# GPT-4o HumanEval Performance Discrepancies Across Evaluation Protocols

Assignee Research

May 30, 2026

**Abstract**

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: Benchmark archaeology: investigate HumanEval score discrepancy for GPT-4o — reported 27.7%–86.2% (spread 58.5pp) across 2 papers. Sources: 'HumanEval-V: Benchmarking High-Level Vis' (27.7%);. Prompt engineering reduces reasoning mistakes in Large Language Models (LLMs). However, its effectiveness in mitigating vulnerabilities in LLM-generated code remains underexplored. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: Benchmarking Prompt Engineering Techniques for Secure Code Generation with GPT Models. Research question: Benchmark archaeology: investigate HumanEval score discrepancy for GPT-4o — reported 27.7%–86.2% (spread 58.5pp) across 2 papers. Sources: 'HumanEval-V: Benchmarking High-Level Vis' (27.7%); 'FeedbackEval: A Benchmark for Evaluating' (86.2%). Identify evaluation protocol differences (few-shot, prompting, preprocessing)..

## 2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.2/10.

## 3 Results

13 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 5.2/10.

## 4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

## References


- http://arxiv.org/abs/2308.10783v2
- http://arxiv.org/abs/2502.06039v1
- http://arxiv.org/abs/2410.12381v3