Llama-2-7B and Llama-3-8B Retrieval Performance on SQuAD 2.0 Under Shrinking Context Windows

Assignee Research

May 29, 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

1 Introduction

This paper examines: Evaluating Multi-Hop Reasoning in RAG Systems: A Comparison of LLM-Based Retriever Evaluation Strategies. Research question: How do Llama-2-7B and Llama-3-8B models compare on SQuAD 2.0 retrieval performance when context window size decreases from 4096 to 1024 tokens?.

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 5.2/10.

3 Results

16 papers retrieved. 13 claims extracted; 4 independently verified. Quality review score: 5.2/10.

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
CARE consistently outperforms existing methods for evaluating multi-hop reasoning in RAG systems.	✓	0.40
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.	✓	0.32
The indirect evaluation approach led to a significant improvement in F1-Score for the small LLaMa model.	×	0.05
The direct approach resulted in a decline in F1-Score for the reasoning model o4-mini.	×	0.02
For CARE, the reasoning model o4-mini exhibited a decrease in accuracy, F1-Score, and recall compared to GPT-4.1, while	×	0.04
The LLaMa 3.1-8b model experienced a significant decline in overall performance, with substantial drops in both F1-Score	×	0.05
CARE consistently outperformed other approaches across all models except for the LLaMa 3.1-8b model.	×	0.08
The indirect method [24] labels a context i as non-relevant if the LLM I cannot answer the query q using only i.	×	0.03
The direct method [23] evaluates whether a context i is crucial to answering a query q with the ground-truth answer a*.	×	0.03
The CARE method evaluates whether a context i is crucial to answering a query q with the ground-truth answer a*, given	×	0.04
The experiments were conducted using the HotPotQA, MuSiQue, and SQuAD datasets.	×	0.12
The complete data of the experiments is available at https://github.com/lorenzbrehme/CARE.	✓	0.17

References

- http://arxiv.org/abs/2506.08827v1

- http://arxiv.org/abs/2504.19754v1
- http://arxiv.org/abs/2604.18234v1