Adaptive Reasoning Suppression and Speculative Decoding Performance on GSM8K with Llama-3-8B

Assignee Research

June 5, 2026

Abstract

This report synthesises findings from 5 peer-reviewed papers addressing the following research question: How does Adaptive Reasoning Suppression compare to Speculative Decoding in terms of GSM8K accuracy and throughput for Llama-3-8B models. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: ARS: Adaptive Reasoning Suppression for Efficient Large Reasoning Language Models. Research question: How does Adaptive Reasoning Suppression compare to Speculative Decoding in terms of GSM8K accuracy and throughput for Llama-3-8B models?.

2 Methodology

Systematic literature search across multiple databases yielded 5 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.4/10.

3 Results

5 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.4/10.

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
ARS consistently achieves superior length reduction while maintaining competitive accuracy across all model scales.	×	0.11
ARS operates through three core components: (1) Multi-checkpoint certainty estimation, (2) Progressive threshold adaptat	×	0.15
ARS establishes multiple checkpoints {c1, c2, . . . , ck} at regular intervals during generation.	×	0.02
At each checkpoint ci, ARS estimates model certainty through tentative answer probing.	×	0.05
The heuristic difficulty estimation is used in the ARS algorithm.	×	0.02
The maximum token limit per response in ARS is set to 1200 tokens.	×	0.02
ARS aims to minimize the expected output length E[T] while preserving reasoning accuracy.	×	0.08
ARS uses three different policies: CoDFastPolicy, ElasticModeratePolicy, and DeepReflectPolicy.	×	0.01
ARS uses difficulty thresholds d1 and d2, and confidence thresholds c1, c2, and c3.	×	0.03
ARS uses a trigger set T = {"Wait", "But", "Alternatively", . . .} to identify reflection behaviors.	×	0.01

References

- http://arxiv.org/abs/2510.00071v2
- http://arxiv.org/abs/2508.17739v2
- http://arxiv.org/abs/2312.17080v4