FlowKV and SnapKV Performance on Needle-in-a-Haystack Under Sub-2GB Memory Constraints

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the comparative performance of FlowKV versus SnapKV on the Needle-in-a-Haystack retrieval task for Llama-3-8b when constrained to extreme memory budgets below 2GB. 13 claims were extracted from source literature; 5 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.6/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Retrieval Models Aren't Tool-Savvy: Benchmarking Tool Retrieval for Large Language Models. Research question: What is the comparative performance of FlowKV versus SnapKV on the Needle-in-a-Haystack retrieval task for Llama-3-8b when constrained to extreme memory budgets below 2GB?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.6/10.

3 Results

15 papers retrieved. 13 claims extracted; 5 independently verified. Quality review score: 5.6/10.

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
IR models exhibit poor performance on TOOLRET, with low retrieval quality degrading the task pass rate of tool-use LLMs.	✓	0.35
A large-scale training dataset with over 200k instances substantially optimizes the tool retrieval ability of IR models.	✓	0.36
Strong retrievers like colbertv2 struggle to retrieve target tools effectively.	×	0.03
TOOLRET is the first large-scale tool retrieval benchmark comprising 7.6k diverse retrieval tasks and a corpus of 43k to	✓	0.34
The collected data for TOOLRET is curated to cover a wide range of practical tool requirements, comprising diverse types	×	0.09
The best model (NV-embedd-v1) achieves an nDCG@10 of only 33.83 in the TOOLRET benchmark.	×	0.02
Two key factors contributing to the performance gap in tool retrieval tasks are lower term overlap between queries and t	✓	0.15
TOOLRET-train is a large-scale training dataset containing more than 200k retrieval tasks.	✓	0.15
The pass rate with pre-annotated toolset (oracle) decreases by 10.1 for e5-large-v2, bge-base-v1.5, and bge-large-v1.5.	×	0.05
The Rouge-L score between generated instruction and seed instruction follows a normal distribution with a kernel density	×	0.00
BM25 achieves a score of 18.98, while COLT achieves 15.43, Colbert achieves 22.40, contriever-msmarco achieves 21.15, gt	×	0.01
NV-Embed-v1 achieves a score of 60, while bm25 achieves 55, e5-small-v2 achieves 55, gte-large-en-v1.5 achieves 50, e5-b	×	0.02
COLT achieves a score of 28.91, Colbert achieves 16.67, and contriever-msmarco achieves 23.48 in the benchmark.	×	0.01

References

- http://arxiv.org/abs/2107.12246v2
- http://arxiv.org/abs/2503.01763v2
- http://arxiv.org/abs/2208.13771v2