Quantization Impact on DeepSeek R1 Throughput and CVE Classification Accuracy

Assignee Research

June 4, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does the quantization of Deepseek R1 impact its throughput and false positive rate when classifying CVEs in the Big-Vul benchmark. 16 claims were extracted from source literature; 5 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Quantitative Analysis of Performance Drop in DeepSeek Model Quantization. Research question: How does the quantization of Deepseek R1 impact its throughput and false positive rate when classifying CVEs in the Big-Vul benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.8/10.

3 Results

13 papers retrieved. 16 claims extracted; 5 independently verified. Quality review score: 5.8/10.

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
DeepSeek-R1 achieves an accuracy of 79.8% on AIME 2024 in FP8 (Reported).	×	0.08
DeepSeek-R1 achieves an accuracy of 97.3% on MATH 500 in FP8 (Reported).	×	0.09
DeepSeek-R1 achieves an accuracy of 71.5% on GPQA in FP8 (Reported).	×	0.07
DeepSeek-R1 achieves an accuracy of 90.8% on MMLU in FP8 (Reported).	×	0.07
DeepSeek-R1 achieves an accuracy of 91.8% on C-Eval in FP8 (Reported).	×	0.07
The average accuracy of DeepSeek-R1 across all benchmarks in FP8 (Official API) is 83.48%.	×	0.05
The weighted average accuracy of DeepSeek-R1 across all benchmarks in FP8 (Official API) is 85.82%.	×	0.05
The accuracy drop for DeepSeek-R1 in Q4 K M (llama.cpp) compared to FP8 (Official API) is 0.68%.	×	0.06
The accuracy drop for DeepSeek-R1 in Q3 K M (llama.cpp) compared to FP8 (Official API) is 1.80%.	×	0.06
The accuracy drop for DeepSeek-R1 in UD-Q2 K XL (Unsloth) compared to FP8 (Official API) is 0.94%.	×	0.05
The accuracy drop for DeepSeek-R1 in DQ3 K M (Ours) compared to FP8 (Official API) is 0.34%.	×	0.08
4-bit quantization maintains little performance degradation versus FP8 while enabling single-machine deployment on stand	✓	0.40
DQ3 K M is a dynamic 3-bit quantization method that significantly outperforms traditional Q3 K M variant on various benc	✓	0.24
DQ3 K M is comparable with 4-bit quantization (Q4 K M) approach in most tasks.	✓	0.16
DQ3 K M supports single-machine deployment configurations for both NVIDIA H100/A100 and Huawei 910B.	✓	0.30
The implementation of DQ3 K M is released at https://github.com/UnicomAI/DeepSeek-Eval.	✓	0.21

References

- http://arxiv.org/abs/2304.11130v1
- http://arxiv.org/abs/2505.02390v2
- http://arxiv.org/abs/2503.10486v2