Optimized Text Embedding Models and Benchmarks for Amharic Passage Retrieval

Anonymous ACL submission

Abstract

Recent work has introduced several families of neural retrieval approaches that use transformer-based pre-trained language models to improve multilingual and cross-lingual retrieval. Their effectiveness for low-resource, morphologically rich languages such as Amharic remains underexplored and often limited 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 architectures. Our proposed RoBERTa-Base-Amharic-Embed (with a modest 110M parameters) outperforms the strongest multilingual model, Arctic Embed 2.0 (568M parameters), with a 5.01% relative improvement in MRR@10 and a 3.34% gain in Recall@10. Even more compact variants that we introduce, such as RoBERTa-Medium-Amharic-Embed (with just 42M parameters), remain competitive despite being 14x smaller. We benchmark our proposed models against sparse and dense retrieval approaches to systematically evaluate retrieval performance in Amharic. We reveal fundamental challenges in low-resource settings, underscoring the need for language-specific adaptation. Our work demonstrates the importance of optimizing retrieval models for morphologically complex languages and establishes a strong foundation for future research. To facilitate further advancements in low-resource information retrieval, we release our dataset, codebase, and trained models at our public repository.

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 use lexical similarity techniques like TF-IDF and BM25 (Robertson and Walker, 1997; Robertson and Zaragoza, 2009), efficiently match queries to documents using lexical similarity but 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. Neural retrieval models have significantly advanced document ranking by using transformer-based pre-trained language models, achieving state-of-the-art performance in question-answering benchmarks such as MS MARCO (Campos et al., 2016) and Natural Questions (Kwiatkowski et al., 2019). These models 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 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 (Tay et al., 2022; Metzler et al., 2021) uses pre-trained encoder-decoder models to consolidate indexing, retrieval, and ranking into a single generative framework. While promising, it lags behind dense retrieval in handling large-scale datasets and dynamic updates, requiring further study of its scalability and effectiveness (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, with prior studies focusing on transfer learning 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 fill the gap identified above, we focus on Amharic and introduce Amharic-optimized retrieval models and benchmarks, contributing in the following key areas: (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 retrieval performance baselines. (iii) *A language-specific vs. multilingual analysis*: we demonstrate that Amharic-optimized models outperform general-purpose multilingual embeddings, validating the need for linguistic specialization. (iv) *A public benchmark dataset*: we adapt the Amharic News dataset into an MS MARCO-style passage ranking corpus, enabling further reproducible research here.

2 Motivation

Recent studies highlight systemic shortcomings in low-resource language technologies, leading to retrieval failures, biased outputs, and exposure to harmful content (Shen et al., 2024; Nigatu and Raji, 2024). For example, Nigatu and Raji (2024) find that Amharic-speaking YouTube users frequently encounter policy-violating content due to retrieval systems misinterpreting 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, exacerbating inequities for underrepresented linguistic communities (Shen et al., 2024).

Amharic, Ethiopia's official language and the second most spoken Semitic language (Gezmu et al., 2018), presents unique challenges for IR. Its root-based templatic morphology allows a single root to generate multiple derived forms. These morphological variations, combined with the Ge'ez script (an Abugida with 33 base characters and complex syllable formations), make it structurally distinct from Indo-European and other well-resourced languages, rendering conventional retrieval models ineffective. 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 is not on retrieval optimization. Our work directly addresses this gap by optimizing retrieval methods to better accommodate Amharic's structural complexities, ultimately improving access to reliable and unbiased information in low-resource linguistic contexts.

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 ranking with more computationally 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 (Robert-

son and Zaragoza, 2009). It struggles with vocabulary mismatch and morphological variability, which are particularly problematic in morphologically rich languages like Amharic. Learned sparse retrieval (LSR) methods (Formal et al., 2021b,a) mitigate these issues by dynamically weighting and expanding terms, enhancing relevance while maintaining interpretability (Dai and Callan, 2020). LSR is constrained in low-resource settings due to limited annotated data, unseen dialectal diversity, and morphological complexity (e.g., Amharic's templatic morphology), which requires specialized subword tokenization or morphological analyzers that are often unavailable.

