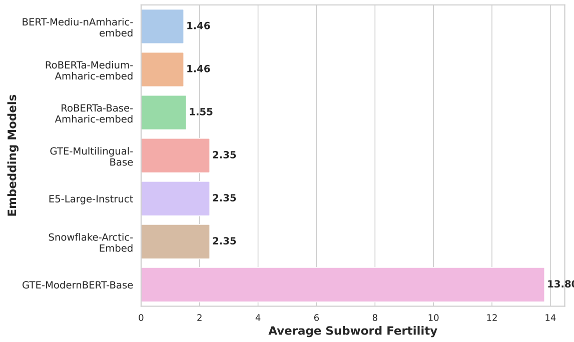

Figure 1: Average subword fertility across embedding models. Lower fertility indicates better alignment with word boundaries, while higher fertility suggests excessive segmentation, which can harm retrieval accuracy.

Tokenizer	Tokenized Output	Subword Count	Subword Fertility
RoBERTa-Base-Amharic-embed	['▁የ', 'ተጃጋገመ', 'ው', '▁የመሬት', '▁መንቀጥቀጥ', 'ና', '▁የእ', 'ሳተ', '▁ገሞራ', '▁ምልክት', '▁በአፋር', '▁ክልል']	12	1.50
snowflake-arctic-embed-l-v2.0	['▁የተ', 'ጃጋ', 'ገ', 'መው', '▁የመ', 'ሬት', '▁መን', 'ቀጥ', 'ቀጥ', 'ና', '▁የእ', 'ሳተ', '▁ገ', 'ሞ', 'ራ', '▁ምልክት', '▁በአ', 'ፋር', '▁ክልል']	19	2.38

Figure 2: Subword tokenization comparison for a representative Amharic sentence. *RoBERTa-Base-Amharic-Embed* produces more compact and linguistically meaningful tokens than *snowflake-arctic-embed-l-v2.0*, reducing subword fragmentation and improving semantic representation quality.

6.4 Model Size vs. Performance in Late Interaction Retrieval

We investigate the trade-off between model size and retrieval effectiveness by comparing three Amharic encoder models within a late interaction framework using ColBERT: *BERT-Medium-Amharic*, *RoBERTa-Medium-Amharic*, and *RoBERTa-Base-Amharic*. Figure 3 summarizes performance across five retrieval metrics, highlighting how encoder size influences ranking accuracy and recall in

Type	Model	MRR@10	NDCG@10	Recall		
				@10	@50	@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

Table 2: Performance of retrieval models on the Amharic Passage Retrieval Dataset. *ColBERT-RoBERTa-Base-Amharic* is a late interaction model that builds on the *RoBERTa-Base-Amharic-Embed* encoder. Best results are shown in **bold**. Statistically significant improvements ($p < 0.05$) over the strongest baseline are marked with †, based on a paired t-test.

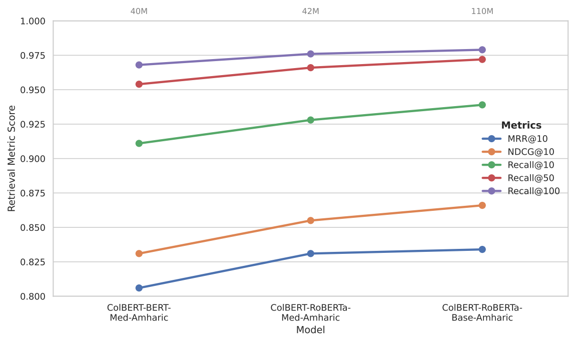

Figure 3: Effect of base model size on ColBERT performance in Amharic passage retrieval. The figure presents retrieval effectiveness across five ranking metrics for ColBERT models initialized with different Amharic base encoders. Lines connect performance metrics per model to highlight comparative trends.

Amharic passage retrieval.

(i) *ColBERT-RoBERTa-Base-Amharic* (`110M`) achieves the best overall performance (MRR@10: 0.843, NDCG@10: 0.866, Recall@10: 0.939), suggesting that scaling up the encoder benefits token-level retrieval, likely due to increased representational capacity. (ii) *RoBERTa-Medium-Amharic* (`42M`) remains highly competitive (MRR@10: 0.831, Recall@10: 0.928), achieving a 1.5% relative performance difference from its larger counterpart while being 62% smaller, demonstrating strong efficiency in resource-constrained scenarios. (iii) *BERT-Medium-Amharic* (`40M`) also performs strongly (MRR@10: 0.806), showing that compact models remain viable for retrieval in low-resource settings.

While larger models boost ColBERT's performance, well-optimized medium-sized encoders strike a more favorable balance between accuracy and efficiency, making them ideal for compute-constrained, low-resource settings.

