Mistral 7B and Llama 3.1 Inference Performance in Multi-Agent RAG for PyPI Security Analysis

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: What is the difference in inference latency and token throughput between Llama 3.1 and Mistral 7B when integrated into a multi-agent system for analyzing PyPI package security with RAG. 15 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Collaborative Multi-Agent Approach to Retrieval-Augmented Generation Across Diverse Data. Research question: What is the difference in inference latency and token throughput between Llama 3.1 and Mistral 7B when integrated into a multi-agent system for analyzing PyPI package security with RAG?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

10 papers retrieved. 15 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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
Mistral 7B has a context window of 4096 - 16K (Sliding Windows).	✓	0.17
Zephyr has a context window of 8192.	×	0.01
Phi-2 has a context window of 2048.	×	0.01
Llama 3 has a context window of 8192.	×	0.08
GPT-4 has a context window of 128,000.	×	0.01
GPT-4 Turbo has a context window of 128,000.	×	0.01
GPT-3.5 Turbo has a context window of 16,385.	×	0.01
Gemini 1.5 Flash has a context window of 1,048,576.	×	0.01
Gemini 1.5 Pro has a context window of 2,097,152.	×	0.01
Local models are better suited for scenarios where data privacy and control are paramount.	×	0.06
API-based models excel in environments that demand scalability and ease of integration.	×	0.05
Hybrid approaches that combine local processing for sensitive data with cloud-based solutions for less critical tasks re	×	0.04
Prompt engineering is a crucial aspect of the Multi-Agent RAG System, as it directly influences the quality and accuracy	×	0.08
Few-shot prompting is particularly effective in this context, as it provides the agent with examples of how to interact	×	0.04
A well-designed prompt includes the user's query, the database schema, and a few-shot example illustrating how similar q	×	0.04

References

- http://arxiv.org/abs/2107.12699v2

- http://arxiv.org/abs/2305.16615v1
- http://arxiv.org/abs/2412.05838v1