Comparative Efficiency of Inference Optimization Techniques on HLE-Verified Benchmarks

Assignee Research

June 3, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the comparative efficiency of different inference optimization techniques when evaluating frontier models on the revised HLE-Verified benchmark in terms of throughput and accuracy trade-offs. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: UniComp: A Unified Evaluation of Large Language Model Compression via Pruning, Quantization and Distillation. Research question: What is the comparative efficiency of different inference optimization techniques when evaluating frontier models on the revised HLE-Verified benchmark in terms of throughput and accuracy trade-offs?.

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 6.5/10.

3 Results

15 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.5/10.

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.

References

- http://arxiv.org/abs/2602.09130v5
- http://arxiv.org/abs/2603.21389v1
- http://arxiv.org/abs/2602.13964v3