Per-Token Compute Density and Error Rates in BigBench Hard Logical Deduction Tasks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: What is the correlation between per-token compute density and error rates on the BigBench Hard logical deduction tasks when using dynamic compute allocation. 7 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: ASTER: Adaptive Spatio-Temporal Early Decision Model for Dynamic Resource Allocation. Research question: What is the correlation between per-token compute density and error rates on the BigBench Hard logical deduction tasks when using dynamic compute allocation?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

10 papers retrieved. 7 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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
ASTER consistently outperforms all baseline methods across diverse datasets and evaluation metrics.	×	0.03
ASTER achieves an average 11.15% improvement in success rate over the second-best approaches.	×	0.04
ASTER demonstrates comprehensive superiority in downstream utility metrics such as RUR and CER.	×	0.03
The optimal effect is achieved when the number of attention heads is set to 4.	×	0.02
Increasing the number of attention heads beyond 4 leads to a decline in decision performance.	×	0.02
ASTER outperforms STAEformer in terms of total reward in the NYPD dataset case study.	×	0.01
ASTER outperforms STAEformer in terms of total reward in the EMS dataset case study.	×	0.01

References

- http://arxiv.org/abs/2407.11310v2
- http://arxiv.org/abs/2601.01982v1
- http://arxiv.org/abs/2506.17929v1