Retrieval-Augmented Llama3-70B and Llama-13B in Code Vulnerability Classification: Efficiency and Accuracy Trade-offs

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the computational efficiency trade-off when applying retrieval augmentation to Llama3-70B for code vulnerability classification, and how does it compare to smaller models like Llama-13B in. With many real-world applications of Natural Language Processing (NLP) comprising of long texts, there has been a rise in NLP benchmarks that measure the accuracy of models that can handle longer input sequences. However, these benchmarks do not consider the trade-offs between. 11 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Characterizing the Efficiency vs. Accuracy Trade-off for Long-Context NLP Models. Research question: What is the computational efficiency trade-off when applying retrieval augmentation to Llama3-70B for code vulnerability classification, and how does it compare to smaller models like Llama-13B in terms of inference latency and accuracy?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.8/10.

3 Results

14 papers retrieved. 11 claims extracted; 3 independently verified. Quality review score: 5.8/10.

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
GovReport, SummScreenFD, and QMSum are evaluated using Rouge, as is standard for summarization.	×	0.04
Qasper is evaluated using a token-level F1 score after normalizing both the predicted and ground-truth answer strings.	×	0.03
For Rouge, following SCROLLS, we calculated the geometric mean of three different types of rouge to provide a single val	×	0.02
For efficiency metrics, we explored the training power efficiency (number of samples trained per second per Watt), total	×	0.06
The training and inference speeds are provided by the HuggingFace library while the total energy consumed and the power	×	0.04
Power efficiency has a strong inverse correlation with the size of the input sequence lengths, with small variations acr	×	0.08
Big Bird-large model has similar power efficiency to LED-large model across the input sequence lengths, but Big Bird's R	✓	0.17
On GovReport and QMSum, LED-large with sequence length 1024 is more efficient and has higher accuracy than each of the L	✓	0.15
Increasing the sequence length for LED-large further increases this accuracy while still often being more efficient than	×	0.14
Increasing model size is a more energy efficient way of increasing accuracy as compared to increasing sequence length fo	✓	0.18
If inference speed is the main efficiency metric of interest, then smaller models should be preferred.	×	0.09

References

- http://arxiv.org/abs/2402.12317v2
- http://arxiv.org/abs/2204.07288v1
- http://arxiv.org/abs/2408.15301v2