Gemma3 Benchmark Performance Across Reasoning Mathematics Coding and Language Tasks

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: What are the benchmark performance scores of Gemma3 on reasoning mathematics coding and language understanding tasks. 10 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Can Small and Reasoning Large Language Models Score Journal Articles for Research Quality and Do Averaging and Few-shot Help?. Research question: What are the benchmark performance scores of Gemma3 on reasoning mathematics coding and language understanding tasks.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

8 papers retrieved. 10 claims extracted; 2 independently verified. Quality review score: 4.5/10.

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
All medium sized LLMs give similar correlations to the cloud-based LLMs and, based on the confidence interval estimates,	×	0.14
The reasoning models Magistral Small, Qwen3 32b, DeepSeek R1 32b perform well but do not have a clear advantage over the	✓	0.19
The overall pattern is similar within UoAs 1 to 6.	×	0.04
Choosing a smaller LLM in a family tends to weaken the correlations but not substantially, except for Gemma 3 1b.	×	0.04
The minimum practical size for an LLM may be between 4b and 1b, although the cutoff may be different for LLM families ot	×	0.06
The patterns within UoAs 1 to 6 are broadly consistent with the overall pattern, except that within UoAs 2 and 5, low co	×	0.03
The research design was to create a large dataset of articles from multiple fields with expert scores, then apply the di	×	0.10
Previous research has shown that LLM accuracy is low because LLM scores tend to cluster at a particular value, so the mo	×	0.10
Previous studies using research quality for journal articles have either used small datasets of under 100 articles with	✓	0.17
The current study initially followed the commonly used departmental mean score approach above, although the largest REF2	×	0.05

References

- http://arxiv.org/abs/2602.02842v1
- http://arxiv.org/abs/2510.24514v1
- http://arxiv.org/abs/2510.22389v2