Query Generation Augmentation for Robust Dense Retrieval Across Language Families in Low-Resource XQuAD Subsets

Assignee Research

June 12, 2026

Abstract

The dual-encoder model is a dense retrieval architecture, consisting of two encoder models, that has surpassed traditional sparse retrieval methods for open-domain retrieval [1]. But, room exists for improvement, particularly when dense retrievers are exposed to unseen passages or queries. Considering out-of-domain queries, i.e., queries originating from domains other than the one the model was trained on, the loss in accuracy may be significant. A main factor for this is the mismatch in the information available to the context encoder and the query encoder during training. Common retireval tr

1 Introduction

This paper examines: Dense Passage Retrieval: Architectures and Augmentation Methods. Research question: Does the query generation augmentation method enhance robustness against language family divergence in dense retrievers as measured by retrieval accuracy drops on low-resource language subsets of XQuAD?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.2/10.

3 Results

16 papers retrieved. 7 claims extracted; 6 independently verified. Quality review score: 7.2/10.

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 dual-encoder model is a dense retrieval architecture consisting of two encoder models.	✓	0.30
The dual-encoder model has surpassed traditional sparse retrieval methods for open-domain retrieval.	✓	0.30
Dense retrievers experience significant loss in accuracy when exposed to out-of-domain queries.	×	0.15
A main factor for accuracy loss in out-of-domain scenarios is the mismatch in information available to the context encod	✓	0.24
Common retrieval training datasets contain an overwhelming majority of passages with only one query per passage.	✓	0.27
Training a DPR model on data containing multiple queries per passage improves the generalizability of the model.	✓	0.26
Training on passages with multiple queries leads to models that generalize better to out-of-distribution and out-of-doma	✓	0.33

References

- https://www.semanticscholar.org/paper/f11b9b89a6bf298881586b889c259d3c85287f57
- http://arxiv.org/abs/2311.05800v2
- http://arxiv.org/abs/2305.03950v1