State-of-the-Art Language Models for HumanEval Code Generation: A Systematic Survey

Assignee Research

June 3, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: HumanEval code generation state of the art language model survey. 16 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Survey of AIOps in the Era of Large Language Models. Research question: HumanEval code generation state of the art language model survey.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

15 papers retrieved. 16 claims extracted; 1 independently verified. Quality review score: 4.5/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
The survey is organized around four research questions targeting critical aspects of how large language models are resha	×	0.10
RQ1 focuses on data sources and preprocessing techniques in AIOps.	×	0.12
RQ2 examines the evolution of AIOps tasks with the advent of LLMs.	×	0.14
RQ3 reviews LLM-based methods in AIOps.	✓	0.18
RQ4 discusses evaluation methodologies in AIOps adapted to the integration of LLMs.	×	0.09
The systematic review process includes defining the search strategy and scope, setting inclusion and exclusion criteria,	×	0.05
The search strategy includes five databases: Scopus, Web of Science, IEEE Xplore, ACM Digital Library, and arXiv.	×	0.05
The search is limited to papers published after 2020.	×	0.06
761 papers were identified from the databases search: 221 from ACM, 238 from Scopus, 81 from WoS, 110 from IEEE, and 111	×	0.02
614 unique papers remained after removing duplicates.	×	0.01
333 papers were excluded based on exclusion criteria: 51 for no large language model, 115 for not related to software, 3	×	0.04
395 papers were selected for full text review.	×	0.03
232 papers were removed during full text review: 98 for no large language model, 72 for not related to software, 4 for n	×	0.04
163 new studies met the inclusion criteria.	×	0.07
The raw log parsing event templates include entries such as 'instruction cache parity error corrected' and 'generating c	×	0.01
The software system tasks include remediation, analysis, answer query, assisted questioning, failure incident, OCE, larg	×	0.04

References

- http://arxiv.org/abs/2509.16679v1
- http://arxiv.org/abs/2507.12472v1
- http://arxiv.org/abs/2402.12317v2