SOVEREIGN: Can Vendi-RAG maintain answer accuracy on the HotpotQA benchmark when the diversity weight hyperparameter is s

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: Can Vendi-RAG maintain answer accuracy on the HotpotQA benchmark when the diversity weight hyperparameter is scaled across a range of 0.1 to 1.0, measured against the Tree of Reviews evaluation metric?.

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

6 papers retrieved. 4 claims extracted, 4 verified. Tribunal: 7.7/10 $\rightarrow$ 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 natural language tasks sin	✓	0.35
LLMs' ability of general-purpose language understanding and generation is acquired by training billions of model's param	✓	0.35
LLMs are trained on massive amounts of text data as predicted by scaling laws	✓	0.20
GPT, LLaMA, and PaLM are three popular LLM families	✓	0.23

References

- https://doi.org/10.48550/arxiv.2402.06196
- https://doi.org/10.1109/access.2021.3140175
- https://doi.org/10.18653/v1/2024.acl-long.642