Impact of 4-bit Quantization on Llama-3 LongBench Accuracy Under Strict Edge Memory Constraints

Assignee Research

June 11, 2026

Abstract

Context lengths of Large Language Models (LLMs) have exploded in recent years, with 128k-token context becoming a standard and million-token context becoming a reality. Efficiently supporting long-context inference remains challenging as the memory that must be allocated in key-value (KV) cache for a generation scales with its context length, limiting the number of long-context requests that can be served concurrently under a given memory budget. KV cache compression can mitigate this issue by removing under-utilized KVs from each attention head's cache and reducing its memory footprint. Highe

1 Introduction

This paper examines: KV-Compress: Paged KV-Cache Compression with Variable Compression Rates per Attention Head. Research question: How does 4-bit quantization impact LongBench accuracy for Llama-3 compared to full-precision baselines when evaluated under strict edge memory constraints?.

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 8.3/10.

3 Results

14 papers retrieved. 7 claims extracted; 7 independently verified. Quality review score: 8.3/10.

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
Context lengths of Large Language Models (LLMs) have exploded in recent years, with 128k-token context becoming a standa	✓	0.32
Efficiently supporting long-context inference remains challenging as the memory that must be allocated in key-value (KV)	✓	0.43
KV cache compression can mitigate this issue by removing under-utilized KVs from each attention head's cache and reducin	✓	0.39
Higher theoretical compression rates can be achieved when the number of removed KVs varies across attention heads, but a	✓	0.45
We introduce KV-Compress, a novel compression method that evicts contiguous KV blocks within a PagedAttention framework,	✓	0.44
Our method achieves state-of-the-art performance on LongBench for both Mistral-7B-Instruct-v0.2 and Llama-3.1-8B-Instruc	✓	0.42
Evaluations on Llama-3.1-8B-Instruct and Llama-3.1-70B-Instruct-FP8 achieve compression rates up to 8x with negligible i	✓	0.40

References

- https://doi.org/10.48550/arxiv.2312.15234
- https://doi.org/10.48550/arxiv.2410.00161
- https://doi.org/10.3390/fi17090417