SOVEREIGN: What is the impact of synthetic data generation scaling (e.g., number of training examples) on retriever MRR@1

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Retrieval-augmented generation (RAG) enhances large language models (LLMs) with external knowledge to answer questions more accurately. However, research on evaluating RAG systems-particularly the retriever component-remains limited, as most existing work focuses on single-context retrieval rather than multi-hop queries, where individual contexts may appear irrelevant in isolation but are essential when combined. In this research, we use the HotPotQA, MuSiQue, and SQuAD datasets to simulate a RAG system and compare three LLM-as-judge evaluation strategies, including our proposed Context-Awar

1 Introduction

Analysis of: Evaluating Multi-Hop Reasoning in RAG Systems: A Comparison of LLM-Based Retriever Evaluation Strategies. Research goal: What is the impact of synthetic data generation scaling (e.g., number of training examples) on retriever MRR@10 for multi-hop queries compared to single-hop queries in domain-specific RAG optimization?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

8 papers retrieved. 9 claims extracted, 4 verified. Tribunal: 6.7/10 → REVISE (revision_round=1). Policy: ESCALATE_TO_OWNER.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
The CARE method consistently outperforms existing methods for evaluating multi-hop reasoning in RAG systems.	✓	0.34
Experiments with LLMs from OpenAI, Meta, and Google demonstrate that CARE consistently outperforms existing methods.	✓	0.25
The performance gains of CARE are most pronounced in models with larger parameter counts and longer context windows.	✓	0.23
Single-hop queries show minimal sensitivity to context-aware evaluation.	✓	0.30
On the HotPotQA dataset, the CARE method achieved an accuracy of 0.827±0.02.	×	0.04
On the MuSiQue dataset, the CARE method achieved an F1-Score of 0.678±0.03.	×	0.03
The indirect evaluation approach led to a significant improvement in F1-Score for the small LLaMa model.	×	0.02
The direct approach resulted in a decline in F1-Score for the reasoning model o4-mini.	×	0.02
CARE consistently outperformed other approaches across all models except for the LLaMa 3.1-8b model.	×	0.04

References

- http://arxiv.org/abs/2604.18234v1
- http://arxiv.org/abs/2507.23334v2

- http://arxiv.org/abs/2404.14464v1