Fact-Chaining Accuracy of Llama-3-8B-128K vs. Qwen-8B and Mistral-8B on BABILong Across Context Lengths

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How does the fact-chaining accuracy of Llama-3-8B-128K compare to Qwen-8B and Mistral-8B on the BABILong benchmark when context length increases 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. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. 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 fact-chaining accuracy of Llama-3-8B-128K compare to Qwen-8B and Mistral-8B on the BABILong benchmark when context length increases 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 4.5/10.

3 Results

15 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 4.5/10.

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.

References

- http://arxiv.org/abs/2604.05114v1
- http://arxiv.org/abs/2605.13831v1
- http://arxiv.org/abs/2406.10149v2