SOVEREIGN: S*: Test Time Scaling for Code Generation

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

Increasing test-time compute for LLMs shows promise across domains but remains underexplored in code generation, despite extensive study in math. In this paper, we propose S*, the first hybrid test-time scaling framework that substantially improves the coverage and selection accuracy of generated code. S* extends the existing parallel scaling paradigm with sequential scaling to push performance boundaries. It further leverages a novel selection mechanism that adaptively generates distinguishing inputs for pairwise comparison, combined with execution-grounded information to robustly identify co

1 Introduction

Analysis of: S*: Test Time Scaling for Code Generation. Research goal: Does the adversarial robustness gap between DeepSeek-R1 and o1-preview on legal reasoning tasks generalize to code generation benchmarks under negation-based token perturbations?.

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

12 papers retrieved. 7 claims extracted, 7 verified. Tribunal: 7.5/10 $\rightarrow$ 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
S* is the first hybrid test-time scaling framework that substantially improves the coverage and selection accuracy of ge	✓	0.38
S* extends the existing parallel scaling paradigm with sequential scaling to push performance boundaries.	✓	0.29
S* leverages a novel selection mechanism that adaptively generates distinguishing inputs for pairwise comparison, combin	✓	0.38
S* consistently improves performance across model families and sizes, enabling a 3B model to outperform GPT-4o-mini.	✓	0.35
S* enables non-reasoning models to surpass reasoning models - GPT-4o-mini with S* outperforms o1-preview by 3.7% on Live	✓	0.37
S* further boosts state-of-the-art reasoning models - DeepSeek-R1-Distill-Qwen-32B with S* achieves 85.7% on LiveCodeBen	✓	0.37
Code will be available under https://github.com/NovaSky-AI/SkyThought.	✓	0.24

References

- http://arxiv.org/abs/2412.16117v1
- https://www.semanticscholar.org/paper/56c2d8a39ec396c54e0d42c9beab88f45e24886c
- http://arxiv.org/abs/2305.00866v2