Model Size and Retrieval Method Effects on FAITH Scores in Religious Domain RAG

Assignee Research

June 12, 2026

Abstract

Retrieval-Augmented Generation (RAG) is a prevalent approach to infuse a private knowledge base of documents with Large Language Models (LLM) to build Generative Q\&A (Question-Answering) systems. However, RAG accuracy becomes increasingly challenging as the corpus of documents scales up, with Retrievers playing an outsized role in the overall RAG accuracy by extracting the most relevant document from the corpus to provide context to the LLM. In this paper, we propose the 'Blended RAG' method of leveraging semantic search techniques, such as Dense Vector indexes and Sparse Encoder indexes, ble

1 Introduction

This paper examines: Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers. Research question: What is the impact of model size (7B vs. 70B) on the FAITH score in retrieval-augmented generation (RAG) for religious domain QA, particularly when comparing dense retrieval vs. sparse retrieval methods?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.5/10.

3 Results

11 papers retrieved. 23 claims extracted; 17 independently verified. Quality review score: 7.5/10.

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
On the NQDataset, the Blended RAG approach achieved a P@20 score of 0.633.	×	0.09
On the NQDataset, the Blended RAG approach achieved an F1 score of 79.6.	×	0.07
On the Trec Covid dataset, the Blended RAG approach achieved an NDCG@10 score of 80.4.	✓	0.22
On the HotpotQA dataset, the Blended RAG approach achieved an F1 and EM score of 0.85.	×	0.10
The study evaluated three search strategies: keyword-based similarity search, dense vector-based search, and semantic-ba	✓	0.19
The study utilized BM25 for keyword-based indexing.	×	0.10
The study utilized KNN for vector-based indexing.	×	0.08
The study utilized Elastic Learned Sparse Encoder (ELSER) for sparse encoder-based semantic search.	✓	0.20
All RAG pipeline experiments were performed using the FLAN-T5-XXL model to avoid the effect of LLM size or type.	✓	0.21
On the Trec-covid dataset, the Blended RAG pipeline achieved an NDCG@10 score of 0.87.	✓	0.28
On the Trec-covid dataset, the COCO-DR Large model achieved an NDCG@10 score of 0.804.	✓	0.29
On the NQ dataset, the monoT5-3B model achieved an NDCG@10 score of 0.633.	✓	0.29
On the NQ dataset, the Blended RAG pipeline achieved an NDCG@10 score of 0.67.	✓	0.29
On the SqUAD dataset, the KNN+BF configuration achieved a Top-5 retrieval accuracy of 94.89.	✓	0.16
On the SqUAD dataset, the KNN+BF configuration achieved a Top-10 retrieval accuracy of 97.43.	✓	0.16
On the SqUAD dataset, the KNN+BF configuration achieved a Top-20 retrieval accuracy of 98.58.	✓	0.15
On the COQA dataset, the SE+BF (Sparse Encoder + Best Fields) configuration achieved a Top-5 retrieval accuracy of 49.94	×	0.11
The RAG-original variant achieved an Exact Match (EM) score of 28.12 and an F1 score of 39.42.	✓	0.18
The RAG-end2end variant achieved an Exact Match (EM) score of 40.02 and an F1 score of 52.63.	✓	0.17
The Blended RAG pipeline achieved an Exact Match (EM) score of 57.63 and an F1 score of 68.4	✓	0.16

References

- http://arxiv.org/abs/2507.23334v2
- http://arxiv.org/abs/2404.07220v2
- http://arxiv.org/abs/2510.22344v1