

# LiveCodeBench Pro

## How Do Olympiad Medalists Judge LLMs in Competitive Programming?

*Presenter: Zerui Cheng (Princeton University)*

*<https://www.zerui-cheng.com>*



The Ultimate Test for AI in Deep Algorithmic Reasoning

Leaderboard: <https://www.livecodebenchpro.com>

Tech Report: <https://arxiv.org/pdf/2506.11928>

# Our Team

An expert team of ICPC Gold medalists, World Finalists, experienced problem setters,  
and top-tier AI researchers with over 1,000 papers and 200,000 citations in total.

Zihan Zheng <sup>1,\*</sup>, Zerui Cheng <sup>2,\*</sup>, Zeyu Shen <sup>2,\*</sup>, Shang Zhou <sup>3,\*</sup>, Kaiyuan Liu <sup>4,\*</sup>, Hansen He <sup>5,\*</sup>,  
Dongruixuan Li <sup>6</sup>, Stanley Wei <sup>2</sup>, Hangyi Hao <sup>7</sup>, Jianzhu Yao <sup>2</sup>, Peiyao Sheng <sup>8</sup>, Zixuan Wang <sup>2</sup>,  
Wenhao Chai <sup>2,†,§</sup>, Aleksandra Korolova <sup>2,†</sup>, Peter Henderson <sup>2,†</sup>, Sanjeev Arora <sup>2,†</sup>,  
Pramod Viswanath <sup>2,8,†</sup>, Jingbo Shang <sup>3,†,‡</sup>, Saining Xie <sup>1,†,‡</sup>

<sup>1</sup> New York University

<sup>2</sup> Princeton University

<sup>3</sup> University of California San Diego

<sup>4</sup> University of Washington

<sup>5</sup> Canyon Crest Academy

<sup>6</sup> University of Waterloo

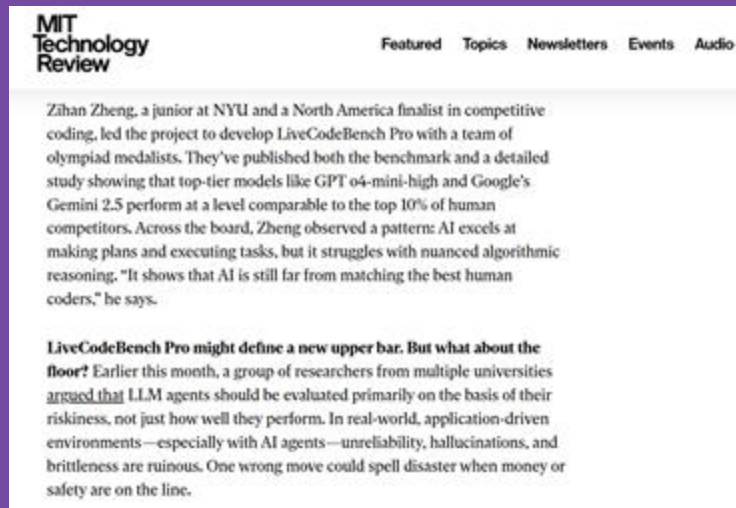
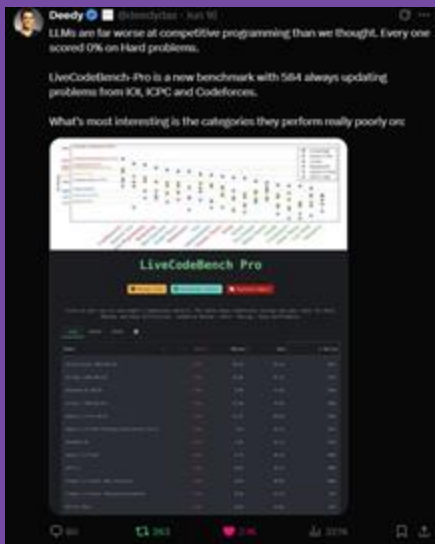
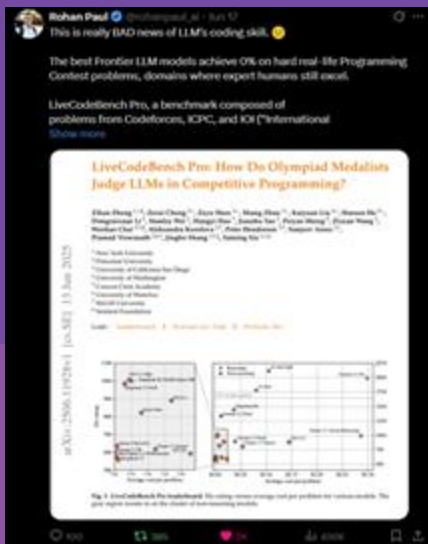
<sup>7</sup> McGill University

