AdaptToken-Lite-8B Benchmark Performance Across Reasoning Mathematics Coding and Language Tasks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What are the benchmark performance scores of AdaptToken-Lite-8B on reasoning mathematics coding and language understanding tasks. 11 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: Probing Vision-Language Understanding through the Visual Entailment Task: promises and pitfalls. Research question: What are the benchmark performance scores of AdaptToken-Lite-8B on reasoning mathematics coding and language understanding tasks.

2 Methodology

Systematic literature search across multiple databases yielded 16 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

16 papers retrieved. 11 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
The OFA model achieves state-of-the-art performance for the VE task on the SNLI-VE dataset with an accuracy of 91.2% on	×	0.14
OFA-X achieves state-of-the-art performance for the VE task on the e-SNLI-VE dataset with an accuracy of 80.9% on the te	×	0.12
The CLOSE model achieves similar performance to the image model on the SNLI-VE dataset without using images.	×	0.07
The Llama 3.2 Vision 11B model achieved 75.2% accuracy on the VQAv2 benchmark.	×	0.06
The Llama 3.2 Vision 11B model achieved 91.1% accuracy on the AI2 Diagram benchmark.	×	0.06
The Llama 3.2 Vision 11B model achieved 51.5% accuracy on the MathVista (testmini) benchmark.	×	0.05
The e-SNLI-VE dataset contains 29,783 images in the training split, 1,000 images in the development split, and 1,000 ima	×	0.03
The e-SNLI-VE dataset contains 131,023 entailment labels in the training split, 5,254 in the development split, and 5,21	×	0.05
The e-SNLI-VE dataset contains 125,902 neutral labels in the training split, 3,442 in the development split, and 3,801 i	×	0.03
The e-SNLI-VE dataset contains 144,792 contradiction labels in the training split, 5,643 in the development split, and 5	×	0.03
The e-SNLI-VE dataset contains 401,717 total labels in the training split, 14,339 in the development split, and 14,740 i	×	0.03

References

- http://arxiv.org/abs/2507.17467v1
- http://arxiv.org/abs/2504.14693v2
- http://arxiv.org/abs/2503.20786v1