Dense retrieval. Dense retrieval encodes queries and documents into a shared semantic space for efficient approximate nearest neighbor search (Johnson et al., 2019; Karpukhin et al., 2020; Xiong et al., 2021). While it mitigates lexical mismatches, its effectiveness in low-resource languages is limited by the need for large-scale labeled data. Multilingual base 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 pre-training. But their effectiveness in morphologically complex languages, such as Amharic, has not been thoroughly investigated.

Recent advances in unsupervised contrastive learning, such as Contriever (Izacard et al., 2022), have shown strong zero-shot and multilingual retrieval performance, particularly in cross-lingual transfer and retrieval without labeled data. But their effectiveness in morphologically complex languages like Amharic remains unexplored, as existing evaluations do not account for challenges posed by 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) suggests that excessive segmentation in morphologically rich languages introduces noise into subword representations, thus degrading performance in downstream tasks.

Despite recent advances in multilingual dense retrieval, state-of-the-art models,[1] such as Arctic Embed 2.0 (Yu et al., 2024) and Multilingual E5 Text Embeddings (Wang et al., 2024), continue to struggle with highly inflected languages. These models often produce suboptimal tokenization, 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 impact 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). Systematic evaluations of both sparse and dense retrieval architectures remain absent, making principled comparisons difficult and leaving the effectiveness of different retrieval paradigms in Amharic IR largely unexamined.

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 News Text Classification Dataset (AMNEWS) (Azime and Mohammed, 2021), formatted in the MSMARCO passage retrieval style (see Section 5). A detailed analysis of 2AIRTC, its limitations, and our evaluations on this dataset is presented in Appendix A.

Our work addresses these gaps by introducing Amharic-specific optimizations (Section 4.2), including the use of stronger and more compact encoder base models to train embedding models that better accommodate Amharic's morphological complexity, along with contrastive training. Additionally, we train a late-interaction Amharic-specific ColBERT model and benchmark Amharic passage retrieval across both sparse and dense architectures. This enables rigorous and well-founded comparisons across retrieval paradigms.

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

4 Methodology

In this section, we outline our approach to Amharic dense retrieval. We first review dense retrieval models and ColBERT, which underpin our retrieval framework. We then introduce our Amharic embedding models, detailing 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 using a transformer-based encoder $\text{Enc}(\cdot)$:

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

The relevance of a passage p to a query q is then determined using cosine similarity or dot product, computed as $f(q, p) = \text{sim}(q_{\text{enc}}, p_{\text{enc}})$, where $\text{sim}(\cdot, \cdot)$ denotes similarity in the shared embedding space.

ColBERT: Late interaction retrieval

ColBERT (Khattab and Zaharia, 2020) improves query-document interactions by preserving token-level embeddings:

$$q_{\text{enc}} = [h_q^1, h_q^2, \ldots, h_q^m],\ p_{\text{enc}} = [h_p^1, h_p^2, \ldots, h_p^n] \tag{2}$$

where h_q^i and h_p^j represent contextualized token embeddings from the query and passage encoders, respectively. Relevance is computed via maximum similarity pooling across token embeddings:

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

This enables fine-grained matching while maintaining efficiency.

4.2 Amharic Text Embedding Models

We design three transformer-based dense retrieval models for Amharic, each with a distinct parameter size and a common context length of 512 to optimize the trade-off between retrieval effectiveness and computational efficiency.

(1) **RoBERTa-Base-AM-Embed** (110M parameters): A 12-layer transformer with a hidden dimension of 768, built upon the XLM-RoBERTa architecture (Conneau et al., 2020). This model leverages deep contextualized representations while remaining compatible with standard retrieval pipelines.

(2) **RoBERTa-Medium-AM-Embed** (42M parameters): A more compact variant employing an 8-layer transformer with a hidden dimension of 512, optimized for efficiency without significant performance degradation.

(3) **BERT-Medium-AM-Embed** (40M parameters): The most compact model among our proposed models, based on the BERT architecture (Devlin et al., 2019), featuring 8 layers with a hidden dimension of 512. This configuration is designed for latency-sensitive retrieval scenarios.

Embedding vector generation: For passage representation, we employ the following transformations to the last hidden states of the pre-trained base models: (i) Mean Pooling: Aggregates the last hidden state into a fixed-length vector representation: $\mathbf{h} = \frac{1}{T}\sum_{t=1}^{T} h_t$ where h_t represents token embeddings, and T is the sequence length. (ii) L2 Normalization: Constrains embeddings to unit sphere for cosine similarity computation: $\mathbf{h}_{\text{norm}} = \frac{\mathbf{h}}{||\mathbf{h}||_2}$.