6.5 Fine-Tuning Multilingual Models with Amharic Supervision

While our primary comparison focuses on zero-shot multilingual models, we also investigate the impact of retrieval-specific supervised fine-tuning. To this end, we fine-tune the strongest multilingual baseline, *Snowflake-Arctic-Embed* (`568M` parameters), using Amharic query–passage pairs. The resulting model, *snowflake-arctic-embed-l-v2.0-AM*, shows substantial performance improvements: MRR@10 increases from 0.659 to 0.827, and Recall@10 rises from 0.831 to 0.942 (Table 3).

These results highlight two key insights: (i) Even large, multilingual embedding models are suboptimal for low-resource retrieval tasks when used out-of-the-box. (ii) Retrieval-specific supervision with in-language data significantly improves ranking effectiveness, especially at top ranks (MRR@10: +25.5%). This underscores the importance of task-aligned and language-specific adaptation. Notably, retrieval fine-tuning enhances semantic alignment more effectively than general-purpose multilingual pretraining, even without modifying the underlying architecture.

6.6 Key Challenges in Amharic Passage Retrieval

While Table 1 shows that Amharic-optimized models like *RoBERTa-Base-Amharic-Embed* consistently outperform multilingual baselines, several persistent challenges reveal the underlying complexity of Amharic IR: (i) Morphological complexity: Amharic's templatic morphology results in diverse word forms. Despite improved tokenization in language-specific models, subword over-segmentation, especially for inflected or compound words, still fragments semantics and limits retrieval accuracy. (ii) Data scarcity: Amharic models are pretrained on just `300M` tokens, far fewer than for high-resource languages. This restricts generalization, particularly for rare terms or specialized domains, and contributes to residual retrieval errors even in strong models. (iii) Evaluation noise: The Amharic passage retrieval dataset lacks human-annotated relevance labels, relying instead on

Model	MRR@10	NDCG@10	Recall		
			@10	@50	@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**†

Table 3: Effect of Amharic-specific fine-tuning on multilingual retrieval performance. *snowflake-arctic-embed-l-v2.0-AM* denotes the fine-tuned variant trained on Amharic passage-level supervision. † indicates statistically significant improvements ($p < 0.05$) over the zero-shot version, based on a paired t-test.

headline–article pairs as heuristic signals. While practical, this weak supervision introduces noise and limits the granularity of relevance modeling. (iv) Qualitative observations: Manual inspection of top-ranked outputs shows that Amharic-optimized dense models generally retrieve more contextually appropriate content. However, even the best models struggle with negation, temporal shifts, and nuanced entailment. For instance, given the query *"Was the planned protest not held?"*, the model retrieved a passage stating *"The planned protest was held,"* ranking it highly despite the semantic contradiction. Sparse models, by contrast, often favor surface-level keyword overlap (e.g., matching on *"protest"*), yet fail to account for polarity or temporal context. These observations highlight that retrieval effectiveness still hinges on capturing deeper semantic and discourse-level nuances, an open challenge in low-resource settings.

These challenges are further discussed in the limitations (Section 8) and illustrated with qualitative error analysis in (Appendix B.2), highlighting fundamental issues in low-resource IR and emphasizing the need for better tokenization, richer training corpora, and curated evaluation benchmarks.

7 Conclusion

We introduced dense retrieval models and established the first systematic benchmark for Amharic passage retrieval. Our models consistently outperform multilingual baselines, underscoring the importance of linguistic adaptation for morphologically rich, low-resource languages. We also show that tokenization quality, especially subword fertility, significantly impacts retrieval performance: compact segmentations improve ranking accuracy, while over-segmentation harms semantic alignment. Our main experiments use the Amharic Passage Retrieval Dataset (derived from AMNEWS using heuristic labels), and we include supplementary results on 2AIRTC in the appendix.

However, both datasets present evaluation challenges: the former lacks gold-standard relevance judgments, and the latter has incomplete labeling. These limitations underscore the need for more robust evaluation resources and motivate future research directions.

To address these gaps, future work should focus on: (i) designing morphology-aware or byte-level tokenizers tailored to Amharic's templatic structure, (ii) improving training with hard negative mining and curriculum-based strategies, and (iii) extending evaluation to document-level and multi-hop retrieval. Creating a high-quality, human-annotated benchmark with expert-labeled relevance, dialect variation, and morphological features, through collaboration with local institutions will be critical for aligning IR systems with real-world Amharic information needs.

Acknowledgements

We are grateful to our reviewers for their thoughtful comments and insightful feedback, which helped us improve the quality of this work.

