SOVEREIGN: MMLU benchmark results multiple language models comparison GPT-4 Claude Gemini scores accuracy 2024

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Abstract Large language models (LLMs) have demonstrated impressive capabilities, but the bar for clinical applications is high. Attempts to assess the clinical knowledge of models typically rely on automated evaluations based on limited benchmarks. Here, to address these limitations, we present MultiMedQA, a benchmark combining six existing medical question answering datasets spanning professional medicine, research and consumer queries and a new dataset of medical questions searched online, HealthSearchQA. We propose a human evaluation framework for model answers along multiple axes including

1 Introduction

Analysis of: Large language models encode clinical knowledge. Research goal: MMLU benchmark results multiple language models comparison GPT-4 Claude Gemini scores accuracy 2024.

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, 10 verified. Tribunal: 9.3/10 $\rightarrow$ 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
Large language models (LLMs) have demonstrated impressive capabilities, but the bar for clinical applications is high.	✓	0.26
Attempts to assess the clinical knowledge of models typically rely on automated evaluations based on limited benchmarks.	✓	0.30
MultiMedQA is a benchmark combining six existing medical question answering datasets spanning professional medicine, res	✓	0.39
A human evaluation framework for model answers along multiple axes including factuality, comprehension, reasoning, possi	✓	0.31
Pathways Language Model 1 (PaLM, a 540-billion parameter LLM) and its instruction-tuned variant, Flan-PaLM 2 are evaluat	✓	0.29
Using a combination of prompting strategies, Flan-PaLM achieves state-of-the-art accuracy on every MultiMedQA multiple-c	✓	0.43
Flan-PaLM achieves 67.6% accuracy on MedQA (US Medical Licensing Exam-style questions), surpassing the prior state of th	✓	0.34
Human evaluation reveals key gaps in the performance of Flan-PaLM.	✓	0.20
Instruction prompt tuning is introduced as a parameter-efficient approach for aligning LLMs to new domains using a few e	✓	0.26
The resulting model, Med-PaLM, performs encouragingly, but remains inferior to clinicians.	✓	0.25

References

- https://doi.org/10.48550/arxiv.2210.11416
- https://doi.org/10.48550/arxiv.2403.05530

- https://doi.org/10.1038/s41586-023-06291-2