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 30K Amharic query-passage pairs. Training is conducted for four epochs using the AdamW optimizer (lr=5e-5) and a cosine learning rate decay schedule. Model performance is evaluated using Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (NDCG), and Recall@K on Amharic passages. For further implementation details, refer to Section 5.2.

Multiple negatives ranking loss (MNRL). Following (Reimers and Gurevych, 2019), we optimize model parameters using in-batch negative sampling. Given a q, positive passage p^+ and hard negatives $\mathcal{N}$, the loss function $\mathcal{L}$ is formulated as:

$$-\log \frac{\exp(f(q,p^+))}{\exp(f(q,p^+)) + \sum_{p^- \in \mathcal{N}} \exp(f(q,p^-))}, \tag{4}$$

This objective maximizes the relative margin between positive and negative passages in the embedding space.

5 Experimental Setup

5.1 Training Data

For our experiments, we utilize the Amharic News Text Classification Dataset (AMNEWS) (Azime and Mohammed, 2021), which originally comprises 50,706 Amharic news articles categorized into six domains: Local News, Sport, Politics, International News, Business, and Entertainment. The article bodies serve as retrieval passages, while headlines function as queries, simulating real-world search scenarios. Since the dataset lacks explicit relevance judgments, we adopt a weak supervision approach, assuming each article to be relevant to its corresponding headline. To ensure data quality, we preprocess the dataset by removing duplicates using MD5 hashing. To align with standard IR benchmarks, we reformat it into an MS MARCO-style passage retrieval dataset. After preprocessing, this results in a dataset of `30K` query-passage pairs, which we then split 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. The embedding models were trained on a single `A100 40GB` GPU for `4` epochs using the Sentence Transformer Trainer from the Sentence Transformers Python library.[2] We used a learning rate of `5e-5`, a batch size of `128`, a cosine learning rate scheduler, and the multiple negatives ranking loss for optimization.

Sparse retrieval baselines. For BM25-based retrieval, we utilize LlamaIndex's BM25Retriever,[3]

Dense retrieval baseline. We implemented ColBERT using its official repository,[4] adapting it for Amharic with `RoBERTa-Medium-Amharic`, an Amharic encoder-based model. The model was trained with a learning rate of `3e-5` and a batch size of `32`, using five negatives sampled from the top `100` BM25-ranked documents.

Evaluation metrics. We evaluate retrieval effectiveness using established common metrics in IR, that capture ranking quality and relevance, including (i) MRR@k, which measures ranking effectiveness by averaging reciprocal ranks. (ii) NDCG@k, which captures ranking quality with a logarithmic discount factor. (iii) Recall@K, which measures how often the relevant passage appears in the top retrieved results.

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

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

[4] https://github.com/stanford-futuredata/ColBERT

6 Experimental Evaluation and Results

This section presents an empirical evaluation to address the following research questions:

RQ1 In Amharic passage retrieval, how effectively can language-specific embeddings enhance ranking accuracy compared to general multilingual models? (Section 6.1)

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

RQ3 How does tokenization quality impact retrieval effectiveness in Amharic dense retrieval models, particularly considering subword segmentation challenges? (Section 6.3)?

RQ4 To what extent does the base model's size influence the ranking performance of neural retrieval models in low-resource Amharic settings? (Section 6.4)

6.1 Multilingual vs. Amharic-Optimized Embedding Models

This section examines how effectively language-specific embedding models enhance ranking accuracy compared to general multilingual models in Amharic passage retrieval. To address this, Amharic-optimized models are evaluated against state-of-the-art multilingual embeddings, using standard retrieval metrics. As shown in Table 1, Amharic-specific models consistently outperform multilingual models across all metrics. The best-performing multilingual model, Snowflake-Arctic-Embed (568M parameters), achieves an MRR@10 of 0.719, while RoBERTa-Base-Amharic-Embed (110M parameters) surpasses it with an MRR@10 of 0.755, a 5.0% relative improvement. While larger multilingual models may alleviate some tokenization inefficiencies by learning richer subword representations, this does not necessarily translate to outperforming well-optimized language-specific models. Similarly, in Recall@10, the highest-scoring multilingual model reaches 0.868, whereas RoBERTa-Base-Amharic-Embed achieves 0.897, marking a 3.3% gain in top-ranked retrieval ac-

