SOVEREIGN: Large language model leaderboard benchmark comparison scores GPT-4o Claude-3 Gemini-1.5 LLaMA-3 performance 20

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

In this paper, we explore the capabilities of state-of-the-art large language models (LLMs) such as GPT-4, GPT-4o, Claude 3.5 Sonnet, Claude 3 Opus, Gemini 1.5 Pro, Llama 3, and Llama 3.1 in solving some selected undergraduate-level transportation engineering problems. We introduce TransportBench, a benchmark dataset that includes a sample of transportation engineering problems on a wide range of subjects in the context of planning, design, management, and control of transportation systems. This dataset is used by human experts to evaluate the capabilities of various commercial and open-source

1 Introduction

Analysis of: Benchmarking the Capabilities of Large Language Models in Transportation System Engineering: Accuracy, Consistency, and Reasoning Behaviors. Research goal: Large language model leaderboard benchmark comparison scores GPT-4o Claude-3 Gemini-1.5 LLaMA-3 performance 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

4 papers retrieved. 0 claims extracted, 0 verified. Tribunal: 1.0/10 $\rightarrow$ REJECT (revision_round=0). 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.

References

- https://arxiv.org/abs/2507.01955
- https://www.semanticscholar.org/paper/5a100bd03cfc4d08db7fe229b3669ebb5e774541
- https://arxiv.org/abs/2408.08302