# Attention Mechanism Variants in Llama-3 and Their Impact on Multilingual Retrieval Robustness

Assignee Research

June 9, 2026

**Abstract**

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How do different attention mechanism variants in Llama-3 impact the robustness of information retrieval across varying context lengths in the MultiLingual Needle-in-a-Haystack evaluation. 3 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: Can LLMs reason over extended multilingual contexts? Towards long-context evaluation beyond retrieval and haystacks. Research question: How do different attention mechanism variants in Llama-3 impact the robustness of information retrieval across varying context lengths in the MultiLingual Needle-in-a-Haystack evaluation?.

## 2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

## 3 Results

11 papers retrieved. 3 claims extracted; 0 independently verified. Quality review score: 3.7/10.

## 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 |
|---|---|---|
| The best-performing method at baseline, CoT, achieves only a moderate 60% success rate on MLRBench. | × | 0.02 |
| Prompt-only methods struggle to maintain accuracy even at moderate contexts. | × | 0.03 |
| Among prompt-only methods, CoT and FS outperform ZS and ICT across all context lengths and languages consistently, with | × | 0.04 |

## References


- http://arxiv.org/abs/2504.12845v1
- http://arxiv.org/abs/2408.10151v1
- http://arxiv.org/abs/2601.15305v1