This work was partially supported by the Dutch Research Council (NWO) under project EINF-9550, with computations performed on the Snellius supercomputer (SURF). Additional support was provided by the Dutch Research Council (NWO) under project numbers 024.004.022, NWA.1389.20.183, and KICH3.LTP.20.006; and the European Union's Horizon Europe program under grant agreement No 101070212.

The content of this paper reflects the views of the authors and does not necessarily represent the official position of their affiliated institutions or sponsors.

8 Limitations

Dataset and evaluation. We rely on the Amharic Passage Retrieval Dataset (derived from AMNEWS), which lacks human-annotated relevance judgments. Our assumption that headlines reflect document relevance introduces weak supervision noise. Furthermore, the dataset's limited scale constrains generalizability. Future work should consider collecting crowd-sourced labels or leveraging Amharic language models for automatic annotation to enhance evaluation fidelity.

Pretraining data. Our Amharic base models were pre-trained on a relatively modest corpus of 300 million tokens from web, news, and social media sources. This is substantially smaller than the corpora used for high-resource language, e.g., English BERT (3.3B) and RoBERTa (30B). Such data limitations may affect model generalization and downstream retrieval performance.

Domain generalization. The main experiments were conducted within the news domain. The effectiveness of our retrieval models in other domains (e.g., medical, legal, or technical) remains untested and would likely require further domain adaptation.

Tokenization and morphology. Amharic's templatic morphology poses tokenization challenges, which we analyze using subword fertility. However, our models do not incorporate explicit morphological analyzers, lemmatizers, or segmentation tools. Instead, we rely on standard tokenization and language-specific fine-tuning. Tokenization inconsistencies introduce over-segmentation, degrading semantic coherence and retrieval accuracy. These limitations open avenues for future work, including the integration of morphology-aware tokenizers, hybrid word–subword representations, and explicit linguistic preprocessing pipelines.

Fine-tuning strategy. We employed full-parameter fine-tuning to maximize retrieval effectiveness in our monolingual Amharic setup, where preserving multilingual capabilities was not a priority. While this approach yields strong performance, future work should explore parameter-efficient alternatives such as LoRA or lightweight adapters, especially in cross-lingual settings where model compactness and multilingual retention are essential.

9 Ethical Considerations

Our study aims to improve passage retrieval for Amharic, a low-resource language. While our models show substantial performance gains, we acknowledge potential ethical concerns regarding data biases, fairness, and deployment risks.

Use of publicly available data. We use two public datasets: AMNEWS (Azime and Mohammed, 2021), comprising news articles, and 2AIRTC (Yeshambel et al., 2020), a TREC-style IR dataset. All data is publicly available, and no new data was collected, ensuring compliance with ethical standards.

Base models and pretraining data. Our Amharic embeddings are derived from models pre-trained on `300M` tokens of publicly accessible Amharic web, news, and tweet data. We use existing checkpoints from Hugging Face and rely on their accompanying documentation for data provenance.

Bias and fairness considerations. Like many datasets sourced from online news content, the AMNEWS dataset may contain inherent biases related to reporting styles, topic framing, and regional representation. Retrieval models trained on this dataset may inherit and reflect these biases, particularly for politically or socially sensitive topics. While our study does not explicitly mitigate bias, we recognize this as an important challenge and encourage future work on fairness-aware retrieval and debiasing strategies.

Algorithmic challenges in low-resource languages. Amharic is a low-resource, morphologically rich language, making it susceptible to algorithmic disparities due to data sparsity and tokenization challenges. While we highlight these issues, our approach does not introduce direct mitigation techniques beyond language-specific fine-tuning. Future work should explore improved tokenization and linguistic adaptation methods to enhance retrieval fairness.

Responsible deployment and transparency. We follow ACL's ethical standards and stress that models should not be deployed in high-stakes applications without rigorous auditing. We support transparency in sharing model limitations and advocate for careful, informed use of our publicly released models and datasets.

We encourage the community to use our models and datasets responsibly, and to continue advancing equitable IR systems that serve linguistically diverse users.

References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2023.

SERENGETI: Massively multilingual language models for Africa. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1498–1537, Toronto, Canada. Association for Computational Linguistics.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Buschhoff, Charvi Jain, Alexander Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, Stefan Kesselheim, and Nicolas Flores-Herr. 2024. Tokenizer choice for LLM training: Negligible or crucial? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3907–3924, Mexico City, Mexico. Association for Computational Linguistics.

Israel Abebe Azime, Mitiku Yohannes Fuge, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Aman Kassahun Wassie, Eyasu Shiferaw Jada, Yonas Chanie, Walelign Tewabe Sewunetie, and Seid Muhie Yimam. 2024a. Enhancing Amharic-LLaMA: Integrating task specific and generative datasets. *arXiv preprint arXiv:2402.08015*.