curacy. Beyond accuracy improvements, parameter efficiency provides further insight into the advantage of language-specific models. RoBERTa-Medium-Amharic-Embed (42M parameters) remains competitive, achieving 0.707 MRR@10 and 0.861 Recall@10, despite being 14x smaller than Snowflake-Arctic-Embed. This suggests that scaling multilingual models does not necessarily translate to better performance in low-resource settings. RoBERTa-Base-Amharic-Embed, at 110M parameters, outperforms all multilingual baselines while being 5x smaller than the strongest competitor, reinforcing the importance of language-specific fine-tuning over brute-force scaling. These findings highlight the inefficiency of general multilingual models in Amharic retrieval and the significant gains from adapting models specifically for the language. Even with fewer parameters, Amharic-optimized models achieve comparable or superior results, confirming that language-specific adaptation is both effective and computationally efficient.

6.2 Sparse vs. Dense Retrieval Performance

This section evaluates how term-based and dense retrieval models compare in effectiveness, establishing a benchmark for Amharic passage retrieval. As shown in Table 2, BM25 achieves competitive performance (0.657 MRR@10, 0.774 Recall@10), confirming its value as a baseline. However, dense retrieval methods demonstrate substantial improvements, particularly in ranking effectiveness.

Among the dense retrieval models, the ColBERT-AM model (which uses RoBERTa-Medium-Amharic as its backbone) enhances retrieval quality, achieving 0.754 MRR@10 and 0.858 Recall@10, effectively outperforming BM25 by leveraging late interaction mechanisms. The RoBERTa-Base-Amharic-embed model achieves the best performance, reaching 0.755 MRR@10 and 0.897 Recall@10, surpassing both BM25 and ColBERT. The advantage is even more pronounced in Recall@100, where RoBERTa achieves 0.971 an 11.5% improvement over BM25 demonstrating its ability to retrieve relevant documents in large candidate pools. While BM25 remains competitive, dense retrieval models provide better ranking accuracy, particularly in retrieving the most relevant documents at top positions. The gains from RoBERTa-Base-Amharic-embed suggest that bi-encoder models, when trained on Amharic-specific data, can outperform both term-based and late interaction retrieval methods. These findings emphasize the value of language-specific pretraining, as both dense models leverage Amharic-optimized architectures, with RoBERTa's bi-encoder design offering the best balance of precision and recall.

6.3 Tokenization Quality and Retrieval Performance

This section examines the impact of tokenization quality on retrieval performance in Amharic dense retrieval models, focusing on subword fertility, the average number of tokens per word (Pietra et al., 1997). Figure 1 compares subword fertility across models, highlighting its effect on retrieval accuracy. This analysis is conducted using a subset of `10k` articles from the Amharic news dataset.

Higher subword fertility increases computational costs and degrades retrieval accuracy due to excessive segmentation disrupting word representations (Ali et al., 2024). Table 1 reflects this: *gte-modernbert-base*, with the highest fertility (13.80), exhibits the weakest retrieval performance (MRR@10 = 0.019). This supports the hypothesis that over-segmentation undermines semantic coherence, aligning with prior findings (Alajrami et al., 2023). In contrast, Amharic-optimized models, such as RoBERTa-Base-Amharic-Embed, achieve lower fertility (1.46) and superior retrieval results (MRR@10 = 0.755). Similar patterns emerge across other Amharic-specific models (RoBERTa-Medium-Amharic-Embed, MRR@10 = 0.707; BERT-Medium-Amharic-Embed, MRR@10 = 0.657), reinforcing the benefits of Amharic-specific optimizations in capturing the language's morphological complexity for improved retrieval performance.

