UC San Diego | NEW YORK UNIVERSITY | UNIVERSITY of WASHINGTON | PRINCETON UNIVERSITY

AutoCode: LLMs as Problem Setters for Competitive Programming

Shang Zhou [1,*], Zihan Zheng [2,*], Kaiyuan Liu [3,*], Zeyu Shen [4,*], Zerui Cheng [4,*], Zexing Chen [1], Hansen He [5], Jianzhu Yao [4], Huanzhi Mao [7], Qiuyang Mang [7], Tianfu Fu [6], Beichen Li [8], Dongruixuan Li [9], Wenhao Chai [4,†], Zhuang Liu [4,†], Aleksandra Korolova [4,†], Peter Henderson [4,†], Natasha Jaques [3,†], Pramod Viswanath [4,10,†], Saining Xie [2,†], Jingbo Shang [1,†]

[1] UC San Diego [2] New York University [3] University of Washington [4] Princeton University [5] Canyon Crest Academy [6] OpenAI [7] UC Berkeley [8] MIT [9] University of Waterloo [10] Sentient Labs

Motivation

Problem setting requires a **deeper understanding of algorithms** than solving. It includes all challenges of solving and even more. Current benchmarks suffer from:

- **High FPR:** Incorrect solutions pass weak test suites.
- **High FNR:** Valid solutions crash on bad inputs.

This **distorts evaluation** and pollutes RL training data. AutoCode solves these problems to build better datasets.

AutoCode Framework

A closed loop **Validator, Generator, and Checker** framework. It mirrors how human experts create contest problems.

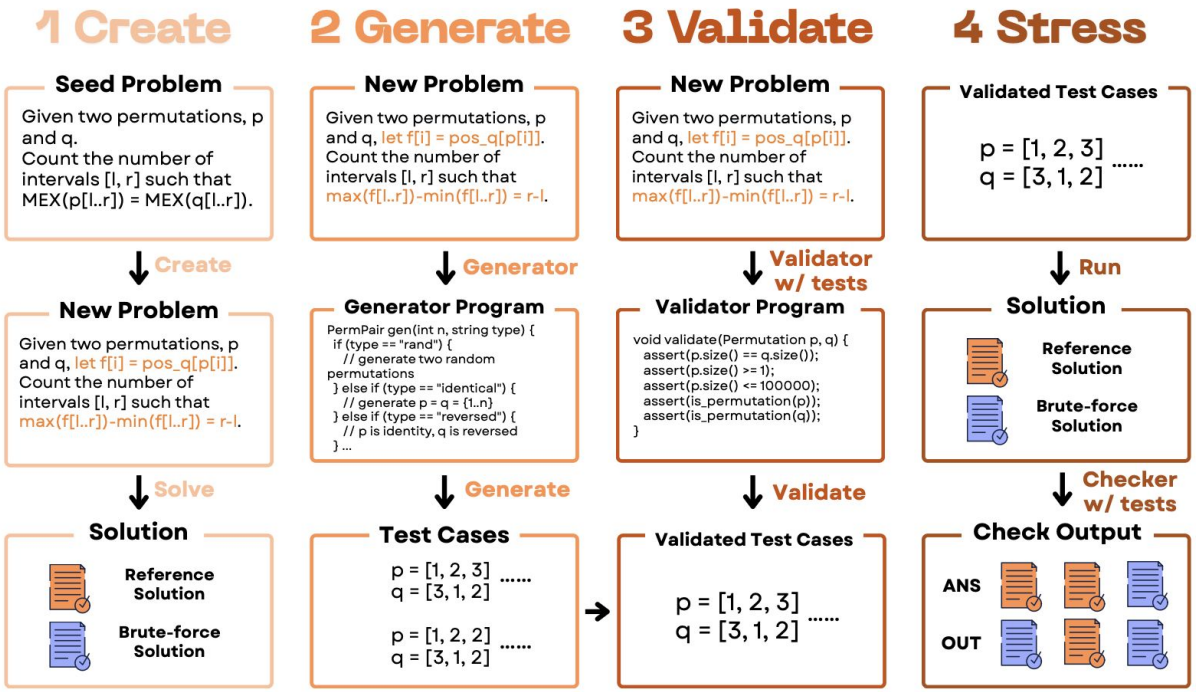

Figure 1. AutoCode pipeline: (1) Create from seed, (2) Generate test cases, (3) Validate inputs, (4) Stress test via dual verification.

Test Case Generation Results

Table 1. **7,538 problem benchmark.** 195,988 submissions, 50% correct / 50% incorrect. AutoCode uses o3.

