AmharicIR+Instr: A Two-Dataset Resource for Neural Retrieval and Instruction Tuning

Anonymous ACL submission

Abstract

Neural retrieval and GPT-style generative models rely on large, high-quality supervised data, which is scarce for low-resource languages such as Amharic. We release an Amharic data resource consisting of two datasets that supports research on (i) neural retrieval-ranking and (ii) instruction-following text generation. The retrieval–ranking dataset contains 1,091 manually verified query-positive-negative document triplets drawn from diverse Amharic sources and constructed to support contrastive training and benchmarking of neural retrievers. Triplets are created through a combination of expert-curated queries, web-derived queries, and LLM-assisted generation, with positive/negative documents selected from the web or synthesized by LLMs and then validated by native speakers. The instruction prompt–response dataset comprises 6,285 Amharic prompt–response pairs spanning multiple domains and instruction types, generated with several LLMs and refined through manual review and correction. We release both datasets (https://anonymous.4open.science/r/AmharicIR-Instr/README.md) with standardized splits and formats (CSV/JSON/JSONL). We provide baselines for Amharic retrieval and generative modelling and show the importance of high quality data for low resource settings. These datasets also come with a methodology that can be generalized to other low-resource languages.

1 Introduction

Information retrieval (IR) aims to return documents that satisfy a user's information need. Traditional ranking functions such as TF–IDF and BM25 rely primarily on lexical matching, which can miss relevant content when queries and documents use synonyms and semantic variations, even after morphological normalization. Recent neural IR models have addressed this limitation by learning representations that capture semantic similarity and by optimizing retrieval models directly from relevance supervision.

Neural retrieval models encode queries and documents into vectors and scores them with a similarity function. Dense dual-encoder models (DPR-style) learn a single embedding per query and document, enabling efficient approximate nearest-neighbor search (Karpukhin et al., 2020). Late-interaction models such as the ColBERT family retain token-level representations and compute finer-grained matching scores at retrieval time, often improving effectiveness at higher storage/compute cost (Khattab and Zaharia, 2020). Sparse neural retrievers such as the SPLADE family produce sparse, vocabulary-aligned representations through learned term expansion, combining neural modelling with inverted-index retrieval (Formal et al., 2021; Lassance et al., 2023). Despite their architectural differences, these methods typically rely on supervised contrastive signals—e.g., query–positive–negative triplets or graded relevance judgments—for training and reproducible benchmarking. These models are effective, however, their performance depends critically on large, high-quality datasets, and models often fail to generalize, especially in low-resource languages (LRLs) (Zhang et al., 2023; Zhao et al., 2024).

Large-scale retrieval datasets such as MS MARCO (Bajaj et al., 2016), BEIR (Thakur et al., 2021), and Natural Questions (Kwiatkowski et al., 2019) provide millions of labelled query—document pairs or triplets, enabling both training and standardized evaluation. More recent multilingual adaptations extend retrieval research beyond English, but they often rely on translation or weak supervision due to limited annotations and may not reflect the linguistic and domain characteristics of truly LRLs (Litschko et al., 2022; Hu et al., 2023; Goworek et al., 2025).

Amharic, spoken by over 80 million people, is a morphologically rich Semitic language written in

the Gez script, with substantial surface-form variation due to inflection and derivation (Eberhard et al., 2023). From an IR perspective, this variation increases vocabulary mismatch and makes tokenization and normalization decisions (e.g., segmentation, handling orthographic variants, punctuation and diacritics) particularly consequential for both sparse and neural retrievers (Yeshambel et al., 2020b; Demlew and Yohannes, 2022; Yohannes et al., 2025). In addition, publicly available preprocessing tools and curated benchmarks are limited compared to high-resource languages, which complicates reproducibility and slows progress in neural retrieval and ranking for Amharic (Yeshambel et al., 2020a; Tonja et al., 2023). While recent Amharic IR work has explored dataset creation and evaluation, existing resources remain limited in scale and structure, and a manually verified triplet dataset designed for contrastive training and benchmarking is still lacking (Yeshambel et al., 2025; Mekonnen et al., 2025).

