Dynamic Chunking vs. Fixed-Size Segmentation in RAG Systems for QuranQA Retrieval and Faithfulness

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does dynamic chunking based on semantic segmentation compare to fixed-size (sentence/paragraph) chunking in terms of retrieval recall and answer faithfulness in RAG systems evaluated on the. 16 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Rethinking Chunk Size For Long-Document Retrieval: A Multi-Dataset Analysis. Research question: How does dynamic chunking based on semantic segmentation compare to fixed-size (sentence/paragraph) chunking in terms of retrieval recall and answer faithfulness in RAG systems evaluated on the QuranQA benchmark?.

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 4.2/10.

3 Results

12 papers retrieved. 16 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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 study evaluates retrieval performance using a fixed-size chunking strategy with chunk sizes of 64, 128, 256, 512, an	✓	0.16
The experiments were conducted without overlapping tokens.	×	0.07
The embedding models used in the experiments are stella_en_1.5B_v5 and snowflake-arctic-embed-l-v2.0.	×	0.08
In the SQuAD dataset, 64-token chunks yield a Recall@1 of 64.1%.	×	0.05
In the SQuAD dataset, increasing chunk size to 512 tokens reduces Recall@1 by 10-15% compared to smaller chunks.	×	0.10
In the NewsQA dataset, peak Recall@1 is achieved at a chunk size of 512 tokens with a value of 55.9%.	×	0.06
In the NarrativeQA dataset, Recall@1 increases from 4.2% at 64 tokens to 10.7% at 1024 tokens.	×	0.05
In the TechQA dataset, Recall@1 improves from 16.5% at 128 tokens to 61.3% at 512 tokens.	×	0.06
For the COVID-QA dataset using the Stella model, the highest Recall@1 is 52.1% at a chunk size of 64 tokens.	×	0.07
For the COVID-QA dataset using the Snowflake model, Recall@1 peaks at 54.2% with a chunk size of 1024 tokens.	×	0.08
For the COVID-QA dataset using the Snowflake model, Recall@5 is 80.2% at a chunk size of 1024 tokens.	×	0.09
The RAG system used in the study is built using LlamaIndex.	×	0.04
Document segmentation was performed using LlamaIndex's TokenTextSplitter.	×	0.04
Retrieval is performed by calculating cosine similarity between the query embedding and chunk embeddings.	×	0.09
The study utilizes the datasets NarrativeQA, Natural Questions (NQ), NewsQA, COVID-QA, TechQA, and SQuAD.	×	0.03
Documents were filtered by performing a string match comparison between the expected answer and the document text to ens	×	0.04

References

- http://arxiv.org/abs/2505.21700v2
- http://arxiv.org/abs/2504.19754v1
- http://arxiv.org/abs/2605.22834v2