SOVEREIGN: State of large language models benchmark evaluation GPT-4 Claude Gemini performance comparison 2024 2025

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Large Language Models (LLMs) have drawn a lot of attention due to their strong performance on a wide range of natural language tasks, since the release of ChatGPT in November 2022. LLMs' ability of general-purpose language understanding and generation is acquired by training billions of model's parameters on massive amounts of text data, as predicted by scaling laws \cite{kaplan2020scaling,hoffmann2022training}. The research area of LLMs, while very recent, is evolving rapidly in many different ways. In this paper, we review some of the most prominent LLMs, including three popular LLM families

1 Introduction

Analysis of: Large Language Models: A Survey. Research goal: State of large language models benchmark evaluation GPT-4 Claude Gemini performance comparison 2024 2025.

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: 9.0/10 $\rightarrow$ 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 drawn a lot of attention due to their strong performance on a wide range of natural la	✓	0.40
LLMs' ability of general-purpose language understanding and generation is acquired by training billions of model's param	✓	0.39
The research area of LLMs, while very recent, is evolving rapidly in many different ways.	✓	0.27
This paper reviews some of the most prominent LLMs, including three popular LLM families (GPT, LLaMA, PaLM).	✓	0.28
The paper discusses the characteristics, contributions, and limitations of GPT, LLaMA, and PaLM.	✓	0.17
The paper gives an overview of techniques developed to build and augment LLMs.	✓	0.20
The paper surveys popular datasets prepared for LLM training, fine-tuning, and evaluation.	✓	0.26
The paper reviews widely used LLM evaluation metrics.	✓	0.20
The paper compares the performance of several popular LLMs on a set of representative benchmarks.	✓	0.22
The paper concludes by discussing open challenges and future research directions.	✓	0.20

References

- https://doi.org/10.3390/informatics11030057
- https://doi.org/10.4230/oasics.icpec.2025.4
- https://doi.org/10.48550/arxiv.2402.06196