Chain-Based Retrieval Accuracy of Llama-3-8B-128K vs. Qwen-8B and Mistral-8B on BABILong

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does the chain-based retrieval accuracy of Llama-3-8B-128K compare to Qwen-8B and Mistral-8B on HotPotQA when varying the maximum context length from 32K to 128K. In recent years, the input context sizes of large language models (LLMs) have increased dramatically. However, existing evaluation methods have not kept pace, failing to comprehensively assess the efficiency of models in handling long contexts. 13 claims were extracted from source literature; 13 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 9.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: BABILong: Testing the Limits of LLMs with Long Context Reasoning-in-a-Haystack. Research question: How does the chain-based retrieval accuracy of Llama-3-8B-128K compare to Qwen-8B and Mistral-8B on HotPotQA when varying the maximum context length from 32K to 128K?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 9.2/10.

3 Results

15 papers retrieved. 13 claims extracted; 13 independently verified. Quality review score: 9.2/10.

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
The input context sizes of large language models (LLMs) have increased dramatically in recent years.	✓	0.26
Existing evaluation methods fail to comprehensively assess the efficiency of models in handling long contexts.	✓	0.23
The BABILong benchmark is designed to test language models' ability to reason across facts distributed in extremely long	✓	0.32
BABILong includes a diverse set of 20 reasoning tasks.	✓	0.24
BABILong reasoning tasks include fact chaining, simple induction, deduction, counting, and handling lists/sets.	✓	0.27
Evaluations show that popular LLMs effectively utilize only 10-20% of the context.	✓	0.25
Popular LLM performance declines sharply with increased reasoning complexity.	✓	0.21
Retrieval-Augmented Generation methods achieve a modest 60% accuracy on single-fact question answering.	✓	0.27
Retrieval-Augmented Generation accuracy on single-fact question answering is independent of context length.	✓	0.23
Among context extension methods, recurrent memory transformers demonstrate the highest performance after fine-tuning.	✓	0.20
Fine-tuned recurrent memory transformers enable the processing of lengths up to 50 million tokens.	✓	0.18
The BABILong benchmark is extendable to any length.	✓	0.20
The authors provide BABILong benchmark splits up to 10 million token lengths.	✓	0.22

References

- https://doi.org/10.48550/arxiv.2406.07887
- https://doi.org/10.48550/arxiv.2406.10149

- https://doi.org/10.48550/arxiv.2505.10063