Method	Consistency ↑	FPR ↓	FNR ↓
CodeContests	72.9%	7.7%	46.3%
CodeContests+	79.9%	8.6%	31.6%
TACO	80.7%	11.5%	26.9%
HardTests	81.0%	12.1%	25.8%
AutoCode (Ours)	**91.1%**	**3.7%**	**14.1%**

720 Problem Benchmark (Unfiltered Codeforces)

98.7%	**1.3%**	**1.2%**
Consistency	FPR	FNR

Ablation Study

Table 2. **Ablation on 720 problems.** GPT 5 High, 33 submissions per problem.

Configuration	Consist. ↑	FPR ↓	FNR ↓
w/o Exhaustive	98.4%	1.7%	1.3%
w/o Random/Extreme	98.4%	1.6%	1.3%
w/o TLE Inducing	98.6%	1.4%	1.3%
w/o Prompt Optimization	98.0%	1.8%	2.9%
Full Framework	**98.7%**	**1.3%**	**1.2%**

Novel Problem Generation

AutoCode creates new variants from seed problems with reference solutions for verification.

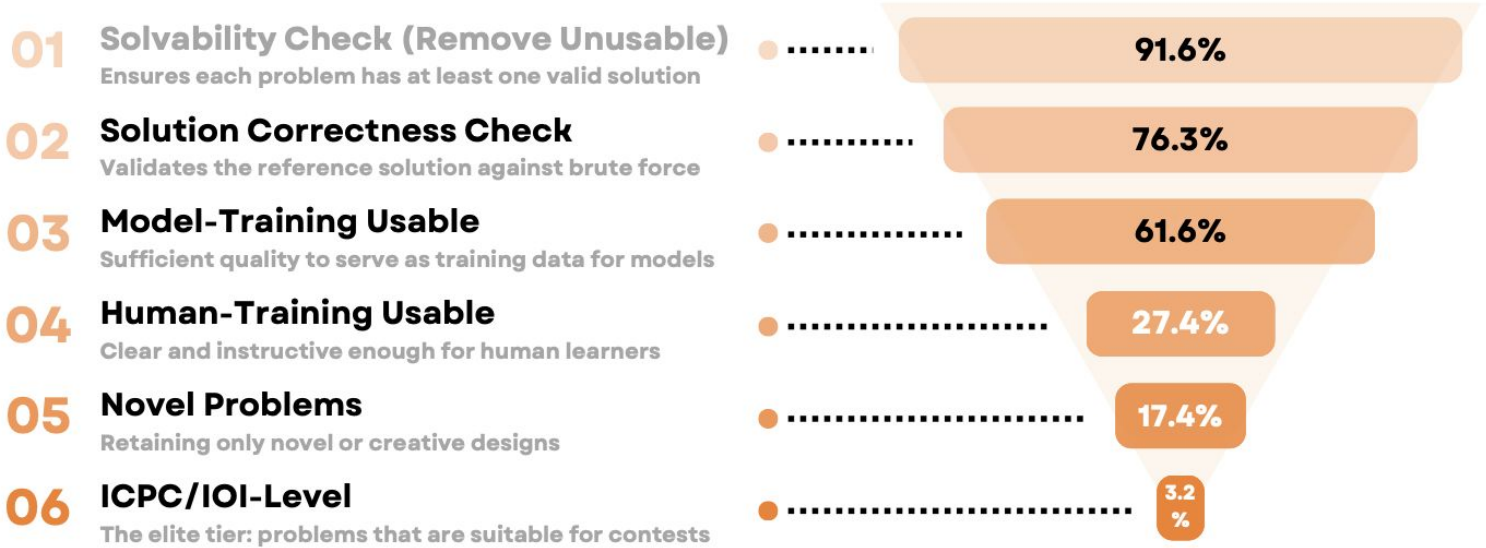

Figure 2. **Quality levels.** 61.6% suitable for training; 3.2% reach ICPC/IOI.

Key Findings

- **Beyond limits:** LLMs create problems they cannot solve (4.2%).
- **Combining ideas:** LLMs mix existing frameworks over inventing new ones.
- **Elo gain:** Novel problems yield +498 Elo compared to +108.
- **Evaluation gap:** Human and LLM quality correlation is only $r = 0.07$.
- **Difficulty proxy:** Difficulty aligns well with human rating ($r = 0.60$).

Conclusion

- **98.7%** match with official judgments
- **50%** drop in both FPR and FNR
- **61.6%** of problems are good for training