What is the performance gap in multi-step arithmetic reasoning between quantized 4-bit LLaMA-2 models and thei

Assignee Research

June 10, 2026

Abstract

In this work, we introduce a novel evaluation paradigm for Large Language Models (LLMs) that compels them to transition from a traditional question-answering role, akin to a student, to a solution-scoring role, akin to a teacher. This paradigm, focusing on "reasoning about reasoning," hence termed meta-reasoning, shifts the emphasis from result-oriented assessments, which often neglect the reasoning process, to a more comprehensive evaluation that effectively distinguishes between the cognitive capabilities of different models. By applying this paradigm in the GSM8K dataset, we have developed

1 Introduction

This paper examines: MR-GSM8K: A Meta-Reasoning Benchmark for Large Language Model Evaluation. Research question: What is the performance gap in multi-step arithmetic reasoning between quantized 4-bit LLaMA-2 models and their full-precision counterparts on the GSM8K benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.8/10.

3 Results

11 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.8/10.

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.

References

- http://arxiv.org/abs/2603.13931v1
- http://arxiv.org/abs/2509.25160v1
- http://arxiv.org/abs/2312.17080v4