AdaptToken-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-8B on reasoning mathematics coding and language understanding tasks. 15 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: Adversarial GLUE: A Multi-Task Benchmark for Robustness Evaluation of Language Models. Research question: What are the benchmark performance scores of AdaptToken-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. 15 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 data curation phase serves as a comprehensive benchmark over existing adversarial attack methods.	×	0.15
The data curation phase provides a fair standard for all adversarial attacks and systematic human annotations to evaluat	×	0.09
Table 4 reports model performance on the AdvGLUE test set for various models including BERT (Large) and RoBERTa (Large).	×	0.05
For MNLI, the test accuracy on the matched and mismatched test sets is reported.	×	0.03
For QQP, accuracy and F1 are reported.	×	0.02
For other tasks, accuracy is reported.	×	0.04
Table 9 reports model performance on the AdvGLUE test set and GLUE dev set for various models.	×	0.06
The macro-average (Avg) of per-task scores for different models is reported.	×	0.05
The first four typo strategies guarantee the word edit distance between the typo word and its original word to be 1, and	×	0.02
In Strategy (i), a space is inserted into a word only when the word contains less than 6 characters.	×	0.03
In Strategy (v), characters in a word are swapped only when the word has more than 4 characters.	×	0.03
For sentiment analysis tasks, the cosine similarity threshold is set to be 0.8.	×	0.02
Table (p4) lists various adversarial attack methods including Embedding, TextBugger, TextFooler, BERT-ATTACK, Sememe-PSO	×	0.07
Table (p17) reports the average performance of various models on SST-2, MNLI, RTE, QNLI, and QQP tasks for both GLUE and	×	0.07
Table (p20) reports the average performance of various word-level attacks on SST-2 and MNLI tasks, including metrics suc	×	0.04

References

- http://arxiv.org/abs/2111.02840v2
- http://arxiv.org/abs/2509.25160v1
- http://arxiv.org/abs/2009.13570v2