Beyond retrieval, recent advances in large language models (LLMs) have broadened information access to include generative question answering, instruction following, and conversational assistants. These models are commonly trained and evaluated using prompt–response datasets (or pairs), which are also scarce for Amharic, limiting progress in instruction tuning, generative evaluation, and retrieval-augmented generation (Alebachew et al., 2025). This motivates resources that jointly support retrieval/ranking and instruction-following generation under a shared Amharic setting.

In this paper, we release an Amharic data resource consisting of two datasets that supports research on (i) neural retrieval–ranking and (ii) instruction-following text generation. It combines (a) 1,091 manually verified query–positive–negative triplets for retrievers (Section 4), and (b) 6,285 GPT-style prompt–response pairs with manual quality control for generative models (Section 5). Both datasets consist in standardized formats (CSV/JSON/JSONL), predefined splits, and clear dataset statements to enable reproducible research. We also provide baselines. Beyond the datasets, we present a reusable methodology for low-resource languages (LRLs) (Section 3).

Our main contributions are:

- A native-speaker-validated data construction pipeline for LRLs, combining expert authoring, web-derived harvesting, and LLM-assisted generation.
- A manually verified Amharic retrieval dataset (1,091 triplets) for contrastive training.
- An Amharic instruction-following dataset (6,285 pairs) for instruction tuning.
- Standardized formats and documentation for reproducible comparisons.

2 Related Work

MS MARCO is a large-scale machine reading comprehension and IR dataset built from real user queries sampled from Bing search logs (Bajaj et al., 2016). It contains 1M anonymized queries and 9M passages extracted from retrieved documents. Passages marked as *is_selected* serve as positive examples for training, while the remaining passages act as implicit negatives, reflecting realistic retrieval noise.

Building on MS MARCO, Karpukhin et al. (2020) show that dense retrievers can be trained at scale using weak supervision from QA datasets over Wikipedia. They create training triples by pairing each question with a positive passage (from gold context or top BM25 retrieval) and sampling negatives from the corpus, including random, BM25 hard negatives (high lexical overlap but no answer), and in-batch negatives. The collection consists of 21M fixed-length (100-word) passages from English Wikipedia, with training queries from datasets like Natural Questions, TriviaQA, and SQuAD. This approach enables large-scale training without manual annotation and has become a standard for building dense retrieval datasets.

ANCE is a dense retrieval training approach that focuses on how negative documents are mined during contrastive learning (Xiong et al., 2021). Using existing datasets such as MS MARCO, Natural Questions, and TriviaQA, ANCE treats relevance judgments as positive and constructs hard negatives globally from the entire corpus. Negatives are retrieved using an approximate nearest neighbor (ANN) index over dense document embeddings, retrieving top-ranked but non-relevant documents for each query. The ANN index is asynchronously refreshed using periodically updated model checkpoints, ensuring that negatives are challenging throughout training.

ANCE highlights the importance of mining globally hard negatives to better match real retrieval errors. This issue is even more central in weakly supervised settings where explicit negatives are

not provided—such as recent Amharic passage retrieval benchmark (Mekonnen et al., 2025). This benchmark repurposes an existing text classification corpus (Azime and Mohammed, 2021) which contains more than 50k Amharic news articles spanning six domains. News headlines are treated as queries and the corresponding article bodies as passages, yielding approximately 45K query–passage pairs after preprocessing. Since explicit relevance judgments are unavailable, each headline is assumed as relevant to its associated article. Positive examples consist of headline–article pairs; there is no explicit negative examples.

Zhang et al. (2023) provide a best-practices guide for multilingual dense retrieval under different resource scenarios (model/data availability), highlighting the benefits of MS MARCO-style pre–fine-tuning, hard negatives, distillation, and hybrid sparse–dense retrieval—recommendations that motivate our focus on building reliable Amharic supervision and evaluation resources.

Such a dataset for Amharic was first created by Yeshambel et al. (2025). The authors constructed a manually curated Amharic dense retrieval training dataset consisting of 152 queries only, each paired with one positive document and one negative document. Each document is associated with a single topic, and relevance is defined by topic alignment. Because queries are non-overlapping, documents relevant to one query can serve as non-relevant (in-batch) negatives for others.

Existing Amharic resources are either weakly supervised or very small; none offers manually verified triplets at the size of ours, along with instruction pairs with quality control.