Among multilingual models, snowflake-arctic-embed-l-v2.0 demonstrates the highest retrieval performance (MRR@10 = 0.719) despite having the same subword fertility (2.35), as gte-multilingual-base and multilingual-e5-large-instruct. This suggests that while larger model size (e.g., 568M parameters in Snowflake-Arctic-Embed) can help mitigate some inefficiencies in multilingual tokenization strategies, it does not fully compensate for the advantages of language-specific adaptation, as evidenced by the superior performance of Amharic-optimized models. In contrast, gte-modernbert-base, which exhibits significantly higher fertility, performs poorly, highlighting the negative impact of excessive segmentation on retrieval. Similarly, gte-multilingual-base and multilingual-e5-large-instruct, both with moderate

	Model	Params	MRR@10	NDCG@10	Recall @10	Recall @50	Recall @100
Multilingual models	gte-modernbert-base	149M	0.019	0.022	0.030	0.054	0.065
	gte-multilingual-base	305M	0.649	0.684	0.794	0.876	0.904
	multilingual-e5-large-instruct	560M	0.713	0.747	0.853	0.924	0.946
	snowflake-arctic-embed-l-v2.0	568M	0.719	0.755	0.868	0.941	0.957
Ours	BERT-Medium-Amharic-embed	40M	0.657	0.696	0.817	0.916	0.945
	RoBERTa-Medium-Amharic-embed	42M	0.707	0.744	0.861	0.941	0.963
	RoBERTa-Base-Amharic-embed	110M	**0.755**†	**0.790**†	**0.897**†	**0.957**†	**0.971**†

Table 1: Performance comparison on the Amharic News dataset between Amharic-optimized and multilingual dense retrieval models, all based on a bi-encoder architecture. 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**. Statistically significant improvements ($p < 0.05$) over the strongest baseline are marked with †, determined using a paired t-test.

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	ColBERT-AM	0.754	0.777	0.858	0.917	0.931
Dense retrieval	RoBERTa-Base-Amharic-embed	**0.755**	**0.790**†	**0.897**†	**0.957**†	**0.971**†

Table 2: Performance of retrieval models on the Amharic News dataset. ColBERT-AM uses RoBERTa-Medium-Amharic as its backbone model. The best results are highlighted in bold, and statistically significant improvements ($p < 0.05$) over the strongest baseline are marked with †, determined using a paired t-test.

fertility (2.35), achieve better retrieval performance than gte-modernbert-base but fall slightly behind snowflake-arctic-embed-l-v2.0, reinforcing the hypothesis that larger model size may mitigate some inefficiencies of multilingual tokenization strategies. However, they still fall short of Amharic-specific models, reinforcing the importance of low subword fertility for improving retrieval efficiency. Furthermore, Amharic-specific models consistently outperform their multilingual counterparts, validating the importance of linguistic specialization in embedding design. These results align with prior research (Toraman et al., 2023; Ali et al., 2024), emphasizing the critical role of tokenization strategies, particularly for morphologically complex languages, in enhancing computational efficiency and ultimately improving downstream retrieval performance.

6.4 Base Model Size vs. Efficiency in Amharic Neural Retrieval

Table 1 presents the influence of base model size on retrieval performance in low-resource Amharic settings. We evaluate ColBERT's effectiveness using three Amharic base models, BERT-Medium-Amharic, RoBERTa-Medium-Amharic, and RoBERTa-Base-Amharic on a 50k articles.

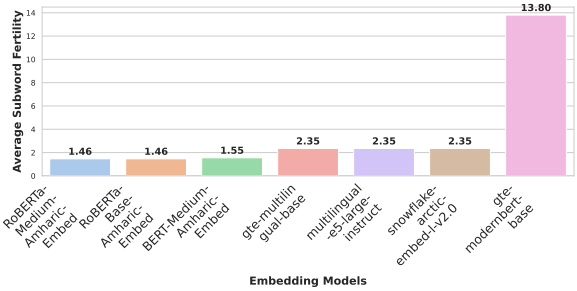

Figure 1: Average subword fertility across embedding models. Lower fertility preserves word integrity, while higher fertility may lead to excessive segmentation, affecting retrieval performance.