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

Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Mitiku Yohannes Fuge, Aman Kassahun Wassie, Eyasu Shiferaw Jada, Yonas Chanie, Walelign Tewabe Sewunetie, and Seid Muhie Yimam. 2024b. Walia-LLM: Enhancing Amharic-LLaMA by integrating task-specific and generative datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 432–444, Miami, Florida, USA. Association for Computational Linguistics.

Tadesse Destaw Belay, Abinew Ayele, and Seid Muhie Yimam. 2021. The development of pre-processing tools and pre-trained embedding models for Amharic. In *Proceedings of the Fifth Workshop on Widening Natural Language Processing*, pages 25–28, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Fernando Campos, Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, Li Deng, and Bhaskar Mitra. 2016. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Antoine Chaffin and Raphaël Sourty. 2024. Pylate: Flexible training and retrieval for late interaction models.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yiqun Liu, Yixing Fan, and Xueqi Cheng. 2023. A unified generative retriever for knowledge-intensive language tasks via prompt learning. In *SIGIR*. ACM.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Zhuyun Dai and Jamie Callan. 2020. Context-aware term weighting for first stage passage retrieval. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 1533–1536.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021a. Splade v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021b. 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*, pages 2288–2292.

Andargachew Mekonnen Gezmu, Binyam Ephrem Seyoum, Michael Gasser, and Andreas Nürnberger. 2018. Contemporary Amharic corpus: Automatically morpho-syntactically tagged Amharic corpus. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 65–70, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. In *TMLR*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

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*, page 39–48, New York, NY, USA. Association for Computing Machinery.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Donald Metzler, Yi Tay, and Dara Bahri. 2021. Rethinking search. *ACM SIGIR Forum*, 55:1 – 27.

Hellina Hailu Nigatu and Inioluwa Deborah Raji. 2024. "i searched for a religious song in amharic and got sexual content instead": Investigating online harm in low-resourced languages on youtube. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, Rio de Janeiro, Brazil. ACM.

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

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with BERT. *arXiv preprint arXiv:1910.14424*.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? No problem! Exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stephen Della Pietra, Mark Epstein, Salim Roukos, and Todd Ward. 1997. Fertility models for statistical natural language understanding. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98/EACL '98, page 168–173, USA. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Ronak Pradeep, Kai Hui, Jai Gupta, Adam Lelkes, Honglei Zhuang, Jimmy Lin, Donald Metzler, and Vinh Tran. 2023. How does generative retrieval scale to millions of passages? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1305–1321, Singapore. Association for Computational Linguistics.

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*. Association for Computational Linguistics.

Stephen E. Robertson and Steve Walker. 1997. On relevance weights with little relevance information. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 16–24, New York, NY, USA. Association for Computing Machinery.

Stephen E. Robertson and Hugo Zaragoza. 2009. *The Probabilistic Relevance Framework: BM25 and Beyond*. Foundations and Trends in Information Retrieval. NOW Publishers.

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.

Walelign Sewunetie, Atnafu Tonja, Tadesse Belay, Hellina Hailu Nigatu, Gashaw Gebremeskel, Zewdie Mossie, Hussien Seid, and Seid Yimam. 2024. Gender bias evaluation in machine translation for Amharic, Tigrigna, and Afaan Oromoo. In *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 1–11, Sheffield, United Kingdom. European Association for Machine Translation (EAMT).

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of LLMs in multilingual contexts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2668–2680, Bangkok, Thailand. Association for Computational Linguistics.

Yi Tay, Vinh Quang Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. 2022. Transformer memory as a differentiable search index. In *NeurIPS*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Cagri Toraman, Eyup Halit Yilmaz, Furkan Şahinüc, and Oguzhan Ozcelik. 2023. Impact of tokenization on language models: An analysis for Turkish. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(4).

Ahmet Üstün, Gosse Bouma, and Gertjan van Noord. 2019. Cross-lingual word embeddings for morphologically rich languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1222–1228, Varna, Bulgaria. INCOMA Ltd.

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*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *ICLR*.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: BERT and beyond. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 1–4, 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.

Tilahun Yeshambel, Josiane Mothe, and Yaregal Assabie. 2021. Morphologically annotated Amharic text corpora. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2349–2355.

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

Qingcheng Zeng, Lucas Garay, Peilin Zhou, Dading Chong, Yining Hua, Jiageng Wu, Yikang Pan, Han Zhou, Rob Voigt, and Jie Yang. 2023. GreenPLM: cross-lingual transfer of monolingual pre-trained language models at almost no cost. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23.

