Blended RAG Performance Scaling Across Multi-Domain Benchmarks and Dataset Sizes

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 5 peer-reviewed papers addressing the following research question: How does the performance of Blended RAG scale with increasing dataset sizes on multi-domain benchmarks like MMLU or HELM, compared to baseline RAG methods, when evaluated using exact match accuracy. Retrieval-Augmented Generation (RAG) enhances Large Language Models (LLMs) by integrating them with an external knowledge base to improve the answer relevance and accuracy. In real-world scenarios, beyond pure text, a substantial amount of knowledge is stored in tables, and user. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: RAG over Tables: Hierarchical Memory Index, Multi-Stage Retrieval, and Benchmarking. Research question: How does the performance of Blended RAG scale with increasing dataset sizes on multi-domain benchmarks like MMLU or HELM, compared to baseline RAG methods, when evaluated using exact match accuracy?.

2 Methodology

Systematic literature search across multiple databases yielded 5 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

5 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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
T-RAG achieves accuracy improvements ranging from 1.2% to 11.4% and recall gains from 1.5% to 12.5% when compared to tab	×	0.06
T-RAG achieves improvements of up to 9.4% in recall@50 on TFV and 8.2% in recall@10 on Multi-hop TQA compared to Table-t	×	0.09
T-RAG consistently improves cross-table question answering performance, yielding an average gain of 11.2% compared to th	×	0.08
The latency of T-RAG for TFV is 133.1 minutes, for Single-hop TQA is 78.8 minutes, and for Multi-hop TQA is 34.6 minutes	×	0.04
The number of tables remaining after coarse-grained multi-head retrieval for TFV is 4204, for Single-hop TQA is 2184, an	×	0.04
The number of tables remaining after fine-grained subgraph retrieval for TFV, Single-hop TQA, and Multi-hop TQA is 10.	×	0.04
The accuracy of DTR on TFV is 21.1% at 10, 27.8% at 20, and 36.2% at 50.	×	0.02
The recall of Table-LLaMA on Multi-hop TQA is 45.8% at 10, 49.1% at 20, and 61.8% at 50.	×	0.03
The EM@10 of Phi-3.5-mini on TFV is 22.3, on Single-hop TQA is 26.2, and on Multi-hop TQA is 13.9.	×	0.04
The F1@50 of LLaMA-3.1-70B on TFV is 62.1, on Single-hop TQA is 50.2, and on Multi-hop TQA is 50.2.	×	0.05

References

- http://arxiv.org/abs/2402.07483v2
- http://arxiv.org/abs/2510.25518v1
- http://arxiv.org/abs/2504.01346v4