3 Reusable Low-Resource Data Construction Methodology

In high-resource settings, large retrieval and instruction datasets are constructed from abundant user logs, crowdsourcing at scale, or weak supervision from existing benchmarks (e.g. MS MARCO-style click logs or Wikipedia-based QA pairs). Such settings also make it easier to obtain large pools of implicit negatives or automatically mined training examples on which models can be trained. In contrast, for Amharic and other LRLs, such large-scale logs and annotated benchmarks are scarce, and weak supervision can be brittle due to morphology, orthographic variation, and limited coverage of web sources. As a result, dataset construction must place more emphasis on controlled coverage, explicit supervision, and human validation.

3.1 Overview

We propose a reusable methodology for constructing retrieval and instruction-following datasets in low-resource settings, when signals are unavailable or unreliable. Our methodology combines complementary data sources (expert authoring, web-derived harvesting, and LLM-assisted generation) and enforces native-speaker validation to ensure that positives truly satisfy the information need and that negatives are genuinely non-relevant (including hard negatives with lexical overlap), language-specific normalization to reduce noise from spelling and punctuation variation, and de-duplication. We believe these complementary steps are necessary to obtain dependable training and evaluation data in LRLs.

While we instantiate this methodology for Amharic in Sections 4 and 5, it is language-agnostic and can be adapted to other LRLs.

3.2 Pipeline Steps

The methodology consists in six steps as follows:

- **Specify coverage**. Define target domains, task types, and linguistic phenomena of interest (e.g., morphology, lexical variation, and multi-word expressions).
- **Generate candidates**. Produce candidate instances using three complementary sources: expert authoring, web-derived harvesting, and LLM-assisted generation, to balance quality, realism, and topical breadth.
- **Define explicit supervision**. Establish acceptance criteria for the supervision signal: for retrieval, specify what constitutes a positive vs. negative document (including hard negatives with lexical overlap); for generation, specify what constitutes an acceptable response (e.g., relevance, completeness, and style).
- **Native-speaker validation**. Manually verify candidates for relevance/correctness and language quality (fluency and grammaticality), and check factual plausibility when prompts require factual content.
- **Normalization and de-duplication**. Apply language-specific normalization (punctuation, spelling conventions, white-space, and encoding) and remove duplicates or near-duplicates to reduce redundancy and noise.

- **Release artifacts**. Provide standardized formats, splits, and documentation for reproducible benchmarking and reuse.

3.3 Instantiations and Adaptation to Other Languages

Instantiation for retrieval triplets. To build query-document triplets, Step 1 defines topical coverage and query types. Step 2 collects candidate documents from complementary sources (expert-driven search, web harvesting, and optional LLM-assisted generation). Step 3 specifies explicit criteria for positives and negatives: a positive document must satisfy the information need expressed by the query, whereas a negative document must not satisfy it; we additionally target hard negatives that share lexical overlap or topic proximity to the query. Step 4 performs native-speaker validation to confirm both relevance (d^+) and non-relevance (d^-). Finally, Steps 5–6 normalize text, remove duplicates, and release the triplets with documentation. Section 4 is an instantiation to Amharic language.

Instantiation for GPT-style instruction data. To build prompt–response pairs, Step 1 defines domain coverage and instruction-following behaviors (e.g., task completion, explanations, conversational responses). Step 2 generates candidate prompts and responses with multiple LLMs under domain/task templates. Step 3 replaces positive/negative criteria with response acceptance criteria (e.g., relevance to the prompt, completeness, fluency, and factual plausibility when required). Step 4 applies native-speaker review to correct hallucinations, unnatural phrasing, and inconsistencies while preserving the prompt intent. Steps 5–6 then normalize formatting and remove near-duplicates before releasing the dataset in standard formats. Section 5 is an instantiation to Amharic language.

Other languages. To instantiate the pipeline for another LRL, one can reuse the same steps while replacing the domain/task templates, selecting appropriate web sources, and adjusting the validation guidelines to language-specific conventions.

4 Amharic Query–Document Triplets for Dense Retrieval

Using the methodology of Section 3 we built a manually verified Amharic query–document triplets $\langle q, d^+, d^- \rangle$ for dense retrieval.

4.1 Query Construction

Queries are natural Amharic phrases intended to reflect realistic search behaviour. We collect queries using three complementary approaches to balance linguistic quality, topical coverage, and scale: (i) Expert-curated queries, (ii) Topic-derived web queries, and (iii) LLM-generated queries.