Appendix

A 2AIRTC: Amharic Ad Hoc Information Retrieval Test Collection

2AIRTC (Yeshambel et al., 2020) is the first TREC-style test collection for Amharic Information Retrieval (IR), comprising `12,583` documents and `240` manually assessed search topics. Each topic includes a title, description, and narrative (in both Amharic and English), with relevance judgments provided in standard QREL format. The dataset spans diverse domains (e.g., news, religion, culture, politics) and includes full-length documents sourced from news outlets, Wikipedia, social media, and blogs.

Limitations of 2AIRTC. Despite its foundational role, 2AIRTC presents several limitations that restrict its utility for robust and reproducible evaluation:

(i) **Incomplete relevance judgments:** Many semantically relevant documents remain unjudged, particularly those retrieved by neural models relying on semantic similarity. This leads to underestimated performance, especially for recall-based metrics, and compromises evaluation reliability.

(ii) **Lack of standardized baselines:** The absence of published baselines or leaderboard comparisons limits reproducibility and makes it difficult to benchmark retrieval systems fairly across studies.

These limitations underscore the need for updated, high-coverage Amharic IR benchmarks with exhaustive annotations and unified evaluation protocols to ensure fair, consistent, and progress-driving comparisons in future research.

A.1 Generalization to 2AIRTC: Amharic-Specific vs. Multilingual Models

To assess the generalization capacity of retrieval models trained on the Amharic Passage Retrieval Dataset, we evaluate their zero-shot performance on 2AIRTC, the only publicly available TREC-style benchmark for Amharic ad hoc retrieval. Despite known limitations such as annotation sparsity, 2AIRTC provides a valuable secondary testbed to evaluate retrieval robustness beyond the news domain. Table 4 compares multilingual and Amharic-specific dense retrievers on this corpus.

Amharic-specific models, despite having significantly fewer parameters, demonstrate competitive generalization. For instance, *RoBERTa-Base-Amharic-embed* achieves 0.770 NDCG@100 and 0.910 Recall@200, just one point below the strongest multilingual baseline (*multilingual-e5-large-instruct*) while being over 5× smaller. This highlights the strength of compact, linguistically aligned models for retrieval in low-resource settings.

Interestingly, performance does not scale monotonically with model size. *gte-multilingual-base* (305M) outperforms the larger *snowflake-arctic-embed-l-v2.0* (568M), indicating that architecture and pretraining objectives can outweigh parameter count.

Key Findings:

(i) Language-specific models generalize effectively: Despite smaller model size, Amharic-optimized models closely match multilingual systems, offering efficient and scalable alternatives for retrieval in low-resource languages.

(ii) Cross-benchmark variance reveals sensitivity to evaluation design: Amharic-specific models outperform on the Amharic Passage Retrieval Dataset but achieve comparable rather than dominant performance on 2AIRTC. This reflects differences in domain and the impact of sparse or incomplete relevance annotations.

(iii) Dense models are disadvantaged by annotation sparsity: Dense retrievers rely on semantic similarity, often surfacing relevant but unjudged content. The incomplete supervision in 2AIRTC penalizes these models on recall-based metrics, underestimating their true effectiveness.

These results emphasize the utility of Amharic-specific models for retrieval in low-resource contexts, while also underscoring the need for more complete and semantically annotated benchmarks to fairly assess dense retrievers' performance across domains.

A.2 Impact of Fine-Tuning on Cross-Domain Generalization

To examine whether supervised fine-tuning improves cross-domain generalization, we evaluate *snowflake-arctic-embed-l-v2.0-AM*, a multilingual model fine-tuned on the Amharic Passage Retrieval Dataset, on the 2AIRTC benchmark without any additional adaptation.

Table 5 presents the results. The fine-tuned model improves recall at both @100 and @200, achieving the highest Recall@200 (0.923) with a +2.6 point gain. It also shows a statistically significant increase in NDCG@100 (0.795 vs. 0.781), though MRR@100 slightly decreases. These findings suggest that retrieval-specific supervision on Amharic queries may enhance semantic alignment even across structurally different corpora. However, given 2AIRTC's known limitations, such as sparse relevance annotations these results should be interpreted as indicative rather than conclusive.

A.3 ColBERT with Amharic-Specific Backbones on 2AIRTC

We report the retrieval performance of three ColBERT variants equipped with Amharic-specific encoder backbones on the 2AIRTC dataset. All models were trained on the Amharic Passage Retrieval Dataset and evaluated zero-shot on 2AIRTC. Table 6 summarizes results across standard ranking metrics. Due to known limitations in 2AIRTC, including incomplete relevance judgments and annotation sparsity, we refrain from drawing strong conclusions and present these results as indicative for completeness.

