SOVEREIGN: Measuring Coding Challenge Competence With APPS

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

While programming is one of the most broadly applicable skills in modern society, modern machine learning models still cannot code solutions to basic problems. Despite its importance, there has been surprisingly little work on evaluating code generation, and it can be difficult to accurately assess code generation performance rigorously. To meet this challenge, we introduce APPS, a benchmark for code generation. Unlike prior work in more restricted settings, our benchmark measures the ability of models to take an arbitrary natural language specification and generate satisfactory Python code. S

1 Introduction

Analysis of: Measuring Coding Challenge Competence With APPS. Research goal: Does the adversarial robustness gap between DeepSeek-R1 and o1-preview on legal reasoning tasks under negation-based token perturbations replicate across code generation benchmarks like Codeforces or APPS when using the S* selection mechanism?.

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

10 papers retrieved. 7 claims extracted, 7 verified. Tribunal: 7.8/10 $\rightarrow$ 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
Modern machine learning models still cannot code solutions to basic problems.	✓	0.30
There has been surprisingly little work on evaluating code generation.	✓	0.24
APPS is a benchmark for code generation that measures the ability of models to take an arbitrary natural language specif	✓	0.34
The APPS benchmark includes 10,000 problems, ranging from simple one-line solutions to substantial algorithmic challenge	✓	0.28
The prevalence of syntax errors is decreasing exponentially as models improve.	✓	0.22
Recent models such as GPT-Neo can pass approximately 20% of the test cases of introductory problems.	✓	0.31
Machine learning models are now beginning to learn how to code.	✓	0.25

References

- http://arxiv.org/abs/2504.20834v4
- http://arxiv.org/abs/2407.15549v3
- http://arxiv.org/abs/2105.09938v3