FlowKV and SmoothEvict Memory-Accuracy Trade-offs at 500K-Token Scale in Llama-3-70B

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does the trade-off between memory efficiency and accuracy compare between FlowKV and SmoothEvict when scaling to 500K+ tokens on Llama-3-70B across diverse LongBench tasks. 11 claims were extracted from source literature; 11 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Architecture Selection via the Trade-off Between Accuracy and Robustness. Research question: How does the trade-off between memory efficiency and accuracy compare between FlowKV and SmoothEvict when scaling to 500K+ tokens on Llama-3-70B across diverse LongBench tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.7/10.

3 Results

4 papers retrieved. 11 claims extracted; 11 independently verified. Quality review score: 8.7/10.

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 paper provides a general framework for characterizing the trade-off between accuracy and robustness in supervised le	✓	0.31
The authors propose a method and define quantities to characterize the trade-off between accuracy and robustness for a g	✓	0.39
The authors introduce a simple trade-off curve.	✓	0.20
The authors define and study an influence function that captures the sensitivity, under adversarial attack, of the optim	✓	0.36
Adversarial training regularizes the parameters in an over-parameterized linear model.	✓	0.26
In the context of adversarial training on over-parameterized linear models, LASSO and ridge regression are recovered as	✓	0.20
The proposed framework allows for the theoretical analysis of the behavior of the trade-off curve.	✓	0.15
Experiments demonstrate trade-off curves of neural networks.	✓	0.19
The demonstrated trade-off curves of neural networks vary with respect to the number of layers.	✓	0.21
The demonstrated trade-off curves of neural networks vary with respect to the number of neurons.	✓	0.19
The demonstrated trade-off curves of neural networks vary across different network structures.	✓	0.21

References

- http://arxiv.org/abs/2005.06339v2
- http://arxiv.org/abs/2408.11848v2

- http://arxiv.org/abs/1906.01354v2