Dynamic Suppression of Redundant Reasoning in ARS vs. Static Pruning for Inference Throughput

Assignee Research

June 5, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does the dynamic suppression of redundant reasoning steps in ARS compare to static pruning methods in terms of inference throughput on GSM8K and MATH benchmarks. 14 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. 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 the dynamic suppression of redundant reasoning steps in ARS compare to static pruning methods in terms of inference throughput on GSM8K and MATH benchmarks?.

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 3.5/10.

3 Results

11 papers retrieved. 14 claims extracted; 1 independently verified. Quality review score: 3.5/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
The study evaluates Qwen2.5-Math-1.5B-Instruct, Qwen2.5-Math-7B-Instruct, and DeepSeek-R1-Distill-Qwen-7B models.	×	0.06
The ARS algorithm utilizes difficulty thresholds d1 and d2 to determine the scheduling mode.	×	0.03
In FAST mode, the ARS policy is configured with 2 drafts and 10 tokens per draft.	×	0.01
In MOD mode, the ARS policy uses a budget of 64 tokens.	×	0.01
In the default mode (neither FAST nor MOD), the ARS policy uses a sc_k value of 3.	×	0.01
The generation process in the ARS algorithm terminates if the text length reaches 1200 tokens.	×	0.03
Confidence scores in ARS are computed using entropy confidence on tentative answers.	×	0.02
Token suppression occurs if the next token is in the trigger set and the suppression probability is greater than a rando	×	0.07
The standard reasoning process defines reflection behaviors as being triggered by keywords such as 'Wait', 'But', and 'A	×	0.05
The objective of the ARS framework is to minimize expected output length E[T] while keeping accuracy degradation below a	×	0.03
ARS operates through three core components: Multi-checkpoint certainty estimation, Progressive threshold adaptation, and	✓	0.15
ARS establishes multiple checkpoints at regular intervals during generation, unlike previous methods that rely on single	×	0.05
On the GSM8K dataset, ARS achieves superior length reduction while maintaining competitive accuracy across all tested mo	×	0.12
Performance metrics for the GSM8K dataset include Accuracy (Acc), Latency (Lat), Tokens Per Correct answer (TPC), and Jo	×	0.04

References

- `http://arxiv.org/abs/2103.05861v1`

- http://arxiv.org/abs/2510.00071v2
- http://arxiv.org/abs/2509.25160v1