Ensemble-Based Uncertainty and Functional Correctness in Code Generation Models

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: Do ensemble-based uncertainty estimates in code generation models correlate with functional correctness rates across varying difficulty levels in the HumanEval benchmark. 15 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Ensemble-Based Uncertainty Estimation for Code Correctness Estimation. Research question: Do ensemble-based uncertainty estimates in code generation models correlate with functional correctness rates across varying difficulty levels in the HumanEval benchmark?.

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 4.2/10.

3 Results

15 papers retrieved. 15 claims extracted; 2 independently verified. Quality review score: 4.2/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
In selective generation tasks with strict False Positive Rate (FPR) constraints, Ensemble Semantic Entropy (ESE) yields	✓	0.29
The Cascading Test-Time Scaling framework (Cas) reduces FLOPs consumption by 64.9% compared to standard Best-of-N scalin	✓	0.18
Experiments were conducted on the LiveCodeBench (Release v2) benchmark, which consists of recent contest problems from A	×	0.03
The behavior-based clustering method named 'Func' groups programs by their execution outputs on a problem-specific gener	×	0.05
For each problem in the experiments, M=12 sampled programs were considered.	×	0.02
In the ensemble setting, samples were evenly drawn from two models and merged before uncertainty computation.	×	0.07
For the GLM4-9B model, the Predictive Entropy (PE) score is -0.0899.	×	0.02
For the GLM4-9B model, the Semantic Entropy (SE) score is -0.7352.	×	0.05
For the GLM4-9B model, the Ensemble Semantic Entropy (ESE) score is -0.7391.	×	0.12
For the Qwen3-8B model, the Semantic Entropy (SE) score is -0.6735.	×	0.05
For the Qwen3-14B model, the Semantic Entropy (SE) score is -0.6852.	×	0.05
For the Qwen3-Coder-30B model, the Semantic Entropy (SE) score is -0.6686.	×	0.05
For the Qwen3-Coder-485B model, the Semantic Entropy (SE) score is -0.6304.	×	0.05
For the gpt-oss-20B model, the Semantic Entropy (SE) score is -0.5960.	×	0.05
For the gpt-oss-120B model, the Semantic Entropy (SE) score is -0.6751.	×	0.05

References

- http://arxiv.org/abs/2410.12381v3

- http://arxiv.org/abs/2511.07364v1
- http://arxiv.org/abs/2603.27098v2