A.4 Toward Robust Benchmarks for Amharic Information Retrieval

Although this study provides strong baselines for Amharic dense retrieval, the limitations of 2AIRTC, particularly its small query pool (240 topics) and sparse, sometimes inconsistent relevance annotations, significantly hinder its utility for rigorous evaluation. These limitations especially penalize dense models, which often retrieve semantically relevant but unjudged documents, leading to underreported performance on recall-oriented metrics. To advance Amharic IR evaluation and support more reliable model development, we recommend the following future directions:

- **Refine and expand 2AIRTC:** Improve annotation quality and coverage through iterative assessments, leveraging expert review, crowdsourcing, or semi-automated labeling to address incompleteness and inconsistency.
- **Develop morphology-aware retrieval methods:** Introduce tokenization and matching techniques

Model	Params	MRR@100	NDCG@100	Recall @100	Recall @200
Multilingual Models					
gte-modernbert-base	149M	0.046	0.017	0.021	0.033
gte-multilingual-base	305M	0.879	0.749	0.790	0.865
multilingual-e5-large-instruct	560M	**0.905**	**0.808**	**0.853**	**0.911**
snowflake-arctic-embed-l-v2.0	568M	0.876	0.781	0.830	0.897
Ours					
BERT-Medium-Amharic-embed	40M	0.805	0.667	0.727	0.828
RoBERTa-Medium-Amharic-embed	42M	0.853	0.735	0.798	0.878
RoBERTa-Base-Amharic-embed	110M	$0.861^{\uparrow}$	0.770	0.830	$0.910^{\uparrow}$

Table 4: Performance comparison of Amharic-optimized and multilingual dense retrieval models, all based on a bi-encoder architecture, evaluated on the 2AIRTC dataset. The models snowflake-arctic-embed-l-v2.0 and multilingual-e5-large-instruct (Hugging Face model names) originate from Arctic Embed 2.0 (Yu et al., 2024) and Multilingual E5 Text Embeddings (Wang et al., 2024), respectively. The best-performing results are highlighted in **bold** and the second best in up-arrow $^{\uparrow}$

Model	**MRR@100**	**NDCG@100**	**Recall@100**	**Recall@200**
snowflake-arctic-embed-l-v2.0	**0.876**	0.781	0.830	0.897
snowflake-arctic-embed-l-v2.0-AM	0.865	**0.795**†	**0.856**†	**0.923**†

Table 5: Effect of Amharic domain-specific fine-tuning on cross-domain retrieval performance. *snowflake-arctic-embed-l-v2.0-AM* is fine-tuned on AMNEWS and evaluated on 2AIRTC. † indicates statistically significant improvements ($p < 0.05$) over the zero-shot baseline.

suited to Amharic's templatic morphology, such as lemmatization or hybrid subword–word representations.

- **Enhance query modeling:** Apply Amharic-specific language models for query expansion and pseudo-relevance feedback to mitigate vocabulary mismatch and improve semantic coverage.
- **Establish multi-dataset evaluation standards:** Benchmark systems across across diverse Amharic retrieval datasets to assess robustness and generalizability, enabling more comprehensive evaluations and reproducible progress.

We hope future efforts will establish larger, expert-annotated testbeds that capture Amharic's linguistic diversity, enabling more faithful and equitable IR system development.

B Amharic Passage Retrieval Dataset Limitations and Qualitative Error Analysis

B.1 Dataset Limitations

While Section A discusses 2AIRTC, here we focus on the Amharic Passage Retrieval Dataset used in our main experiments, constructed by pairing news headlines with their corresponding articles. Each headline is treated as a query and its article as a relevant passage. While these headlines often serve as effective proxies for user queries, they are inherently editorial and concise, crafted to capture attention rather than to reflect authentic information-seeking behavior. This introduces a distributional gap between training-time queries and real-world user intent, which may limit generalization to practical retrieval scenarios. Moreover, the dataset lacks explicit relevance judgments or user interaction signals (e.g., clicks, ratings). Negative examples are generated by sampling non-matching articles, but these may still be topically related or semantically similar. This can introduce label noise, weakening the learning signal during contrastive training. To address these gaps, future work should:

- Incorporate curated or user-derived queries (e.g., search logs or community Q&A),
- Employ better hard negative mining strategies, and
- Collect human-annotated relevance labels for robust evaluation.

