

MedGemma-27B Benchmark Performance Across Reasoning Mathematics and Language Tasks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What are the benchmark performance scores of MedGemma-27B on reasoning mathematics coding and language understanding tasks. 18 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Reactor Mk.1 performances: MMLU, HumanEval and BBH test results. Research question: What are the benchmark performance scores of MedGemma-27B 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 4.3/10.

3 Results

16 papers retrieved. 18 claims extracted; 3 independently verified. Quality review score: 4.3/10.

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
Reactor Mk.1 has less than 100B parameters in total.	×	0.13
Reactor Mk.1 achieved a 92% score on the MMLU evaluation.	✓	0.18
Reactor Mk.1 achieved a 95% score on the HumanEval evaluation.	✓	0.15
Reactor Mk.1 achieved an 88% score on the BBH evaluation.	✓	0.18
GPT-4o supports real-time conversations, Q&A, and text generation.	×	0.05
GPT-4o can process and respond to combinations of text, audio, and image files.	×	0.05
Claude Opus can analyze static images, handwritten notes, and graphs.	×	0.04
Claude Opus can generate code for websites in HTML and CSS.	×	0.04
Claude Opus can turn images into structured JSON data.	×	0.04
Claude Opus can debug complex code bases.	×	0.04
Claude Opus can translate between various languages in real-time.	×	0.05
Meta Llama 3 has two models featuring 8 billion and 70 billion parameters.	×	0.09
Meta Llama 3 uses a decoder-only transformer architecture.	×	0.03
Meta Llama 3 features a tokenizer with a 128K vocabulary.	×	0.02
Meta Llama 3 uses grouped query attention across 8B and 70B sizes.	×	0.02
Meta Llama 3 models are trained on sequences of 8,192 tokens.	×	0.03
Gemini models are natively multimodal and capable of understanding and combining text, code, images, audio, and video fi	×	0.04
Gemini models undergo extensive safety testing, including bias assessments.	×	0.03

References

- <http://arxiv.org/abs/2509.25160v1>

- <http://arxiv.org/abs/2406.10515v2>
- <http://arxiv.org/abs/2312.17080v4>