FlowKV and Gradient Checkpointing Trade-offs in Gemma-3-12B for High-Throughput Dialogue Systems

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What are the trade-offs between FlowKV and gradient checkpointing in terms of memory efficiency and response latency when deployed on Gemma-3-12B for high-throughput, domain-specific dialogue systems. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: FlowKV: A Disaggregated Inference Framework with Low-Latency KV Cache Transfer and Load-Aware Scheduling. Research question: What are the trade-offs between FlowKV and gradient checkpointing in terms of memory efficiency and response latency when deployed on Gemma-3-12B for high-throughput, domain-specific dialogue systems (e.g., medical or legal Q&A)?.

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 3.2/10.

3 Results

14 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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
The study compares FlowKV against PD-colocate inference frameworks and existing open-source disaggregated inference syst	×	0.11
Simulated data was generated with predefined input and output lengths to compare maximum throughput.	×	0.03
Real-world data for end-to-end response latency comparison was sampled from the gov_report, multi_news, and qmsum subtas	×	0.04
Requests were simulated using a Poisson arrival process with rates controlled via requests per second (RPS).	×	0.02
The experiments utilized Meta-Llama-3.1-8B-Instruct and Meta-Llama-3.1-70B-Instruct models.	×	0.05
vLLM uses PagedAttention to manage memory for attention keys and values combined with continuous batching.	×	0.01
Homogeneous performance evaluation was conducted on an NVIDIA A100-SXM4-80GB server with 8 GPUs interconnected via NVLin	×	0.02
At 1K/256 input/output tokens and 2.0 RPS, FlowKV achieved a throughput of 507.36, while DistServe achieved 404.55.	×	0.02
At 5K/256 input/output tokens and 2.0 RPS, FlowKV achieved a throughput of 470.68, while DistServe achieved 112.87.	×	0.02
At 10K/256 input/output tokens and 2.0 RPS, DistServe resulted in a Failure, whereas FlowKV achieved a throughput of 285	×	0.02
At 1K/256 input/output tokens and 0.1 RPS, FlowKV achieved a throughput of 27.88, compared to 27.62 for DistServe.	×	0.03
Mooncake and vLLM-Disagg showed identical throughput values of 83.74 at 1K/256 tokens and 0.4 RPS.	×	0.01

References

- http://arxiv.org/abs/2411.19147v3

- http://arxiv.org/abs/2504.03775v1
- http://arxiv.org/abs/2005.10855v1