Model	MRR		NDCG		Recall	
	@100	@200	@100	@200	@100	@200
ColBERT-BERT-Medium-Amharic	0.907	0.907	0.823	**0.842**†	0.880	**0.930**†
ColBERT-RoBERTa-Medium-Amharic	0.909	0.909	0.831	0.840	0.886	0.917
ColBERT-RoBERTa-Base-Amharic	**0.919**†	**0.919**	**0.834**†	0.838	**0.887**†	0.906

Table 6: Retrieval performance of ColBERT models trained with different Amharic encoder backbones, evaluated at @100 and @200 cutoffs for MRR, NDCG, and Recall on 2AIRC dataset. *ColBERT-BERT-Medium-Amharic-AM* uses a medium-sized BERT encoder trained on the Amharic passage retrieval dataset; *ColBERT-RoBERTa-Medium-Amharic* uses a medium RoBERTa encoder trained on the same corpus; *ColBERT-RoBERTa-Base-Amharic* uses a larger RoBERTa base encoder finetuned for Amharic. Best results are marked in **bold** and statistically significant differences ($p < 0.05$) are indicated with †.

B.2 Qualitative Error Analysis

To complement our quantitative evaluation, we conducted a small-scale qualitative analysis to better understand retrieval behaviors. We manually inspected top-ranked passages for selected queries across both sparse and dense systems. Amharic-optimized dense models generally retrieved semantically relevant content, often capturing broader meanings beyond exact keyword matches. In contrast, sparse models like BM25 tended to prioritize surface-level term overlap, sometimes surfacing passages that were topically misaligned despite lexical similarity.

One notable failure pattern involved the handling of negation. Dense models, despite their semantic capabilities, frequently retrieved similar or identical passages for both affirmative and negated versions of a query, failing to reflect the semantic reversal. This indicates that current Amharic embeddings may inadequately model negation, likely due to limited exposure to such constructs during pretraining.

Figure 4 illustrates this issue: despite the presence of negation in Query 2, the model ranks the same passage as for the affirmative Query 1, with nearly identical similarity scores. This suggests insufficient sensitivity to fine-grained semantic shifts like polarity reversal. A broader set of such examples is provided in our Python notebook, available in the public GitHub repository.

C Hyperparameter Sensitivity

We conduct a grid search over learning rate, batch size, and training epochs using *RoBERTa-Medium-Amharic-embed* to analyze the impact of hyperparameters on retrieval effectiveness. Figures 5–10 present six heatmaps showing MRR@10, NDCG@10, and Recall@10 under two epoch settings (*3* and *5*). The results highlight that: (i) increasing training epochs from 3 to 5 yields consistent improvements across all metrics. For example, with a learning rate of `5e-5` and batch size `256`, MRR@10 improves from `0.721` to `0.737`, and Recall@10 rises from `0.875` to `0.887`. (ii) Among learning rates, `5e-5` consistently outperforms `2e-5`, especially at larger batch sizes. (iii) Batch size shows mild impact overall, with stable or slightly improved performance as size increases. The best overall configuration, `5e-5` learning rate, `256` batch size, and `5` epochs, achieves the top scores across all metrics, emphasizing the benefits of sustained training with a moderately aggressive learning rate.

These trends highlight that while batch size offers some flexibility, retrieval quality is more sensitive to learning rate and training duration.

- **Query 1:** "የለንደን ማራቶን ሊሰረዝ ይችላል"
  - **Top Document:** "በቀጣዩ መስከረም ሊካሄድ ቀጠሮ የተያዘለት ‹‹የግሬት ኖርዝ ረን›› ግማሽ ማራቶን ውድድር መሰረዙን ተከትሎ ጥቅምት ላይ እንደሚካሄድ የተነገረው የለንደን ማራቶንም ሊሰረዝ እንደሚችል ጥርጣሬ ፈጥሯል..."
  - **Similarity (Relevance) Score:** 0.7415
- **Query 2:** "የለንደን ማራቶን አልተሰረዘም"
  - **Top Document:** "በቀጣዩ መስከረም ሊካሄድ ቀጠሮ የተያዘለት ‹‹የግሬት ኖርዝ ረን›› ግማሽ ማራቶን ውድድር መሰረዙን ተከትሎ ጥቅምት ላይ እንደሚካሄድ የተነገረው የለንደን ማራቶንም ሊሰረዝ እንደሚችል ጥርጣሬ ፈጥሯል..."
  - **Similarity (Relevance) Score:** 0.729

Figure 4: Negation failure case: The model retrieves the same top passage for both a positive (Query 1) and a negated (Query 2) version of the query, with comparable similarity scores. This reflects a lack of semantic sensitivity to negation.

