GLM-4.5-Air AutoMonitor-Bench Miss Rates in Mathematics Reasoning Tasks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does the AutoMonitor-Bench Miss Rate (MR) of GLM-4.5-Air compare to other state-of-the-art LLMs like GPT-4 and Claude 3 on mathematics reasoning 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 4.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: PediatricsGPT: Large Language Models as Chinese Medical Assistants for Pediatric Applications. Research question: How does the AutoMonitor-Bench Miss Rate (MR) of GLM-4.5-Air compare to other state-of-the-art LLMs like GPT-4 and Claude 3 on mathematics reasoning tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.0/10.

3 Results

4 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 4.0/10.

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
PediatricsGPT is developed upon the Baichuan2-Base models in two versions with 7 and 13 billion parameters.	×	0.03
The model training is accomplished through the PyTorch platform with Accelerate and DeepSpeed packages using eight Nvidi	×	0.03
The ZeRO strategy is employed to alleviate the memory overhead during full parameter training.	×	0.04
The AdamW optimizer is adopted for network optimization, and the bf16 data accuracy is chosen.	×	0.02
Extensive experiments are conducted on three application-oriented benchmarks to assess the model's pediatric medical abi	×	0.05
Each benchmark contains 300 held-out samples to reject data leakage during training.	×	0.03
PedCorpus is constructed through the multi-dimensional corpus across three application-oriented medical tasks, including	×	0.05
Specialized Pediatric Data is extracted from textbooks, guidelines, and knowledge graphs ensuring knowledge professional	×	0.09
Over 500 corresponding disease guidelines are collected, including diagnostic protocols and treatment consensus.	×	0.03
Extensive knowledge entities are sampled from ternary instances in the knowledge graphs.	×	0.03
Real Doctor-patient Conversations are used to avoid the model collapse dilemma.	×	0.01

References

- http://arxiv.org/abs/2402.11651v2
- http://arxiv.org/abs/2405.19266v4
- http://arxiv.org/abs/2310.00034v2