Language Mixing Degrades Long-Context Recall in Llama-3 Multilingual Retrieval

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: To what extent does language mixing within the retrieval context degrade long-context recall performance of Llama-3 models on the MultiLingual Needle-in-a-Haystack test compared to monolingual. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.5/10. 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: To what extent does language mixing within the retrieval context degrade long-context recall performance of Llama-3 models on the MultiLingual Needle-in-a-Haystack test compared to monolingual settings?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.5/10.

3 Results

13 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.5/10.

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/2409.18006v3
- http://arxiv.org/abs/2504.12845v1
- http://arxiv.org/abs/2408.10151v1