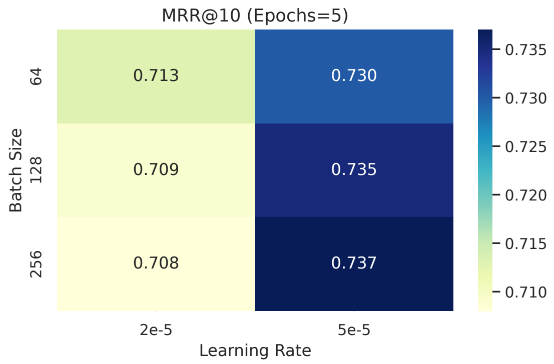

Figure 5: MRR@10 scores with 5 training epochs. The best performance (0.737) is achieved with learning rate 5e-5 and batch size 256. Higher learning rates consistently improve ranking quality across all batch sizes.

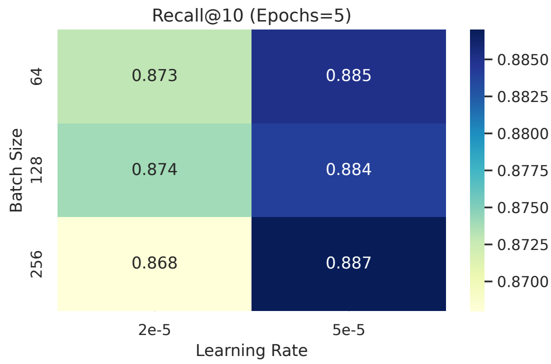

Figure 7: Recall@10 under 5 training epochs. Maximum recall (0.887) is observed at 5e-5/256. Performance improves steadily with training duration and a higher learning rate.

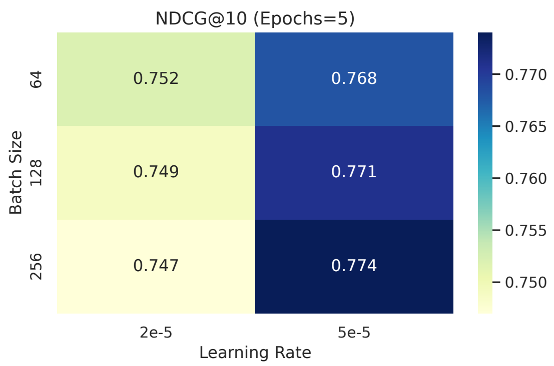

Figure 6: NDCG@10 scores with 5 training epochs. Peak score (0.774) occurs at 5e-5 learning rate and batch size 256. Larger batch sizes generally benefit from more aggressive learning.

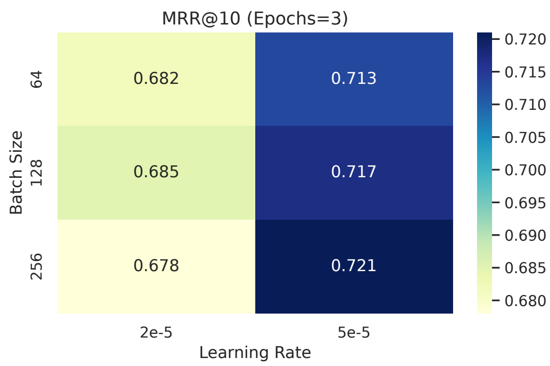

Figure 8: MRR@10 with 3 training epochs. Best score (0.721) is attained at 5e-5/256. Shorter training limits performance, but learning rate remains a strong influence.

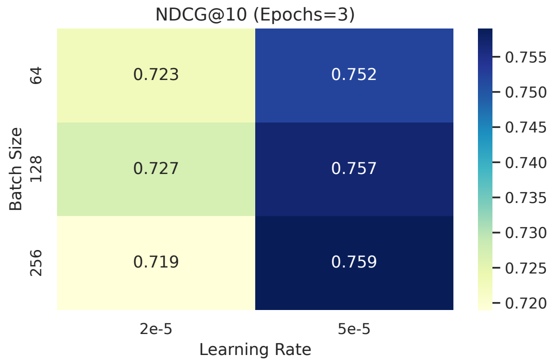

Figure 9: NDCG@10 with 3 training epochs. Performance is highest at 5e−5/128, and all batch sizes benefit from higher learning rates.

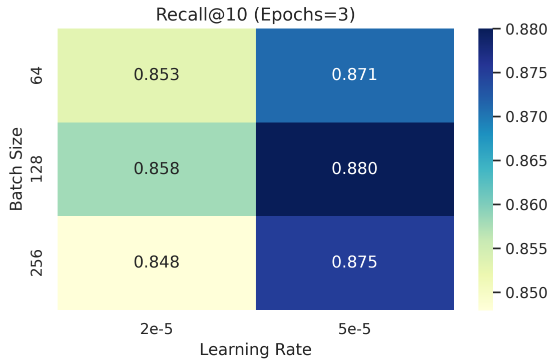

Figure 10: Recall@10 with 3 training epochs. The best score (0.880) is reached at 5e−5/128, with higher learning rates consistently outperforming 2e−5.