AlgoVeri: An Aligned Benchmark for Verified Code Generation on Classical Algorithms

Haoyu Zhao [*1,2] Ziran Yang [*1,2] Jiawei Li [*4] Deyuan He [*2] Zenan Li [*3]
Chi Jin [1,2] Venugopal V. Veeravalli [4] Aarti Gupta [2] Sanjeev Arora [1,2]

1 Princeton Language Intelligence 2 Princeton University 3 ETH Zürich 4 University of Illinois Urbana-Champaign * Core Contributor

Paper: 2602.09464 ***Data:*** zzzzzhy/algoveri

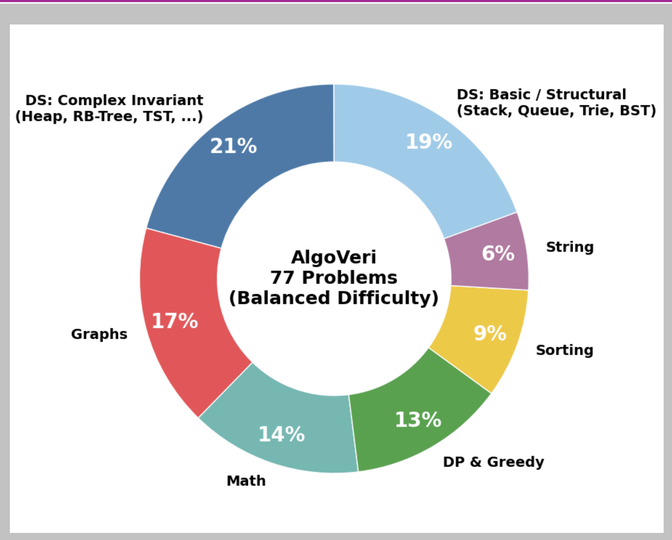
Summary

AlgoVeri consists of code generation tasks with **semantically aligned specifications** of 77 classical algorithms in Dafny, Verus and Lean.

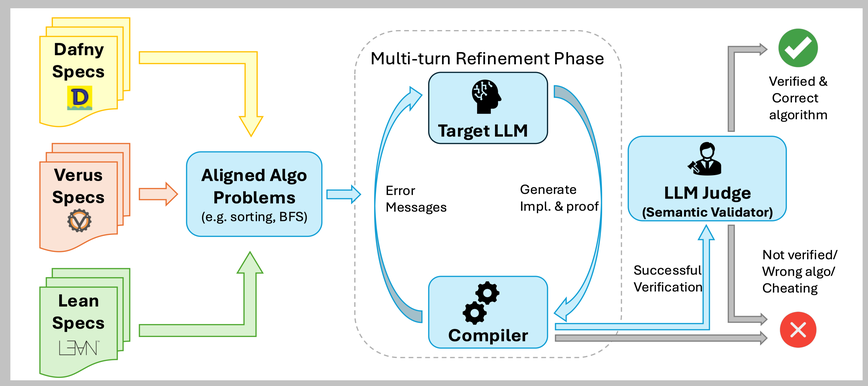
Highlights

- Aligned specifications enables **fair evaluation**
- Expert curation and formal well-formedness proofs guard the quality of specifications
- Includes **challenging global properties** of classical algorithms (e.g., max-flow)
- Enables analysis of **failure dynamics** under different verification language / environment

