Document Redundancy In Retrieval Corpora Impact The Answer Accuracy And Latency Of Joint Optimization Rag Frameworks On

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does document redundancy in retrieval corpora impact the answer accuracy and latency of joint optimization RAG frameworks on the Natural Questions benchmark. Retrieval-Augmented Generation (RAG) systems depend critically on the quality of document preprocessing, yet no prior study has evaluated PDF processing frameworks by their impact on downstream question-answering accuracy. We address this gap through a systematic comparison of. 23 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: From PDF to RAG-Ready: Evaluating Document Conversion Frameworks for Domain-Specific Question Answering. Research question: How does document redundancy in retrieval corpora impact the answer accuracy and latency of joint optimization RAG frameworks on the Natural Questions benchmark?.

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 3.8/10.

3 Results

14 papers retrieved. 23 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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
Docling provides excellent Markdown conversion with the hierarchical PDF strategy.	×	0.13
DeepSeek OCR achieves great precision in extracting native text and complex tables.	×	0.03
The system uses a benchmark dataset of 50 questions to evaluate RAG accuracy.	×	0.07
The LLM used in the experiments was gpt-4o-mini.	×	0.05
The embedding model used was text-embedding-3-small.	×	0.05
Marker delivers excellent results with complex tables and preserves document hierarchy.	×	0.04
High K value (200) was chosen to ensure relevant information is available in retrieval context.	×	0.04
LangChain's PDFLoader sometimes misreads text segments.	×	0.02
Some questions target text segments that LangChain's PDFLoader reads incorrectly.	×	0.04
Document structure is preserved using Docling-hierarchical-pdf.	×	0.10
DeepSeek OCR requires page-to-image conversion and loses document structure.	×	0.07
Configuration-driven pipeline allows transparent selection through the ingestion orchestrator.	×	0.02
All experiments used a fixed RAG configuration to isolate the effect of the data preparation pipeline.	×	0.07
Chunk size used in the experiments was 1000 characters with a 200-character overlap.	×	0.03
Document hierarchy is sometimes mistaken by Docling, requiring additional visual checking.	×	0.04
The research team determined expected answers by analyzing source documents.	×	0.03
High hardware requirements are required for DeepSeek OCR.	×	0.03
Portuguese " is misinterpreted in some cases.	×	0.02
Document hierarchy is lost in some conversion methods.	×	0.05
Cloud-only constraint affects some conversion tools.	×	0.03
Configuration 21 (HTML, K = 200) achieved 79.1% accuracy.	×	0.05
Configuration 20 (Hierarchical + Images, K = 200) achieved 94.1% accuracy.	×	0.07
Configuration 11 (full cleaning, K = 200) achieved 83.2% accuracy.	×	0.05

References

- http://arxiv.org/abs/2604.04948v2
- http://arxiv.org/abs/2409.03708v2
- http://arxiv.org/abs/2502.11228v2