While the embedding model derived from RoBERTa-Base-Amharic, referred to as RoBERTa-Base-Amharic-Embed, achieves the highest standalone dense retrieval performance (shown in Table 1) (MRR@10: 0.755, Recall@10: 0.897), integrating the base RoBERTa-Base-Amharic model into ColBERT shows a different trend. As shown in Table 3, RoBERTa-Medium-Amharic (42M) outperforms the larger RoBERTa-Base-Amharic (110M) within ColBERT (MRR@10: 0.754 vs. 0.736). This suggests that increased model size

Base model	Params	MRR@10	NDCG@10
BERT-Med-Amh	40M	0.748	0.771
RoB-Med-Amh	42M	**0.754**	**0.777**
RoB-Base-Amh	110M	0.736	0.760

Table 3: Retrieval performance of ColBERT with different Amharic base models on the Amharic news dataset. BERT-Med-Am refers to BERT-Medium-Amharic-embed, RoB-Med-Am to RoBERTa-Medium-Amharic-embed, and RoB-Base-Amh to RoBERTa-Base-Amharic-embed. The best result in each column is in **bold**.

does not always enhance performance in architectures focused on token-wise interactions. A possible explanation is that the higher parameter count of RoBERTa-Base-Amharic (110M) risks overfitting on moderate-sized datasets, limiting its generalization in ColBERT's token-level retrieval framework. Conversely, RoBERTa-Medium-Amharic balances specificity and generalization more effectively, aligning better with ColBERT's fine-grained token representation needs. These findings highlight a critical trade-off: larger base models offer advantages in standalone dense retrieval but may not consistently improve performance in token-level interaction architectures like ColBERT. In low-resource Amharic settings, RoBERTa-Medium-Amharic emerges as the optimal choice, achieving strong performance (MRR@10: 0.754) with greater efficiency (42M parameters). These results emphasize that model scaling does not universally improve performance; architectures relying on fine-grained token interactions may benefit more from parameter-efficient base models in low-resource scenarios.

6.5 Key Issues in Amharic Passage Retrieval Performance

Table 1 demonstrates that Amharic-optimized models, such as RoBERTa-Base-Amharic-Embed, outperform multilingual models in Amharic passage retrieval. However, several persistent challenges highlight the inherent difficulties in processing Amharic text. (i) One major issue is morphological complexity. While optimized tokenization improves performance over multilingual models, over segmentation particularly in compound or inflected words disrupts semantic coherence, impairing retrieval accuracy. This issue, common in morphologically rich languages, leads to fragmented word representations that hinder effective passage retrieval. Although our language specific fine-tuning mitigates some effects, it does not fully resolve tokenization inconsistencies. (ii) Another key challenge is the size of the pretraining corpus. The Amharic-optimized models were trained on a relatively small dataset of 300 million tokens, significantly fewer than the billions of tokens available for high-resource languages like English. This data scarcity restricts the models' ability to generalize, making it difficult to match the performance of models trained on larger datasets. As a result, RoBERTa-Base-Amharic-Embed, despite outperforming multilingual models, struggles with rare or out-of-context terms, particularly in retrieval tasks. (iii) The AMNEWS dataset, used in this study, lacks human-labeled relevance judgments, introducing noise into model evaluation. The assumption that headlines accurately reflect article relevance, while practical, does not fully capture the nuances of document relevance. This limitation affects the reliability of performance metrics. Additionally, the dataset's relatively small size restricts the generalizability of findings to larger and more diverse Amharic text collections. These challenges, further discussed in the Limitations section (Section 8), underscore fundamental issues in Amharic passage retrieval and the limitations of our approach.

7 Conclusion

We have introduced Amharic-optimized dense retrieval models and established the first systematic benchmark for Amharic passage retrieval. Our findings show that language-specific embeddings outperform multilingual baselines, highlighting the necessity of linguistic adaptation for morphologically complex languages. We also demonstrated that tokenization quality significantly impacts retrieval performance, with over-segmentation degrading accuracy.

These results expose the limitations of existing multilingual retrieval systems and reinforce the need for models tailored to low-resource languages. Despite these advancements, our study is constrained by reliance on a single dataset (AMNEWS) and the absence of standardized relevance judgments, limiting broader generalizability.

Future research should enhance morphological tokenization, extend retrieval to document-level search and question-answering, and explore domain adaptation to further advance Amharic IR.

8 Limitations

While our study establishes strong benchmarks for Amharic passage retrieval, several limitations must be acknowledged.

Dataset and evaluation constraints. Our experiments rely on the Amharic News Text Classification Dataset (AMNEWS), which lacks explicit human-labeled relevance judgments. The weak supervision approach assumes that each article is relevant to its corresponding headline, which may introduce noise in retrieval evaluation. Additionally, our dataset size is limited, restricting generalizability to larger or more diverse collections.

