Parametric-Retrieval Ratio Effects on Llama-3 Needle-in-a-Haystack Accuracy

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How does varying the ratio of parametric parameters to retrieval context size impact exact match accuracy on the Needle-in-a-Haystack benchmark for Llama-3 variants trained with mixed objective. 13 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Multilingual Needle in a Haystack: Investigating Long-Context Behavior of Multilingual Large Language Models. Research question: How does varying the ratio of parametric parameters to retrieval context size impact exact match accuracy on the Needle-in-a-Haystack benchmark for Llama-3 variants trained with mixed objective functions?.

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 4.7/10.

3 Results

16 papers retrieved. 13 claims extracted; 3 independently verified. Quality review score: 4.7/10.

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
This study is the first to investigate the multilingual long-context behavior of Large Language Models (LLMs).	✓	0.34
The Multilingual Needle in a Haystack (MLNeedle) test assesses model performance across seven languages.	✓	0.21
The MLNeedle benchmark evaluates models in both monolingual and cross-lingual settings.	✓	0.15
The experimental setup involves providing a model with a question Q and K documents, where exactly one document contains	×	0.03
The experiments systematically vary the position of the needle, the language of the needle, and the language of the hays	×	0.12
Table (p4) reports a performance score of 0.335 for Llama2-7B-Chat at 4K context length.	×	0.06
Table (p4) reports a performance score of 0.700 for Llama3-8B-Instruct at 8K context length with a 4K needle position.	×	0.07
Table (p5) shows that Mistral-7B-Instruct-v0.2 achieved a score of 0.68 when both the needle and haystack were in Englis	×	0.04
Table (p5) shows that Mistral-7B-Instruct-v0.2 achieved a score of 0.24 when the needle was in Arabic and the haystack w	×	0.03
Table (p5) shows that Llama3-8B-Instruct achieved a score of 0.61 when both the needle and haystack were in English.	×	0.04
Table (p5) shows that Llama3-8B-Instruct achieved a score of 0.15 when the needle was in Chinese and the haystack was in	×	0.03
The study performed ablation studies on the effects of temperature sampling, instruction tuning, and evaluation metrics.	×	0.03
The source code and datasets for the study are available at https://github.com/AmeyHengle/multilingual-needle-in-a-hayst	×	0.08

References

- http://arxiv.org/abs/2408.10151v1
- http://arxiv.org/abs/2411.19360v1
- http://arxiv.org/abs/2505.14302v1