Task Breakdown



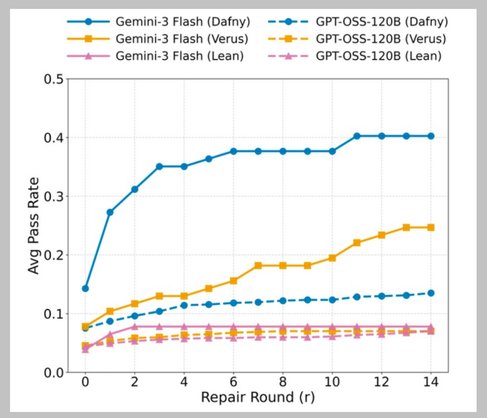
Evaluation Pipeline



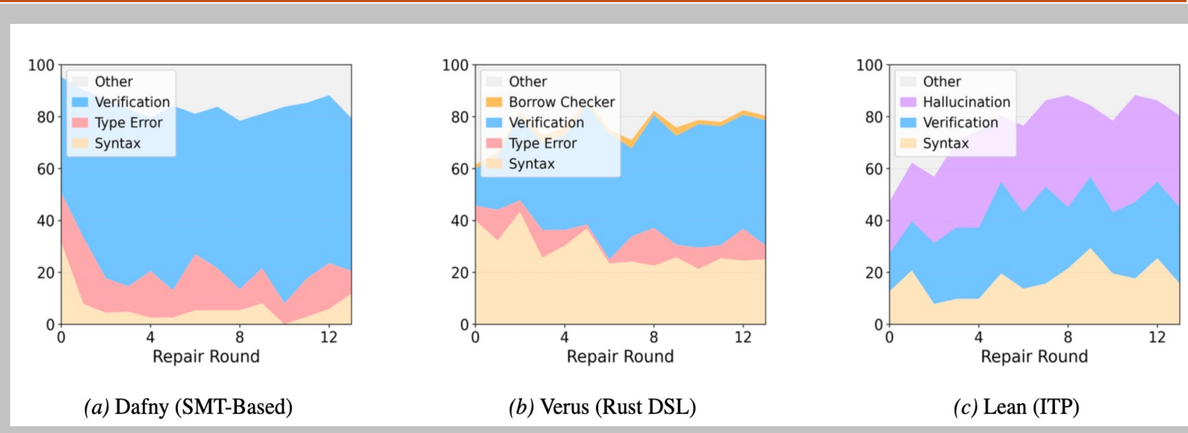
Benchmark Results

	Budget	Dafny		Verus		Lean	
		Compiler Verified	+ Semantic Filtered	Compiler Verified	+ Semantic Filtered	Compiler Verified	+ Semantic Filtered
GPT-5.3 Codex	1×8	49.35	42.86	14.29	11.69	23.38	11.69
Gemini-3 flash	1×15	55.84	40.26	25.97	24.68	9.09	7.79
GPT-5 mini	1×15	41.56	30.47	7.79	6.49	5.19	5.19
GPT-OSS-120B	1×15	21.04±2.23	13.51±1.66	7.66±0.91	7.01±1.04	12.60±2.25	7.01±1.19
	10×15	44.16	28.57	12.99	10.39	25.97	14.29
Qwen3-235B-A22B-Instruct	1×15	25.32±2.19	18.31±2.43	9.48±1.43	9.09±1.54	3.77±1.23	3.77±1.23
	10×15	32.47	29.87	12.99	12.99	6.49	6.49
Qwen3-Next-80B-A3B-Thinking	1×15	23.77±2.02	17.27±2.18	7.66±0.91	6.49±0.82	3.64±0.97	3.51±1.17
	10×15	35.06	33.77	11.69	10.39	5.19	5.19
Devstral-2-123B-Instruct	1×15	9.09±1.93	6.10±1.43	8.05±0.97	7.27±1.19	3.51±1.17	3.38±1.19
	10×15	33.77	18.18	14.29	12.99	6.49	6.49

Scaling Analysis and Failure Dynamics



Key finding 1: closed-source LLMs are better at horizontal scaling in iterative repairing.



(a) Dafny (SMT-Based) *(b)* Verus (Rust DSL) *(c)* Lean (ITP)

Key finding 2: Dafny offers smoother repair process; LLMs face *syntax barrier* in Verus (Rust) and *search barrier* in Lean (e.g., proof search and lemma hallucination).

PRINCETON Language + Intelligence

PRINCETON UNIVERSITY

ETH zürich