SOVEREIGN: Cross-benchmark generalization of PRISM framework's robustness to irrelevant context: how do Llama-3, Mistral,

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 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: Cross-benchmark generalization of PRISM framework's robustness to irrelevant context: how do Llama-3, Mistral, and Qwen model backends compare on F1 and precision metrics when evaluated on standardized multi-hop reasoning benchmarks with context windows from 1k to 32k tokens?.

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

7 papers retrieved. 4 claims extracted, 2 verified. Tribunal: 6.7/10 → REVISE (revision_round=1). Policy: ESCALATE_TO_OWNER.

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
ChatGPT was released in November 2022	×	0.09
LLMs are trained with billions of model parameters on massive amounts of text data	✓	0.22
Scaling laws predict LLM performance based on model size and training data	×	0.11
GPT, LLaMA, and PaLM are three popular LLM families	✓	0.21

References

- https://doi.org/10.1038/s41746-024-01390-4
- https://doi.org/10.48550/arxiv.2402.06196
- https://doi.org/10.48550/arxiv.2307.13721