Expert-curated queries are crafted by a native Amharic speaker to reflect diverse information needs across domains. During query authoring, we explicitly consider Amharic linguistic properties, including complex morphology, lexical variety, and multi-word expressions whose meanings are not compositional (is distinct from individual words). This design ensures linguistic diversity and realistic search behaviour.

Topic-derived web queries are crafted by an expert and native Amharic speaker to reflect realistic information needs across diverse domains. They are built from Amharic web pages across domains such as news, blogs, government, technology, and related sources to capture natural user interests and topical variety. Concretely, we derive query candidates from the page/topic descriptors of documents and use this procedure to complement the manually created set with naturally occurring topics.

LLM-generated queries are produced by prompting LLMs to generate additional queries for specified topics–domains. Controlled prompt templates are used that explicitly specified the language (Amharic), retrieval paradigm (dense retrieval), domain (e.g., sports, medical, agriculture, communication, entertainment, transport, etc.), and query characteristics (concise, information-seeking, and unambiguous). To ensure sufficient domain coverage and balance, we enforce a minimum of 50 generated queries per domain. All queries are then manually reviewed and lightly edited for grammaticality, spelling, and fluency.

4.2 Positive and Negative Documents

For each query, we collect one positive and one negative document. While we categorize query according to the generation strategy use, we organize document collection by the source from which documents are obtained:

Search engine–retrieved documents. For expert–curated queries, we execute each query using Google Search. Positive documents are selected by carefully reading the top results and choosing the most relevant document(s). Negative documents

Table 1: Dense retrieval triplet dataset statistics.

Statistic	Value
Number of queries	1,092
Number of documents	2,184
Number of sentences	45243
Number of words	679,040
Avg. query length (words)	6
Avg. document length (words)	310.92
Vocabulary size (unique words)	113,598
Avg. #sentences per document	20.72
Min. #sentences per document	5
Max. #sentences per document	278

{

"query": "የአማርኛ ቋንቋ ተናጋሪ ምን ያህል ነው?",

"positive": "አማርኛ በኢትዮጵያ ውስጥ በብዙት የሚነገር ቋንቋ ነው። ቋንቋው የፌደራል መንግስት የስራ ቋንቋ መሆኑ አጠቃቀሙን አስፋፍቶታል። በህዝብ ቆጠራ መረጃ መሰረት እንደ እናት ቋንቋ እና እንደ ሁለተኛ ቋንቋ የሚናገሩ ሰዎች ቁጥር ከ 80 ሚሊየን በላይ መሆኑን ያመለክታል። በከተሞች ውስጥ አማርኛ በብዙ ህዝብ መካከል የመግባቢያ ቋንቋ ሆኗል። እንዲሁም በትምህርት ቤት, በመገናኛ ብዙሃንና በመንግስታዊ አገልግሎቶች ላይ በሰፊው ይጠቅማል። ከኢትዮጵያ ውጭም በዲያስፖራ ማህበርሰብ መካከል የአማርኛ ተናጋሪዎች በብዛት አሉ።",

"negative": "ኢትዮጵያ ብዙ ቋንቋዎች ያሏት ሀገር ናት። ከ 80 በላይ የተለያዩ ቋንቋዎች በሀገሪቷ ይነገራሉ። አያንዳንዱ ቋንቋ የራሱ ባህልና ታሪክ አለው። ኦሮምኛ, ትግረኛና ሶማሊኛ ከሚነገሩ ቋንቋዎች መካከል ይጠቀሳሉ። እነዚህ ቋንቋዎች በክልሎች መሪት የአስተዳደር ቋንቋ ሆነው ያገለግላሉ። ይህ የቋንቋ ብዙት የኢትዮጵያን ቋንቋ ሀብትን ያመለክታል።"

}

Figure 1: Example of a JSON record in the released triplet dataset.

-assess as such by reading the results are either hard negative if they are both top-ranked and share query terms, negative otherwise.

Website-sourced documents. For web-derived queries, we assign the corresponding source page as the positive document, ensuring that the query directly reflects the document content. Negative documents are sampled from the same website but from non-relevant pages, so that negatives preserve realistic topical and stylistic properties.

