RAG Integration Impact on Llama3.1 and Mistral 7B Latency and Memory Efficiency in Battery Management Systems

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 1 peer-reviewed paper addressing the following research question: How does the retrieval-augmented generation (RAG) integration affect the inference latency and memory efficiency of Llama3.1 compared to Mistral 7B on cyber-physical system battery management. 8 claims were extracted from source literature; 8 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Review of Large Language Models for Energy Systems: Applications, Challenges, and Future Prospects. Research question: How does the retrieval-augmented generation (RAG) integration affect the inference latency and memory efficiency of Llama3.1 compared to Mistral 7B on cyber-physical system battery management datasets, measured in tokens per second and GPU memory usage?.

2 Methodology

Systematic literature search across multiple databases yielded 1 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.3/10.

3 Results

1 papers retrieved. 8 claims extracted; 8 independently verified. Quality review score: 8.3/10.

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
Rapid developments in large language models (LLMs) have created new opportunities for their use in the energy sector, fr	✓	0.42
These models help improve decision-making, anomaly detection, and optimization procedures in intricate energy systems by	✓	0.38
This study provides a comprehensive review of the LLM origins, evaluation, and fine-tuning techniques as well as their i	✓	0.48
Their performance in terms of explainability, generalization ability, and scalability for energy-related applications is	✓	0.30
The report also emphasizes significant challenges to the adoption of LLMs, such as the need for computing power, the lac	✓	0.35
Power-efficient models, hybrid artificial intelligence (AI) platforms, and domain-specific fine-tuning are some of the s	✓	0.33
Future areas of interest include multi-modality to obtain maximal forecasting and operational intelligence, real-time ad	✓	0.31
This paper summarizes current developments and provides information on LLM-driven innovation in energy systems while mai	✓	0.34

References

- https://doi.org/10.1109/access.2025.3610994