Ruler Score Discrepancies in Llama-3.1-8B Benchmark Evaluations Across Studies

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 1 peer-reviewed paper addressing the following research question: Benchmark archaeology: investigate Ruler score discrepancy for Llama-3.1-8B — reported 1.9%–85.6% (spread 83.7pp) across 2 papers. Sources: 'MTraining: Distributed Dynamic Sparse At' (1.9%);. 8 claims were extracted from source literature; 8 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 9.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: VISTA: Verification In Sequential Turn-based Assessment. Research question: Benchmark archaeology: investigate Ruler score discrepancy for Llama-3.1-8B — reported 1.9%–85.6% (spread 83.7pp) across 2 papers. Sources: 'MTraining: Distributed Dynamic Sparse At' (1.9%); 'ReST-KV: Robust KV Cache Eviction with L' (85.6%). Identify evaluation protocol differences (few-shot, prompting, preprocessing)..

2 Methodology

Systematic literature search across multiple databases yielded 1 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 9.0/10.

3 Results

1 papers retrieved. 8 claims extracted; 8 independently verified. Quality review score: 9.0/10.

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
Hallucination is defined as generating statements unsupported or contradicted by available evidence or conversational co	✓	0.30
Existing metrics either evaluate isolated responses or treat unverifiable content as errors, limiting their use for mult	✓	0.34
VISTA is a framework for evaluating conversational factuality through claim-level verification and sequential consistenc	✓	0.32
VISTA decomposes each assistant turn into atomic factual claims, verifies them against trusted sources and dialogue hist	✓	0.42
VISTA substantially improves hallucination detection over FACTSCORE and LLM-as-Judge baselines across eight large langua	✓	0.38
Human evaluation confirms that VISTA's decomposition improves annotator agreement and reveals inconsistencies in existin	✓	0.31
VISTA models factuality as a dynamic property of conversation.	✓	0.17
VISTA offers a more transparent, human-aligned measure of truthfulness in dialogue systems.	✓	0.28

References

- https://doi.org/10.48550/arxiv.2510.27052