SOVEREIGN: How does the accuracy of Gemini 1.5 Pro on the MMMU benchmark compare to MoE-LLaVA and dense LLaVA-1.5 when ev

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Excellence in a wide variety of medical applications poses considerable challenges for AI, requiring advanced reasoning, access to up-to-date medical knowledge and understanding of complex multimodal data. Gemini models, with strong general capabilities in multimodal and long-context reasoning, offer exciting possibilities in medicine. Building on these core strengths of Gemini, we introduce Med-Gemini, a family of highly capable multimodal models that are specialized in medicine with the ability to seamlessly use web search, and that can be efficiently tailored to novel modalities using custo

1 Introduction

Analysis of: Capabilities of Gemini Models in Medicine. Research goal: How does the accuracy of Gemini 1.5 Pro on the MMMU benchmark compare to MoE-LLaVA and dense LLaVA-1.5 when evaluated under a fixed 4K token context window?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

11 papers retrieved. 10 claims extracted, 9 verified. Tribunal: 7.9/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
Med-Gemini was evaluated on 14 medical benchmarks.	×	0.14
Med-Gemini established new state-of-the-art (SoTA) performance on 10 out of 14 medical benchmarks.	✓	0.23
Med-Gemini surpasses the GPT-4 model family on every benchmark where a direct comparison is viable.	✓	0.22
The best-performing Med-Gemini model achieved 91.1% accuracy on the MedQA (USMLE) benchmark.	✓	0.21
The MedQA (USMLE) performance of 91.1% was achieved using a novel uncertainty-guided search strategy.	✓	0.22
Med-Gemini was evaluated on 7 multimodal benchmarks, including NEJM Image Challenges and MMMU (health & medicine).	✓	0.24
On 7 multimodal benchmarks, Med-Gemini improved over GPT-4V by an average relative margin of 44.5%.	✓	0.21
Med-Gemini achieved state-of-the-art performance on a needle-in-a-haystack retrieval task from long de-identified health	✓	0.25
Med-Gemini achieved state-of-the-art performance on a medical video question answering task.	✓	0.16
Med-Gemini surpassed prior bespoke methods on needle-in-a-haystack retrieval and medical video QA using only in-context	✓	0.21

References

- https://doi.org/10.48550/arxiv.2404.14219

- `https://doi.org/10.48550/arxiv.2404.18930`
- `https://doi.org/10.48550/arxiv.2404.18416`