LLM-synthesized documents. For each LLM-generated query, the corresponding positive and negative (including hard-negative) documents are generated simultaneously using controlled LLM prompts and stored together with the query in CSV files organized by domain. The prompts explicitly constrains the output language to Amharic and the document length to 5–7 sentences. Positive documents are required to satisfy the query's information need, whereas negative documents are topically related but semantically misaligned.

A native annotator carefully read each candidate document with respect to its query and verify (i) that d^+ satisfies the query and (ii) that d^- does not satisfy the query, including in the presence of lexical overlap (hard negatives).

4.3 Resulting dataset

Dataset format and statistics. In total, the dataset consists of 1,092 triplets from which 200 triplets correspond to manually created queries, 276 triplets are derived from web-collected queries, and 616 triplets are synthesized using large language models. It was created by the first author of the paper. Table 2 reports some dataset-level statistics. Texts are in UTF-8 and written in Amharic (Ge'ez) script. We release the dataset in JSON,JSONL and CSV formats, with one triplet per line. Figure 1 is an example of a JSON record. Prompts used for the generation are also released.

Preprocessing. We preprocess queries and documents to remove duplicates, incomplete entries, non-Amharic content, and noise such as HTML tags and irrelevant metadata. We normalize punctuation and white-space, map sentence boundary markers "::" and "." to the standard Amharic full stop, and collapse long whitespace/newline sequences into a single space. We also perform manual checks for language correctness and fluency and filter sensitive or personal information.

5 Amharic GPT-Style Prompt–Response

Using the same methodology (Section 3), we construct a GPT-style Amharic prompt–response dataset of 6,225 pairs. To maximize coverage while maintaining quality, we generate candidate pairs with multiple large language models (LLMs), including ChatGPT, Gemini, and DeepSeek, and then perform manual validation and correction by native Amharic speakers.

5.1 Data Construction

Domains and task types. To encourage broad generalization, we predefine a set of domains including education, health, agriculture, governance, culture, religion, technology, and daily life. Within each domain, we focus on common instruction-following behaviors, including (i) direct task completion (e.g., "write/translate/summarize"), (ii) explanation (e.g., "explain why/how"), and (iii) conversational responses (e.g., multi-turn assistant-style answers). These domain and task specifications guide prompt design and ensure consistent coverage across different LLMs.

Prompt and response generation. Prompt–response candidates are generated using LLMs under domain- and task-level guidance. Rather than manually authoring each prompt, we provide the

target domain and task type and ask the LLMs to produce diverse Amharic prompts together with appropriate responses. This strategy increases topical diversity while keeping the dataset structure controlled.

Manual annotation, review and quality control. Following Steps 4–5 of Section 3, the annotation, quality control and manual review of the dataset were conducted by a native Amharic speaker with expertise in the language, constructing resources in LRL and information retrieval. The annotation was performed following consistent annotation guidelines that define relevance based on semantic matching between queries and documents. The guidelines specify explicit criteria for selecting positive and negative documents to ensure consistency and quality.

A document is labeled as positive (relevant) if it directly satisfies the information need expressed in the query. This includes cases where the document contains the exact answer, a clear explanation, or highly relevant contextual information. Semantic equivalence is considered even when there is no exact keyword overlap; for example, paraphrases, synonyms, or conceptually similar expressions in Amharic are treated as relevant. Documents that partially address the query but still provide useful and accurate information may also be considered positive.

A document is labeled as negative (non-relevant) if it does not address the query's intent, even if it shares some surface-level keywords. This includes documents that are topically unrelated, contain misleading or incorrect information, or discuss a different aspect of the same general topic without answering the query. Documents with only weak or tangential connections to the query are also marked as negative.

To ensure annotation quality, each triplet (query, positive document, negative document) was manually reviewed, and ambiguous cases were resolved through careful consideration of the query intent and document meaning. This process ensures that relevance judgments are semantically grounded and consistent across the dataset. Furthermore, the annotator reviewed all pairs for grammaticality, fluency, and relevance to the prompt, and correct hallucinations or unnatural phrasing while preserving the prompt intent. Each triplet was examined to verify that the positive and negative documents were logically consistent and factually reasonable with respect to the query. Any sentences that were contradictory, impossible, or clearly inaccurate were corrected or removed, while maintaining the intended meaning of the query.

Table 2: Amharic GPT dataset statistics.

