Hierarchical vs. Sliding Window Chunking in Longformer Long-Range Reasoning on HotpotQA

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does hierarchical chunking compare to sliding window strategies in preserving long-range reasoning accuracy for Longformer on the HotpotQA benchmark. The effectiveness of Retrieval-Augmented Generation (RAG) is highly dependent on how documents are chunked, that is, segmented into smaller units for indexing and retrieval. Yet, commonly used "one-size-fits-all" approaches often fail to capture the nuanced structure and. 13 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Adaptive Chunking: Optimizing Chunking-Method Selection for RAG. Research question: How does hierarchical chunking compare to sliding window strategies in preserving long-range reasoning accuracy for Longformer on the HotpotQA benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.0/10.

3 Results

10 papers retrieved. 13 claims extracted; 1 independently verified. Quality review score: 6.0/10.

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
Most studies measure chunking quality extrinsically via downstream retrieval metrics such as Hits@k, Recall@K, and Norma	×	0.06
There is a pressing need for robust, intrinsic metrics capable of directly assessing chunk quality.	×	0.03
Ideal chunks should be self-contained and logically complete, respectful of length constraints, semantically cohesive, a	×	0.04
Breaking logical units scatters essential information across chunks.	×	0.01
Oversized or undersized chunks degrade embedding quality and waste retrieval slots.	×	0.02
Mixing unrelated topics dilutes semantic signals.	×	0.02
Adaptive Chunking dynamically selects the best chunking method for each document based on the average of five intrinsic	✓	0.17
LangChain Recursive Splitter is a widely used baseline in retrieval systems.	×	0.07
Page-based Chunking is the simplest and most common baseline in production pipelines.	×	0.04
For each document in the corpus, three question–answer pairs are generated using GPT-4.1 with a 10,000-token context win	×	0.03
Retrieval Completeness measures how well the retrieved context supports the ground truth answer.	×	0.02
The corpus consists of 33 PDF documents from the Ekimetrics CLAIR project, spanning the legal, technical, and social sci	×	0.12
The documents were parsed into markdown text using Microsoft Azure AI Document Intelligence.	×	0.03

References

- http://arxiv.org/abs/1101.5396v2
- http://arxiv.org/abs/2603.25333v1
- http://arxiv.org/abs/1503.02466v1