<sup>8</sup> Sentient Foundation



# Media Coverage

- Tweeted by top AI influencers, accumulating over 1M views in total on X
- Covered by MIT Technology Review on June 24, 2025





# TL; DR: What is LiveCodeBench Pro?

LiveCodeBench Pro is a high-quality **competitive programming** benchmark which tests **the genuine deep algorithmic reasoning abilities** of the state-of-the-art AI models with **detailed diagnostics** handcrafted by an expert team of Olympiad medalists and AI researchers.

## Top 3 insights from our evaluation:

- ▶ Models lag humans on hardest problems (0% on Hard tier for any model)
- ▶ Structured logic  $\gg$  creativity for reasoning models
- ▶ Tool use inflates scores (Bayesian Elo w/o tools  $\approx 2116$  vs. 2700+ reported)



# Agenda

- ▶ **Motivation** – What is Competitive Programming (CP) & Why It Matters
- ▶ **Current Gaps** – Limitations of Current CP Benchmarks
- ▶ **Our Solution and Main Results** – A Quick Glance at LiveCodeBench Pro
- ▶ **Deep Diagnostics** – Fine-grained Annotations & Error Analysis
- ▶ **Open Questions & Discussion**
- ▶ **Q&A**

# 1. Motivation

What is Competitive Programming (CP) & Why It Matters



# What is Competitive Programming?

- ▶ Think of LeetCode on extreme steroids - but way, way harder!
- ▶ It's essentially mathematics with code - pure algorithmic reasoning

- ▶ Problems require deep insights from:



Number theory and combinatorics



Graph theory and dynamic programming



Game theory and optimization



Complex data structures



Carefully curated test cases

ensure no guessing

-

only pure reasoning

- ▶ Success requires both mathematical insight and flawless implementation



# Why Is This Perfect for AI Evaluation?

- ▶ **Ultimate objectivity** - fully automated evaluation, no subjective judgment, only pass/fail
- ▶ **Exhaustive hidden test suites** - impossible to game or guess
- ▶ **Pure reasoning challenge** - tests the very edge of human cognitive abilities
- ▶ **Unified environment** - same hardware, same constraints, fully replicable

Unlike ultimate math challenges, the evaluation is

**100% free from human graders,  
fully automated, objective, and robust**

No ambiguity in correctness - either your algorithm works or it doesn't





# Competitive Programming - Industry Gold Standard

The first wide-adopted benchmark in competitive programming, LiveCodeBench, has been used by major AI labs for model evaluation, and reflected in their model release reports.



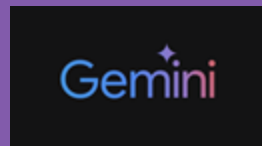
OpenAI

Model cards & releases



Anthropic

Claude evaluations



Google

Gemini assessments

The **de facto standard** for measuring deep algorithmic reasoning in LLMs











## 2. Current Gaps








Is LiveCodeBench good enough? NOT QUITE!



# Limitations of LiveCodeBench

**Low differentiation:** Top reasoning models solve ~80% of tasks while non-reasoning models can solve over 65% of tasks.

1	 o3 	 83.9%
2	 Grok 4	 83.2%
3	 o4 Mini 	 82.2%
4	 Gemini 2.5 Pro Preview	 79.2%

9	 DeepSeek R1	 70.2%
10	 Claude Opus 4 (Thinking)	 70.2%
11	 Grok 3 Mini Fast Low Reasoning	 66.3%
12	 DeepSeek V3 (03/24/2025)	65.5%













# Limitations of LiveCodeBench

No direct comparison with humans or Diagnostics for further improvement:

What does 80% solve rate imply?

