GPT-2-340M Benchmark Performance Across Reasoning Mathematics 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 GPT-2-340M on reasoning mathematics coding and language understanding tasks. 10 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: EchoMind: An Interrelated Multi-level Benchmark for Evaluating Empathetic Speech Language Models. Research question: What are the benchmark performance scores of GPT-2-340M 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 4.2/10.

3 Results

16 papers retrieved. 10 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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
EchoMind is an interrelated multi-level benchmark for evaluating empathetic speech language models.	✓	0.31
EchoMind evaluates models on multiple dimensions including understanding, reasoning, conversation, content, and voice.	×	0.07
EchoMind supports both text and audio inputs and outputs.	×	0.02
EchoMind includes a multi-level evaluation framework.	×	0.08
EchoMind evaluates reasoning, conversation, content, and voice aspects.	×	0.05
EchoMind includes metrics such as WER, SemSim, Acc, NISQA, DNMOS, EmoAlign, and VES for evaluation.	×	0.02
EchoMind evaluates models on text context fit, text conversation naturalness, text colloquial degree, and text speech re	×	0.04
EchoMind includes human evaluation metrics for comparison with model performance.	×	0.05
EchoMind evaluates context fit, speech relevance, and voice emotion similarity.	×	0.04
EchoMind includes a comparison of model performance with human performance.	×	0.07

References

- http://arxiv.org/abs/2510.22758v2
- http://arxiv.org/abs/2410.12381v3
- http://arxiv.org/abs/2604.14140v1