Scaling Multilingual Models with Manifold-Aware Distance Metrics for Cross-Lingual Retrieval

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the effect of scaling multilingual models with manifold-aware distance metrics (e.g., MA-DPR) on cross-lingual retrieval performance across different language families in the MLQA benchmark,. Dense Passage Retrieval (DPR) typically relies on Euclidean or cosine distance to measure query-passage relevance in embedding space, which is effective when embeddings lie on a linear manifold. However, our experiments across DPR benchmarks suggest that embeddings often lie on. 15 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MA-DPR: Manifold-aware Distance Metrics for Dense Passage Retrieval. Research question: What is the effect of scaling multilingual models with manifold-aware distance metrics (e.g., MA-DPR) on cross-lingual retrieval performance across different language families in the MLQA benchmark, measured by accuracy improvements over cosine similarity baselines?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

14 papers retrieved. 15 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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
The system specifications include CPU—Intel(R) Core(TM) i7-14700HX and GPU—NVIDIA GeForce RTX 4070 Laptop GPU.	×	0.01
Average CPU utilization during measurement is $\sim$5%.	×	0.04
All codes and results are available online at github.com/QianfengWen/Manifold_Distance_Retrieval.git.	×	0.05
The experiments evaluate MA-DPR dManifold against several baselines including DPR with dEuclidean, DPR with dEuclidean +	×	0.08
The DPR benchmarks used are MS MARCO, NFCorpus, SciDocs, and ANTIQUE.	×	0.06
Two embedding models are used: msmarco-distilbert-base-tas-b (tas-b) and SciNCL.	×	0.04
MS MARCO is the in-distribution dataset for tas-b and SciDocs is the in-distribution dataset for SciNCL.	×	0.03
All embeddings are 2-normalized.	×	0.04
The evaluation metrics used are Recall, Mean Average Precision (MAP), and Normalized Discounted Cumulative Gain (nDCG) f	×	0.02
The research questions addressed are RQ1: Manifold Hypothesis Validation, RQ2: MA-DPR vs Baseline, RQ3: Design Choice Co	×	0.06
In Figure 2, the alignment between Euclidean distance and manifold-aware distance is evaluated for all query-passage pai	✓	0.24
In-distribution pairs exhibit strong agreement and relevance distinction using both distance metrics.	×	0.08
The remaining OOD settings show more misalignment, where manifold distance sometimes offers improved relevance distincti	×	0.12
The orange 'line' in the lower left of Figure 2 is due to relevant documents that are 1-hop away from the query in the m	×	0.10
The disconnected 'blobs' present in many plots correspond to different numbers of hops from the query in the manifold gr	×	0.05

References

- http://arxiv.org/abs/2408.10536v1
- http://arxiv.org/abs/2505.18673v1
- http://arxiv.org/abs/2509.13562v1