FlowKV Token Eviction Performance in LongBench Code Completion for 70B Models

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: Does the token eviction strategy in FlowKV maintain performance on code completion tasks within LongBench relative to full-context baselines for 70B parameter models. 5 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Mask Tokens as Prophet: Fine-Grained Cache Eviction for Efficient dLLM Inference. Research question: Does the token eviction strategy in FlowKV maintain performance on code completion tasks within LongBench relative to full-context baselines for 70B parameter models?.

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 3.8/10.

3 Results

10 papers retrieved. 5 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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
Under the extreme 32 KV budget, the method outperforms the best competing baseline by 7.02 points on LLaDA-8B.	×	0.03
With a 256 KV budget, the method retains 94.33% of the dLLM-Cache baseline's performance.	×	0.09
At a 32K-token context, the method achieves 31$\times$ faster decoding and 65% lower peak memory than LLaDA.	×	0.05
The method supports up to 8$\times$ longer prompts on an RTX 4090 GPU.	×	0.01
The MaskKV method achieves a performance of 36.27, which is +0.88 higher than the MaskVoting baseline.	×	0.06

References

- http://arxiv.org/abs/2605.09649v1
- http://arxiv.org/abs/2510.09309v1
- http://arxiv.org/abs/2505.15347v2