Pre-training data limitations. The Amharic base models used in this study were pre-trained on 300 million tokens, primarily sourced from webpages, news, and tweets. This is significantly smaller compared to the data used to train encoder models for high-resource languages, such as English BERT (3.3 billion tokens) and RoBERTa (over 30 billion tokens). The relatively small pre-training corpus may constrain the models' ability to generalize and perform on par with retrieval models derived from base models that were trained on larger-scale corpora.

Domain generalization. Our models are trained and evaluated on news articles, which may not fully generalize to other domains such as legal, medical, or conversational retrieval. Their effectiveness outside the news domain remains untested and may require further adaptation.

Tokenization and morphological complexity. Amharic is a morphologically rich language, which poses challenges for subword tokenization. While our study highlights these challenges, it does not propose direct mitigation strategies beyond language-specific fine-tuning. Tokenization inconsistencies can lead to over-segmentation, potentially affecting retrieval accuracy.

These limitations highlight key areas for future research, including expanding training data, incorporating human-labeled relevance judgments, improving tokenization strategies, and broadening linguistic coverage.

9 Ethical Considerations

Our study focuses on improving Amharic passage retrieval. While our models demonstrate strong performance improvements, we acknowledge potential ethical concerns related to data biases, fairness, and responsible deployment.

Use of publicly available dataset. We use the Amharic News Text Classification Dataset (AMNEWS) (Azime and Mohammed, 2021) and the 2AIRTC dataset (Yeshambel et al., 2020), both publicly available and published. AMNEWS comprises news articles from various sources, while 2AIRTC is a TREC-like IR dataset with news articles, topics, and relevance judgments. As no additional data collection was performed, we adhere to ethical guidelines by using only openly accessible and documented resources.

Base model and pretraining data. The base models used to create our embedding models were pretrained on 300 million tokens from publicly available Amharic text, including webpages, news, and tweets. As we did not perform this pre-training ourselves, we rely on prior work for the base model's data collection and training details.

Bias and fairness considerations. Like many datasets sourced from online news content, AMNEWS 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 guidelines and emphasize that Amharic retrieval models should be deployed with caution, especially in sensitive applications. We strongly encourage transparent reporting of retrieval biases and responsible use of our models and dataset.

References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2023. SERENGETI: Massively multilingual language mod-

els for Africa. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1498–1537, Toronto, Canada. Association for Computational Linguistics.

Ahmed Alajrami, Katerina Margatina, and Nikolaos Aletras. 2023. Understanding the role of input token characters in language models: How does information loss affect performance? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9085–9108, Singapore. 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*.

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.

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 Adhoc Information Retrieval Test Collection

A notable contribution for Amharic Information Retrieval (IR) is 2AIRTC, the first Amharic Adhoc Information Retrieval Test Collection (Yeshambel et al., 2020). Developed following TREC-style evaluation methodologies, 2AIRTC consists of 12,583 manually judged documents and 240 search topics, serving as a structured benchmark for Amharic IR research. While this resource facilitates standardized evaluation, it exhibits several critical limitations that hinder its effectiveness as a retrieval benchmark.

Limitations of 2AIRTC. (i) Inconsistencies in Relevance Judgments: A major drawback of 2AIRTC is the inconsistency in relevance annotations, where numerous semantically relevant documents are not labeled as relevant. This misalignment between manual judgments and retrieval model outputs disproportionately affects embedding-based models, which frequently retrieve relevant yet unjudged documents. As a result, recall-based evaluation metrics become unreliable, potentially leading to misleading conclusions regarding retrieval effectiveness. (ii) Lack of Standardized Baseline Benchmarks: The absence of established baseline retrieval benchmarks in 2AIRTC makes systematic comparison across different retrieval architectures challenging. Without well-defined baseline performances, assessing improvements over existing methods remains difficult.

B Performance Comparison of Amharic-Optimized and Multilingual Dense Retrieval Models on 2AIRTC

Despite the limitations of 2AIRTC, it remains, to the best of our knowledge, the only publicly available test collection for Amharic Adhoc Information Retrieval. Therefore, we evaluate multilingual and Amharic-specific dense retrieval models on this benchmark to analyze their generalization ability. Specifically, we assess how well these models retrieve relevant documents when applied to a different dataset than they were trained on, without additional fine-tuning on 2AIRTC.

