SOVEREIGN: How does the performance of DeepSeek-R1 compare to o1-preview on the APPS benchmark when evaluated under negat

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Recently, there is a high demand for deploying DeepSeek-R1 and V3 locally, possibly because the official service often suffers from being busy and some organizations have data privacy concerns. While single-machine deployment offers infrastructure simplicity, the models' 671B FP8 parameter configuration exceeds the practical memory limits of a standard 8-GPU machine. Quantization is a widely used technique that helps reduce model memory consumption. However, it is unclear what the performance of DeepSeek-R1 and V3 will be after being quantized. This technical report presents the first quantita

1 Introduction

Analysis of: Quantitative Analysis of Performance Drop in DeepSeek Model Quantization. Research goal: How does the performance of DeepSeek-R1 compare to o1-preview on the APPS benchmark when evaluated under negation-based token perturbations, and does the S* selection mechanism mitigate observed robustness gaps?.

2 Methodology

Multi-query arXiv search (1 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

6 papers retrieved. 7 claims extracted, 5 verified. Tribunal: 5.0/10 → REVISE (revision_round=1). Policy: ESCALATE_TO_OWNER.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
DQ3 K M quantization method supports single-machine deployment for both NVIDIA H100/A100 and Huawei 910B.	✓	0.23
The DQ3 K M method supports single-machine deployment configurations for both NVIDIA H100/A100 and Huawei 910BB.	✓	0.21
The accuracy drop for 4-bit quantization (Q4 K M) is 0.68% relative to FP8 (Official API) performance.	×	0.10
The DQ3 K M method was found to have a 0.34% accuracy drop compared to the baseline.	×	0.03
The 4-bit quantization (Q4 K M) maintains little performance degradation versus FP8 while enabling single-machine deploy	✓	0.35
The DQ3 K M method outperforms the traditional Q3 K M variant on various benchmarks, and is comparable with 4-bit quanti	✓	0.21
DQ3 K M supports single-machine deployment configurations for both NVIDIA H100/A100 and Huawei 910B.	✓	0.24

References

- http://arxiv.org/abs/2504.07128v3
- http://arxiv.org/abs/2503.10460v4
- http://arxiv.org/abs/2505.02390v2