# DeepSeek-14B Benchmark Performance Across Reasoning Mathematics and Language Tasks

Assignee Research

June 6, 2026

**Abstract**

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What are the benchmark performance scores of DeepSeek-14B on reasoning mathematics coding and language understanding tasks. 20 claims were extracted from source literature; 2 were 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: Reactor Mk.1 performances: MMLU, HumanEval and BBH test results. Research question: What are the benchmark performance scores of DeepSeek-14B on reasoning mathematics coding and language understanding tasks.

## 2 Methodology

Systematic literature search across multiple databases yielded 15 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

15 papers retrieved. 20 claims extracted; 2 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 |
|---|---|---|
| Reactor Mk.1 has less than 100B parameters in total. | × | 0.12 |
| Reactor Mk.1 achieved a 92% score on the MMLU evaluation. | ✓ | 0.16 |
| Reactor Mk.1 achieved a 95% score on the HumanEval evaluation. | × | 0.15 |
| Reactor Mk.1 achieved an 88% score on the BBH evaluation. | ✓ | 0.16 |
| Reactor Mk.1 exhibited superior performance compared to benchmark models like GPT 4o, Claude, Llama 3, Gemini, and Mistr | × | 0.13 |
| GPT-4o supports real-time conversations, Q&A, and text generation, utilizing all modalities in a single model to underst | × | 0.03 |
| GPT-4o can engage in real-time verbal conversations with minimal delay, respond to questions using its knowledge base, a | × | 0.03 |
| GPT-4o processes and responds to combinations of text, audio, and image files. | × | 0.04 |
| Claude Opus can analyze static images, handwritten notes, and graphs. | × | 0.04 |
| Claude Opus can generate code for websites in HTML and CSS. | × | 0.05 |
| Claude Opus can turn images into structured JSON data and debug complex code bases. | × | 0.03 |
| Claude Opus can translate between various languages in real-time, practice grammar, and create multilingual content. | × | 0.04 |
| Meta Llama 3 has two models featuring 8 billion and 70 billion parameters. | × | 0.09 |
| Meta Llama 3 demonstrates state-of-the-art performance on industry benchmarks and offers improved reasoning and code gen | × | 0.06 |
| Meta Llama 3 uses a decoder-only transformer architecture, featuring a tokenizer with a 128K vocabulary and grouped quer | × | 0.01 |
| Meta Llama 3 models are trained on sequences of 8,192 tokens to ensure efficient language encoding and inference. | × | 0.03 |
| Gemini models are natively multimodal and capable of understanding and combining text, code, images, audio, and video fi | × | 0.08 |
| Gemini models can generate code based on various inputs and perform complex reasoning tasks. | × | 0.09 |
| Gemini Pro and Ultra versions outperform previous models in terms of pretraining and post-training improvements. | × | 0.02 |
| Gemini models undergo extensive safety testing, including bias assessments, in collaboration with external experts to id | × | 0.03 |

## References


- http://arxiv.org/abs/2406.10515v2
- http://arxiv.org/abs/2210.09261v1
- http://arxiv.org/abs/2502.19187v2