

# Synthetic-Vulnerable Code Ratios and Their Effect on Code Llama’s Benchmark Performance

Assignee Research

June 4, 2026

## Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: What is the impact of varying the ratio of synthetic-vulnerable to standard code samples in the training data on Code Llama’s performance on HumanEval and MBPP code generation benchmarks, measured by pass@k scores? 15 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. 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: What is the impact of varying the ratio of synthetic-vulnerable to standard code samples in the training data on Code Llama’s performance on HumanEval and MBPP code generation benchmarks, measured by pass@k scores?.

## 2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

## 3 Results

10 papers retrieved. 15 claims extracted; 0 independently verified. Quality review score: 3.3/10.

## 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.	×	0.04
HumanEval+ expands the test cases significantly to an average of 774.8 per problem.	×	0.04
MBPP provides 500 test programming problems with three automated test cases each.	×	0.04
WizardCoder-34B achieved a pass@1 score of 71.5% on the HumanEval benchmark.	×	0.04
WizardCoder-34B achieved a pass@1 score of 61.2% on the MBPP benchmark.	×	0.02
WizardCoder-15B achieved a pass@1 score of 57.3% on the HumanEval benchmark.	×	0.04
WizardCoder-15B achieved a pass@1 score of 51.8% on the MBPP benchmark.	×	0.02
GPT-4 achieved a pass@1 score of 67.0% on the HumanEval benchmark.	×	0.05
CodeLlama-Python (34B) achieved a pass@1 score of 53.7% on HumanEval and 56.2% on MBPP.	×	0.06
WizardCoder models demonstrated superior performance across all 8 evaluated programming languages (Java, JavaScript, C++	×	0.12
The DS-1000 benchmark comprises 1,000 distinct data science workflows spanning 7 libraries.	×	0.05
WizardCoder demonstrates significant superiority over all other models on the DS-1000 benchmark insertion scores.	×	0.06
WizardCoder models were evaluated using hyper-parameters temperature=0.2 and top_p=0.95 for HumanEval and MBPP results.	×	0.05
WizardCoder models were evaluated with n=200 samples to estimate the pass@1 score on HumanEval and MBPP.	×	0.05
MultiPL-E evaluations were conducted with hyper-parameters temperature=0.2, top_p=0.95, max_length=512, and n=50.	×	0.02

## References

- <http://arxiv.org/abs/2412.21199v2>

- <http://arxiv.org/abs/2306.09896v5>
- <http://arxiv.org/abs/2306.08568v2>