SOVEREIGN: How does the performance of EVOR compare to static RAG pipelines on code generation accuracy across multiple p

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4, was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google's PaLM for example)

1 Introduction

Analysis of: Sparks of Artificial General Intelligence: Early experiments with GPT-4. Research goal: How does the performance of EVOR compare to static RAG pipelines on code generation accuracy across multiple programming language domains using the HumanEval benchmark?.

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

8 papers retrieved. 4 claims extracted, 4 verified. Tribunal: 7.5/10 $\rightarrow$ 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
GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, wit	✓	0.32
GPT-4's performance is strikingly close to human-level performance	✓	0.21
GPT-4 often vastly surpasses prior models such as ChatGPT	✓	0.18
GPT-4 was trained using an unprecedented scale of compute and data	✓	0.20

References

- https://doi.org/10.48550/arxiv.2303.12712
- https://doi.org/10.1145/3520312.3534862
- https://doi.org/10.1007/s11704-026-60308-3