What is the trade-off between retrieval depth (top-k) and inferencing speed (latency) for 7B models augmented

Assignee Research

June 10, 2026

Abstract

Retrieval-Augmented Generation (RAG) improves the factual accuracy of large language models by combining document retrieval with text generation. In biomedical question answering, where correctness is critical, the effect of key hyperparameters has not been studied in a systematic way. This paper presents an evaluation of RAG on the COVID-QA dataset with a focus on three retrievers (dense, BM25, hybrid), two retrieval depths (top- $\mathrm{k}=1,3$), and optional reranking with a cross encoder. We use a single biomedical prompt and measure exact match (EM), F1 score, semantic similarity, ground

1 Introduction

This paper examines: HYPER-RAG: Evaluating Hyperparameter Trade-Offs in Biomedical Retrieval-Augmented Generation. Research question: What is the trade-off between retrieval depth (top-k) and inferencing speed (latency) for 7B models augmented with hybrid query-based retrievers on the QuAC benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.3/10.

3 Results

16 papers retrieved. 9 claims extracted; 9 independently verified. Quality review score: 7.3/10.

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
Retrieval-Augmented Generation (RAG) improves the factual accuracy of large language models by combining document retrie	✓	0.35
The effect of key hyperparameters in RAG has not been studied systematically in biomedical question answering.	✓	0.19
This paper evaluates RAG on the COVID-QA dataset with a focus on three retrievers (dense, BM25, hybrid), two retrieval d	✓	0.33
The evaluation uses a single biomedical prompt and measures exact match (EM), F1 score, semantic similarity, groundednes	✓	0.30
A composite score is reported that balances lexical accuracy, semantic similarity, and efficiency.	✓	0.24
Results on a 100-question subset show that reranking improves grounding at the cost of extra latency.	✓	0.26
Increasing top-k improves recall but gives smaller gains after a point.	✓	0.25
Multiple metrics are needed to judge biomedical RAG systems reliably.	✓	0.27
Careful tuning of retrieval and reranking settings can yield practical improvements under compute constraints.	✓	0.30

References

- https://www.semanticscholar.org/paper/21ed39c3e4117e74925430d0e2ba97df5ed71d05
- http://arxiv.org/abs/2404.07220v2
- http://arxiv.org/abs/2504.01346v4