SOVEREIGN: How does GPT-4o's performance on HumanEval compare to other state-of-the-art LLMs like Claude 3 Opus or Llama

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Large Language Models (LLMs) have garnered remarkable advancements across diverse code-related tasks, known as Code LLMs, particularly in code generation that generates source code with LLM from natural language descriptions. This burgeoning field has captured significant interest from both academic researchers and industry professionals due to its practical significance in software development, e.g., GitHub Copilot. Despite the active exploration of LLMs for a variety of code tasks, either from the perspective of natural language processing (NLP) or software engineering (SE) or both, there is

1 Introduction

Analysis of: A Survey on Large Language Models for Code Generation. Research goal: How does GPT-4o's performance on HumanEval compare to other state-of-the-art LLMs like Claude 3 Opus or Llama 3 in terms of pass@1 and pass@k metrics?.

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

9 papers retrieved. 6 claims extracted, 6 verified. Tribunal: 8.8/10 $\rightarrow$ 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 achieved advancements in code generation tasks, generating source code from natural la	✓	0.20
GitHub Copilot is an example of a practical application of LLMs for code generation in software development.	✓	0.17
There is a noticeable absence of a comprehensive and up-to-date literature review dedicated specifically to LLMs for cod	✓	0.24
The survey introduces a taxonomy categorizing developments in LLMs for code generation covering data curation, advances,	✓	0.19
The survey presents an empirical comparison of LLM capabilities using the HumanEval, MBPP, and BigCodeBench benchmarks.	✓	0.18
The empirical comparison in the survey covers various levels of difficulty and types of programming tasks.	✓	0.18

References

- https://doi.org/10.48550/arxiv.2406.00515
- https://doi.org/10.48550/arxiv.2406.12793
- https://doi.org/10.48550/arxiv.2404.14219