

Scaling Retrieval-Augmented Generation for Clinical Notes: Latency-Accuracy Trade-offs

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the trade-off between retrieval latency and generation accuracy in RAG systems when scaling from millions to billions of clinical notes, as measured by response time and NaturalQuestions. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: To Memorize or to Retrieve: Scaling Laws for RAG-Considerate Pretraining. Research question: What is the trade-off between retrieval latency and generation accuracy in RAG systems when scaling from millions to billions of clinical notes, as measured by response time and NaturalQuestions benchmark scores?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.8/10.

3 Results

13 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 7.8/10.

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.

References

- <http://arxiv.org/abs/2404.07220v2>
- <http://arxiv.org/abs/2504.01346v4>
- <http://arxiv.org/abs/2604.00715v1>