Metrics	Value
Number of pairs	6,225
Avg. prompt length (words)	7.32
Avg. response length (words)	15.61
Max. prompt length	30
Max. response length	30
Vocabulary size (unique words)	31,278
Avg. word length (words)	4.13
Total. sentences	15,311
Question prompts (%)	95.89

5.2 Format and Release

After manual correction, we normalize punctuation, spelling conventions, and formatting. We remove duplicates and highly similar pairs to reduce redundancy. The final dataset consists of clean, high-quality Amharic prompt–response pairs suitable for both training and evaluation. The dataset consists of publicly available and synthetically generated text and does not contain any personal or sensitive information. It is intended solely for academic and research purposes and not for commercial use. Therefore, formal ethical approval was not performed or required.

To ensure usability across different platforms and research needs, we release the dataset in both CSV and JSON formats. The CSV format supports tabular inspection and preprocessing, while the JSON format facilitates integration with common LLM training pipelines.

6 Intended Uses and Baselines

Both datasets have been shared with reviewers on GitHub for evaluation. They will be released on Hugging Face and GitHub upon paper acceptance, together with documentation and access instructions under a CC BY 4.0 license. We provide here some baselines on existing -non contaminated- IR collections for neural retrieval.

6.1 Dense retrieval triplets

The Amharic query–document triplet dataset presented in Section 4 enables training and evaluation of retrieval models that rely on explicit positive and negative examples.

Table 3: Retrieval performance -NDCG, MAP, Recall and Precision) of fine-tuned Splade, RoBERTa, and ColBERT models at different cutoff values k for implicit and hard negatives.

		Splade				**RoBERTa**				**ColBERT**			
	k	NDCG@k	MAP@k	R@k	P@k	NDCG@k	MAP@k	R@k	P@k	NDCG@k	MAP@k	R@k	P@k
implicit neg.	1	0.58	0.58	0.58	0.58	**0.75**	**0.75**	**0.75**	**0.75**	0.74	0.74	0.74	0.74
	3	0.65	0.63	0.69	0.23	**0.83**	**0.81**	**0.88**	**0.29**	0.82	0.80	**0.88**	**0.29**
	5	0.68	0.65	0.76	0.15	**0.84**	**0.82**	**0.92**	**0.18**	0.83	0.81	0.90	0.18
	10	0.70	0.66	0.82	0.08	**0.85**	**0.82**	**0.95**	**0.10**	0.84	0.81	0.94	0.09
	100	0.72	0.66	0.95	0.01	**0.86**	**0.82**	**0.99**	**0.01**	0.85	0.81	**0.99**	**0.01**
hard neg.	1	0.69	0.69	0.69	0.69	**0.83**	**0.83**	**0.83**	**0.83**	0.79	0.79	0.79	0.79
	3	0.76	0.74	0.81	0.27	**0.89**	**0.87**	**0.93**	**0.31**	0.85	0.84	0.90	0.30
	5	0.78	0.76	0.87	0.17	**0.90**	**0.88**	**0.95**	**0.19**	0.87	0.85	0.94	0.19
	10	0.80	0.76	0.93	0.09	**0.91**	**0.88**	**0.99**	**0.10**	0.88	0.85	0.96	0.10
	100	0.81	0.77	0.98	0.01	**0.91**	**0.89**	**1.00**	**0.01**	0.88	0.85	0.99	0.01

Models. We consider SPLADE-RoBERTa-Amharic-Medium (sparse expansion), RoBERTa-Amharic-Embed-Medium (dense embeddings), and ColBERT-Amharic-Medium (late interaction) pre-trained Amharic neural retrieval models [1]. We fine-tuned these models on the training/validation part of our dataset. We compare (i) when using the full triplet with hard negative manually checked with (ii) implicit negatives (usual approach in MS Marco and BEIR datasets). The fine-tuned models and the dataset splits are available on GitHub [2].

Hard negatives help. Table 3 reports the results on the test part of our dataset. On implicit negative, all models achieve competitive retrieval performance across most cut-off values. RoBERTa achieves the highest performances on the different cutoff values. ColBERT performed slightly lower but maintains strong Recall, and SPLADE is the less performing. NDCG, MAP and recall improve with increasing k, while precision decreased. The important decrease of precision with cut-off $k > 3$ is because each query has only a small number of relevant documents in qrels, causing more non-relevant documents to appear among the top-ranked results. These results establish a strong baseline for future experiments.