Is it super-human intelligence or just average human level?

How does its reasoning pattern compare with a human at the same level?

1	 o3 	 83.9%
2	 Grok 4	 83.2%
3	 o4 Mini 	 82.2%
4	 Gemini 2.5 Pro Preview	 79.2%

Humans: ? %

Average CS major undergrads: ? %

Senior software engineers at FAANG: ? %

IOI gold medalists: ? %



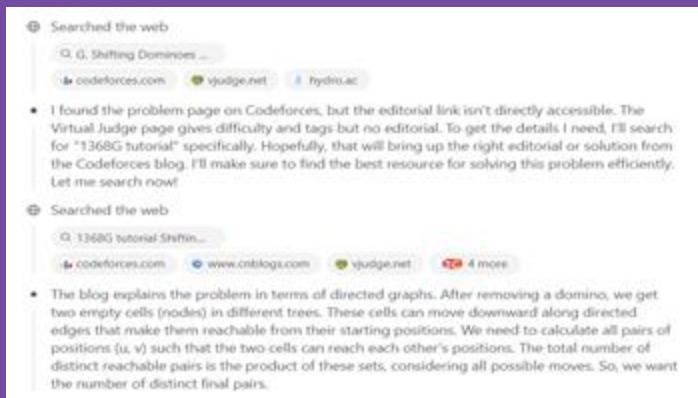
# Limitations of LiveCodeBench

Data contamination and exaggerated liveness claim:

Tasks are updated every 3-6 months for “liveness”,  
but solutions and editorials are out only 1-2 days after release of the tasks.

With tool usage, the solutions can be easily found on the Internet

→ not true deep algorithmic reasoning abilities



# 3. Our Solution and Main Results

A Quick Glance at LiveCodeBench Pro



# Introducing LiveCodeBench Pro

- ▶ 584 high-quality problems (still being updated live) from premier contests (Codeforces, ICPC, IOI)
- ▶ Real-time collection - captured and evaluated before any public solutions to prevent data contamination
- ▶ Bayesian Elo ratings - directly comparable to human levels
- ▶ Fine-grained annotation and analysis of algorithmic categories and failure modes by Olympiad medalists
- ▶ No LeetCode problems - only the hardest, most contamination-free challenges are included, representing the boundaries of human intelligence in algorithmic reasoning



# LiveCodeBench Pro - The Difficulty Spectrum

## Easy

≤2000 Elo Rating  
~15 minutes for world-class competitors

## Medium

2000-3000 Elo Rating  
Multiple algorithms + advanced reasoning required

## Hard

>3000 Elo Rating  
Defeats 99.9% of participants in competitions

Hard problems sometimes remain unsolved even by the strongest competitors during live contests!





# The Reality Check: Model Performance

53%

Best model (o4-mini-high)  
pass@1 on Medium problems

0%

ALL models  
pass@1 on Hard problems

2116

o4-mini-high rating  
vs 2700+ reported with tools  
top 1.5% among human competitors

Significant gap remains to human grandmaster levels, especially without external tools

## 4. Deep Diagnostics

Our findings - A deeper dive into the statistics



# Three Types of Cognitive Challenges



## Knowledge-Heavy

Templates, algorithms, deep mathematical results. Success depends on breadth of knowledge and implementation skill.

---

Examples: Segment Trees, FFT, Graph Algorithms



## Logic-Heavy

Step-by-step mathematical reasoning, systematic derivations, combinatorial analysis.

---

Examples: Dynamic Programming, Combinatorics



## Observation-Heavy

"Aha!" moments, creative insights, deductive leaps that collapse the problem space.

