How does increasing the proportion of mined hard negatives in multilingual dense retrieval training datasets impact Mean

Assignee Research

June 11, 2026

Abstract

Dense retrieval models using a transformer-based bi-encoder architecture have emerged as an active area of research. In this article, we focus on the task of monolingual retrieval in a variety of typologically diverse languages using such an architecture. Although recent work with multilingual transformers demonstrates that they exhibit strong cross-lingual generalization capabilities, there remain many open research questions, which we tackle here. Our study is organized as a "best practices" guide for training multilingual dense retrieval models, broken down into three main scenarios: when a

1 Introduction

This paper examines: Toward Best Practices for Training Multilingual Dense Retrieval Models. Research question: How does increasing the proportion of mined hard negatives in multilingual dense retrieval training datasets impact Mean Reciprocal Rank (MRR) on low-resource language benchmarks compared to standard negative sampling?.

2 Methodology

Systematic literature search across multiple databases yielded 2 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.7/10.

3 Results

2 papers retrieved. 10 claims extracted; 9 independently verified. Quality review score: 7.7/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Dense retrieval models using a transformer-based bi-encoder architecture have emerged as an active area of research.	✓	0.35
Recent work with multilingual transformers demonstrates that they exhibit strong cross-lingual generalization capabiliti	✓	0.30
The study is organized as a 'best practices' guide for training multilingual dense retrieval models, broken down into th	✓	0.40
The three main scenarios are: 'have model, no data'; 'have model and data'; and 'have data, no model'.	✓	0.21
The study gains a better understanding of the role of multi-stage fine-tuning.	✓	0.19
The study examines the strength of cross-lingual transfer under various conditions.	✓	0.18
The study explores the usefulness of out-of-language data.	×	0.13
The study compares the advantages of multilingual vs. monolingual transformers.	✓	0.17
The recommendations offer a guide for practitioners building search applications, particularly for low-resource language	✓	0.27
The work leaves open a number of research questions but provides a solid foundation for future work.	✓	0.26

References

- https://doi.org/10.1145/3613447
- https://doi.org/10.18653/v1/2025.findings-acl.543