Training with hard negative consistently outperforms the implicit negative counter part, across all models (Splade, RoBERTa, ColBERT) and evaluation metrics (NDCG, MAP, Recall, Precision). The most notable differences occur in Recall and MAP, where hard negatives often lead to substantial gains (for example, for Splade at $k = 3$, recall increases by $+0.12$, from 0.69 to 0.81), while precision shows the smallest improvements.

[1] Those models are available on : https://huggingface.co/rasyosef/[ModelName]

[2] https://anonymous.4open.science/r/AmharicIR-Instr

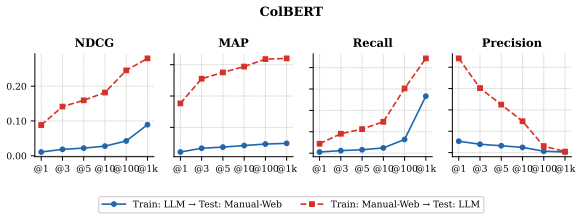

Figure 2: Colbert finetuned on manual and texted on synthetic data (red) and reversly (blue).

Manual is better than synthetic. The dataset contains 475 manual queries and 617 synthetic ones. We made a cross evaluation. Either the manual queries are used for training and the synthetic for test, or the reverse. Figure 2 shows manual queries are much more impactful than synthetic ones, although they are fewer.

Comparison with other triplet collections. The existing Amharic manually curated triplet dataset (Yeshambel et al., 2025) consists of only 152 triplets, collected solely from web sources; it lacks systematic preprocessing. Moreover, the absence of hard negative samples reduces its effectiveness for contrastive learning, and its applicability is largely limited to basic dense retrieval settings. In the same way, (Mekonnen et al., 2025) Amharic dataset, although larger, uses implicit negative, not hard curated negative. We show the importance of the hard negative. Here, we also significantly enhance the diversity and coverage of the dataset by combining triplets from three sources: web-collected, manually constructed, and LLM-synthesized data. Notably, the inclusion of manually created queries introduces more realistic representations of user information needs, which were not captured in prior work.

Our dataset not only exceeds several standard benchmarks in terms of query size but also provides

explicitly constructed triplets, making it directly suitable for training and evaluation. Therefore, we argue that the size, structure, diversity and quality of our dataset are sufficient and aligned with established practices in neural information retrieval research.

6.2 GPT-style prompt–response pairs

The Amharic prompt–response dataset supports instruction tuning, supervised fine-tuning, and controlled evaluation of generative models in Amharic. Potential uses include: training or adapting Amharic-capable LLMs to follow instructions across domains, evaluating generation quality (fluency, relevance, robustness) under domain shifts, and studying retrieval-augmented generation (RAG) by combining the prompt–response dataset with the retrieval triplets as an external knowledge source.

7 Conclusion

Neural retrieval and GPT-style generative models rely on large, high-quality supervised datasets, which are scarce for LRLs like Amharic. We present a new Amharic resource consisting of two datasets for (i) neural retrieval and ranking, and (ii) instruction-following text generation. The retrieval–ranking dataset contains query—positive–negative triplets for contrastive training and benchmarking of neural retrievers, including ColBERT-style late interaction and SPLADE-style sparse models. Triplets are created using expert-curated queries, web-derived queries, and LLM-assisted generation, with documents validated by native speaker. The instruction prompt–response dataset includes Amharic prompt–response pairs across multiple domains, generated with three LLMs, and refined for grammaticality, relevance, fluency, and factual accuracy. Baseline experiments on fine-tuned SPLADE, RoBERTa, and ColBERT models show that RoBERTa achieves the best retrieval performance. The results indicate that fine-tuning on the Amharic triplet dataset yields effective retrieval performance across multiple neural architectures. The comparative performance analysis of fine-tuning with hard negatives consistently improves retrieval performance across all models and evaluation metrics. Both datasets are released with standardized splits and formats (CSV/JSON/JSONL) to enable reproducible research. The methodology can be generalized to other LRLs, helping expand NLP resources beyond widely studied languages.

These datasets and methods address the scarcity of high-quality Amharic data and provide a foundation for future work in neural retrieval and generative modeling. The datasets can be expanded and scaled by associating each query with multiple positive and multiple negative documents.