---

Examples: Greedy, Game Theory, Constructive

Each category tests different cognitive abilities and represents distinct challenges for AI reasoning



# Key Finding #1: The Skill Spectrum



## LLM Strengths

- **Knowledge-Heavy:** Segment trees, data structures
- **Logic-Heavy:** Combinatorics, DP, math
- **Implementation:** Bug-free, syntactically correct



## LLM Weaknesses

- **Observation-Heavy:** Game theory, greedy, ad-hoc
- **Case Work:** Edge cases and corner conditions
- **Interactive:** Dynamic problem-solving dialogue

LLMs excel at **structured reasoning**  
**creative insights**

but struggle with



# Key Finding #2: Error Analysis

Line-by-line analysis of 125 failed submissions from o3-mini vs humans:



## Conceptual Errors

- 64.2% more than humans (87 vs 53 out of 125)
- Algorithm logic errors
- Wrong observations
- Faulty mathematical reasoning



## Implementation Errors

- 62.5% less than humans (15 vs 40 out of 125)
- Syntax errors almost non-existent
- I/O handling consistently correct
- Initialization errors rare

56 out of 125 LLM submissions fail on given sample inputs - 410% more than human submissions!  
Models don't verify basic correctness - easy potential improvement with terminal usage

# 🔍 Key Finding #3: Multiple Attempts Matter

## pass@1 - 1793 Elo

o4-mini-medium performance on a single attempt

top 5% among human competitors

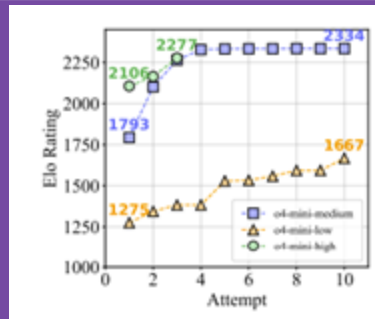
## pass@10 - 2334 Elo

o4-mini-medium performance after 10 attempts

top 1% among human competitors

↑ +541 improvement!

- Observation-heavy problems (Game Theory, Greedy, Case Work) benefit most from pass@k
- Making different hypotheses on different attempts without rigorously proving does the magic
- Points converge on pass@10 - still 400+ point gap to reported performance with tools
- Even with pass@k, 0% success rate on Hard problems





# The Power of External Tools

- ▶ Terminal access & tool calls explain the remaining ~400 Elo gap
- ▶ Local compilation: Catch syntax errors immediately
- ▶ Sample testing: Verify correctness on provided examples
- ▶ Brute-force validation: Generate test cases to find edge case bugs
- ▶ Pattern discovery: Run experiments to find algorithmic insights
- ▶ Search solution from the web: Shortcut to success without reasoning -> Liveness is important in evaluation

Without tools: native reasoning limitations become apparent  
With tools: Models can iteratively debug and improve solutions

# Key Finding #4: Reasoning vs Non-reasoning

Comparing DeepSeek R1 vs V3 and Claude 3.7 Sonnet (reasoning vs non-reasoning):



## Biggest Gains

- Combinatorics: +1400 Elo improvement on R1 vs V3
- Knowledge-Heavy:  
Data structures, segment trees show large gains



## Limited Gains

- Observation-Heavy:  
Game theory, greedy show minimal improvement
- Some categories even show negative improvement

Current reasoning methods excel in structured logic but have inherent limitations for creative problem-solving





# Key Implications from our Evaluation

- ▶ Claims of surpassing elite humans (which is unfortunately not true today) need serious qualification
- ▶ Models excel at implementation precision, not superior reasoning
- ▶ Creative insights and observations remain uniquely human strengths
- ▶ Claimed high performance largely driven by tool augmentation, not reasoning breakthroughs
- ▶ Significant room for improvement in edge case handling and algorithmic creativity
- ▶ Genuine liveness is important for future benchmarks to distinguish native reasoning from tool use

The gap to human grandmaster levels remains significant, especially in areas demanding novel insights and creativity.



## 5. Open Questions & Discussion



# LiveCodeBench Pro: What's next?

- ▶ **Problem creation:**

Could AI craft novel, hard algorithmic problems—and how would we ensure their rigor?

- ▶ **Model self-improvement:**

How might we enable models to test, critique, and refine their own solutions—without human-in-the-loop?

- ▶ **Recursive RL framework:**

What would an end-to-end loop of problem creation → evaluation → targeted improvement look like in practice?

**Thank you!**  
**Any questions?**