C Result Analysis

Table 4 presents a comparative evaluation of multilingual and Amharic-specific dense retrieval mod-

els on the 2AIRTC dataset. The results are inconsistent with our findings and unreliable due to the aforementioned reasons.

C.1 Multilingual vs. Amharic-Specific Models

Multilingual models exhibit the highest retrieval effectiveness, with *multilingual-e5-large-instruct* achieving the best NDCG@100 (0.808) and Recall@200 (0.911). However, despite having significantly fewer parameters, Amharic-specific models demonstrate competitive performance. *RoBERTa-Base-Amharic-embed (NEW 45k)* achieves an NDCG@100 of 0.771 and Recall@200 of 0.903, narrowing the performance gap with the best multilingual model. While multilingual models maintain an advantage, the relatively small margin suggests that language-specific adaptations can effectively compensate for model size disparities, highlighting the efficiency of domain adaptation in retrieval tasks.

This trend contrasts with our findings on the Amharic News dataset, where Amharic-specific models outperformed multilingual ones. The discrepancy suggests that 2AIRTC's domain characteristics and annotation inconsistencies may introduce systematic retrieval bias, influencing evaluation outcomes and limiting the reliability of cross-benchmark comparisons.

C.2 Impact of Model Size

Unlike typical trends in dense retrieval, larger models do not consistently yield better performance on 2AIRTC. *snowflake-arctic-embed-l-v2.0* (568M) underperforms relative to *multilingual-e5-large-instruct* (560M) and *gte-multilingual-base* (305M), reinforcing that pretraining data composition and model architecture can outweigh parameter count in determining retrieval effectiveness. Among Amharic-specific models, smaller architectures such as *BERT-Medium-Amharic-embed* (40M) perform below *RoBERTa-Base-Amharic-embed* (110M) but remain competitive relative to their scale.

However, given the known inconsistencies in 2AIRTC's annotations and domain-specific biases, these results should be interpreted with caution, as dataset-specific factors may influence model rankings and obscure broader retrieval trends.

C.3 Inconsistencies in 2AIRTC Evaluation

The results on 2AIRTC differ from trends observed in the Amharic News dataset, where Amharic-specific models consistently outperformed multilingual ones. This discrepancy raises concerns about the dataset's reliability. The limited 240 topics constrain generalization, while incomplete relevance labels distort recall-based metrics. Dense models, which retrieve documents based on semantic similarity rather than lexical overlap, may suffer disproportionately from missing relevance annotations.

D Future Directions for Amharic Retrieval Evaluation

While this study evaluates dense retrieval models in both multilingual and Amharic-specific settings, the limitations of 2AIRTC, that is, its small dataset size (240 topics) and inconsistencies in relevance annotations undermine the reliability of these evaluations. The limited number of queries restricts the generalizability of results, while incomplete relevance labels distort performance metrics, particularly for embedding-based retrieval models.

To enhance Amharic retrieval evaluation, future work should focus on:

- **Expanding and refining 2AIRTC** through more comprehensive and iterative relevance assessments, potentially leveraging crowdsourcing or semi-automated annotation to improve coverage and consistency.
- **Investigating morphology-aware retrieval techniques** to better handle Amharic's complex word formation processes and rich morphology.
- **Exploring query expansion and pseudo-relevance feedback** to mitigate vocabulary mismatches and enhance document retrieval effectiveness.
- **Benchmarking retrieval models across multiple Amharic datasets** to provide a more robust assessment of generalization and model effectiveness.

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	0.806	0.664	0.723	0.829
RoBERTa-Medium-Amharic-embed	42M	0.875	0.744	0.796	0.880
RoBERTa-Base-Amharic-embed	110M	0.864	0.753	0.816	0.892
RoBERTa-Base-Amharic-embed (NEW 45k)	110M	0.886	0.771	0.827	0.903
snowflake-arctic-embed-l-v2.0-finetuned-amharic	568M	0.760	0.740	0.800	0.868

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 model snowflake-arctic-embed-l-v2.0-finetuned-amharic is a fine-tuned version of Snowflake-Arctic-Embed-v2.0 using the news dataset (30k articles and headlines). The best-performing results are highlighted in **bold**.