DeepSeek R1 and Codestral Memory and Latency with IceCache vs. On-GPU KV-Cache

Assignee Research

June 1, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: How does the memory consumption and latency of DeepSeek R1 and Codestral compare when using IceCache's KV-cache management against traditional on-GPU KV-cache in autoregressive generation tasks with. Key-Value (KV) cache plays a crucial role in accelerating inference in large language models (LLMs) by storing intermediate attention states and avoiding redundant computation during autoregressive generation. However, its memory footprint scales linearly with sequence length. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: IceCache: Memory-efficient KV-cache Management for Long-Sequence LLMs. Research question: How does the memory consumption and latency of DeepSeek R1 and Codestral compare when using IceCache's KV-cache management against traditional on-GPU KV-cache in autoregressive generation tasks with sequence lengths up to 16k tokens?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

9 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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
IceCache was evaluated on Llama-3.1-8B-Instruct and Mistral-7B-Instruct-v0.2, which employ group-query attention (GQA).	×	0.09
IceCache was also evaluated on Qwen3-32B, a larger-scale model, and LongChat-7B-v1.5, which employs standard multi-head	×	0.03
Experiments were conducted on LongBench, which contains long generation tasks such as summarization and code generation.	×	0.09
Experiments were conducted on GSM8K, which requires chain-of-thought reasoning.	×	0.07
Experiments were conducted on RULER under extremely long-context settings.	×	0.06
The experimental platform comprised an NVIDIA A100 40GB PCIe GPU for small models and an NVIDIA H100 80GB PCIe GPU for l	×	0.04
The software stack included CUDA version 12.2, PyTorch version 2.4.1, and HuggingFace Transformers version 4.57.1.	×	0.02
IceCache was implemented on top of HuggingFace Transformers, utilizing FlashInfer for the attention kernel operation.	×	0.02
Neither IceCache nor baseline methods were applied to the first two layers of the models.	×	0.03
IceCache maintained 100% retrieval accuracy across all tested budget sizes in the passkey retrieval task.	×	0.04
IceCache was compared against six state-of-the-art KV cache optimization methods on the LongBench benchmark.	×	0.09

References

- http://arxiv.org/abs/2601.04359v1
- http://arxiv.org/abs/2604.10539v1
- http://arxiv.org/abs/2511.00321v1