Hybrid Embeddings in Tree of Reviews Enhance Robustness in Multi-Hop QA Retrieval

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 1 peer-reviewed paper addressing the following research question: Does the use of hybrid embeddings in Tree of Reviews improve robustness against distractor documents in multi-hop QA datasets compared to single-embedding retrieval methods. The Portable Document Format (PDF) is widely used for enterprise information communication and archival, but its emphasis on visual fidelity presents major barriers for ingestion into Large Language Model (LLM)-based systems. High-quality data ingestion is critical for. 12 claims were extracted from source literature; 11 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.9/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: OPTIMIZING PDF INGESTION FOR LARGE LANGUAGE MODELS IN RAG ARCHITECTURES. Research question: Does the use of hybrid embeddings in Tree of Reviews improve robustness against distractor documents in multi-hop QA datasets compared to single-embedding retrieval methods?.

2 Methodology

Systematic literature search across multiple databases yielded 1 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.9/10.

3 Results

1 papers retrieved. 12 claims extracted; 11 independently verified. Quality review score: 6.9/10.

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 Portable Document Format (PDF) is widely used for enterprise information communication and archival.	✓	0.24
PDF's emphasis on visual fidelity presents major barriers for ingestion into Large Language Model (LLM)-based systems.	✓	0.34
High-quality data ingestion is critical for Retrieval-Augmented Generation (RAG) systems.	✓	0.25
RAG systems increasingly rely on unstructured organizational knowledge.	✓	0.21
Complex PDFs often suffer from context loss and semantic degradation during extraction.	✓	0.20
Context loss and semantic degradation in PDF extraction impair RAG performance.	✓	0.15
Current parsing techniques are often evaluated on simplified benchmarks.	✓	0.19
There is a gap between the capabilities of current parsing techniques and the needs of real-world enterprise documents.	✓	0.24
Key challenges in PDF ingestion include layout interpretation, contextualization of tables and images, OCR noise reducti	✓	0.30
Existing approaches to PDF parsing are categorized into pipeline-based methods, holistic Vision-Language Models (VLMs),	✓	0.31
Analysis of reported performance reveals persistent gaps between model accuracy and human-level understanding in complex	✓	0.28
Current benchmarks for PDF parsing have limitations.	×	0.11

References

- https://www.semanticscholar.org/paper/579afdbff86315a55bc7418cb2c54ac1b1cd0723