Computing Infrastructure

All experiments were conducted on a computing node equipped with an AMD EPYC 9254 CPU (24 cores per CPU), 512 GiB of RAM. For GPU-based computations, we used two NVIDIA H100 GPUs, each with 94 GiB of memory.

Limitations

The dataset was curated exclusively by the first author, a native Amharic speaker of the language. This procedure may have introduced non-intentional bias.

We have proposed initial baselines and comparisons, but others could have been included.

8 Ethical Considerations

No external annotators were involved; hence, no recruitment or compensation procedures were applicable. The corpus was compiled from publicly available sources and augmented with synthetically generated data using large language models. The Web-based and manual-based datasets we collected from publicly accessible web sources in the Amharic language. Specifically, we utilized content from Amharic Wikipedia (https://am.wikipedia.org/), Facebook public pages (https://web.facebook.com/AlAinAmharic/), Amharic blog posts such as Dawit Worku Blog (https://dawitworku.wordpress.com/), and news websites including Ethiopian News Agency (https://www.ena.et/web/amh), Ethiopian Broadcasting Corporation (https://www.ebc.et/), and Al Ain Amharic (https://am.al-ain.com/). All data were collected either manually or through standard web access methods from content that is publicly available without authentication or access restrictions.

The dataset does not include personally identifiable or sensitive information. Given the absence of human subjects and private data, formal ethical approval was not required. However, the construction process adhered to standard data ethics practices,

including quality control, deduplication, and content filtering.

We used AI assistants to improve our phrasing, not to write the paper.

References

Kindie Alebachew, Durga Prasad Sharma, and Messay Samuel. 2025. Navigating the nlp and text-based conversational ai: a survey for problem formulation in amharic texts. *Discover Data*, 3.

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

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, and 1 others. 2016. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Gashaw Demlew and Dessalew Yohannes. 2022. Resolving amharic lexical ambiguity using neural word embedding. In *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 55–60.

DM Eberhard, GF Simons, and Fennig CD. 2023. Cd (eds.)," ethnologue: Languages of the world. *Twenty-sixth edition", Dallas, Texas: SIL International, Turkish Language Ethnologue*.

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

Roksana Goworek, Olivia Macmillan-Scott, and Eda B. Özyiğit. 2025. What drives cross-lingual ranking? retrieval approaches with multilingual language models.

Xiyang Hu, Xinchi Chen, Peng Qi, Deguang Kong, Kunlun Liu, William Yang Wang, and Zhiheng Huang. 2023. Language agnostic multilingual information retrieval with contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9133–9146, Toronto, Canada. Association for Computational Linguistics.

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*, SIGIR '20, 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.

Carlos Lassance, Simon Lupart, Hervé Déjean, Stéphane Clinchant, and Nicola Tonellotto. 2023. A static pruning study on sparse neural retrievers. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 1771–1775, New York, NY, USA. Association for Computing Machinery.

Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2022. On cross-lingual retrieval with multilingual text encoders. *Information Retrieval Journal*, 25:149–183.

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

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *NeurIPS 2021 Datasets and Benchmarks Track*.

Atnafu Lambebo Tonja, Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Moges Ahmed Mehamed, Olga Kolesnikova, and Seid Muhie Yimam. 2023. Natural language processing in Ethiopian languages: Current state, challenges, and opportunities. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 126–139, Dubrovnik, Croatia. Association for Computational Linguistics.

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 *International Conference on Learning Representations Poster*.

Tilahun Yeshambel, Moncef Garouani, Serge Molina, and Josiane Mothe. 2025. Dense retrieval for low resource languages-the case of amharic language. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3098–3100.

Tilahun Yeshambel, Josiane Mothe, and Yaregal Assabie. 2020a. 2airtc: The amharic adhoc information retrieval test collection. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 55–66. Springer.

Tilahun Yeshambel, Josiane Mothe, and Yaregal Assabie. 2020b. Amharic document representation for adhoc retrieval. In *Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*.

Dessalew Yohannes, Yenewondim Biadgie Sinshaw, Surafiel Habib Asefa, and Yaregal Assabie. 2025. Amharic document clustering using semantic information from neural word embedding and encyclopedic knowledge. *Scientific African*, 28:e02657.

Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. 2023. Toward best practices for training multilingual dense retrieval models. *ACM Transactions on Information Systems*, 42(2):1–33.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60.