Open-Source vs. Proprietary Code LLMs: HumanEval Performance in Multi-Language Generation

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: How does the fine-tuning of open-source code LLMs like WizardCoder on Evol-Instruct compare to proprietary models in terms of HumanEval pass@1 accuracy, particularly when evaluated on multi-language. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: WizardCoder: Empowering Code Large Language Models with Evol-Instruct. Research question: How does the fine-tuning of open-source code LLMs like WizardCoder on Evol-Instruct compare to proprietary models in terms of HumanEval pass@1 accuracy, particularly when evaluated on multi-language code generation tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

9 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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
HumanEval comprises 164 problems with an average of 9.6 test cases per problem.	×	0.05
HumanEval+ expands the test cases significantly to an average of 774.8 per problem.	×	0.05
MBPP provides 500 test programming problems with three automated test cases each.	×	0.03
WizardCoder models are compared with closed-source models such as GPT4, Claude, and Bard on the EvalPlus leaderboard.	×	0.07
All models generate code solutions for each problem utilizing a single attempt, and the resulting pass rate percentage i	×	0.03
WizardCoder demonstrates a significant superiority over all other models when tackling data science problems on the DS-1	×	0.07
WizardCoder models are evaluated with the same set of hyper-parameters: temperature=0.2, top p=0.95, max length=512, and	×	0.02
WizardCoder-15B achieves a pass@1 score of 57.3% on HumanEval and 51.8% on MBPP.	×	0.04
WizardCoder-34B achieves a pass@1 score of 71.5% on HumanEval and 61.2% on MBPP.	×	0.04
The DS-1000 benchmark comprises 1k distinct data science workflows spanning 7 libraries.	×	0.05
The MultiPL-E benchmarks encompass 8 distinct programming languages: Java, JavaScript, C++, PHP, R, Julia, Swift, and Ru	×	0.06
WizardCoder models surpass the SOTA open-source Code LLMs across all evaluated programming languages on the MultiPL-E be	×	0.14

References

- http://arxiv.org/abs/2410.12381v3
- http://arxiv.org/abs/2306.08568v2
- http://arxiv.org/abs/2602.06370v1