Retrieval-Augmented Language Models in Knowledge-Intensive Task Performance

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does retrieval augmentation improve language model performance on knowledge-intensive tasks v14. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Multilingual Retrieval-Augmented Generation for Knowledge-Intensive Task. Research question: How does retrieval augmentation improve language model performance on knowledge-intensive tasks v14.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.0/10.

3 Results

12 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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 paper uses three multilingual open-domain question-answering tasks: MLQA, MKQA, and XOR-TyDi QA.	×	0.14
MLQA and MKQA are manually and machine-translated, whereas XOR-TyDi QA is translated by professional annotators.	×	0.01
The retrieval system used is Cohere with Wikimedia_dump as the database.	×	0.02
The multilingual embedding model Cohere_Embed_V3 is used for embedding individual articles.	×	0.03
The top-5 most relevant passages are retrieved and used as in-context knowledge during inference.	×	0.08
Two retrieval approaches are used: monolingual retrieval and multilingual retrieval.	×	0.14
Google Translate is used as the translation system for tRAG and CrossRAG in the main discussion.	×	0.08
Three different LLMs are used: GPT-4o, Llama-3-8b-instruct, and Command-R-35b.	×	0.03
Greedy decoding is used in all experiments with a temperature of 0 and a maximum generation length of 2048.	×	0.05
The paper uses flexible exact-match for evaluation.	×	0.02
English (en) has the highest percentage (46.3%) in the language distribution of the Wikimedia_dump.	×	0.03
The retrieval from W_En+SL for English shows 98.9% from English documents and 1.1% from other languages.	×	0.08
The retrieval from W_ALL for English shows 98.9% from English documents and 1.1% from other languages.	×	0.09

References

- http://arxiv.org/abs/2402.01176v2
- http://arxiv.org/abs/2504.03616v2
- http://arxiv.org/abs/2410.08876v2