UniCode: Augmenting Evaluation for Code Reasoning

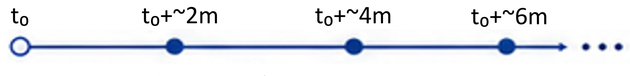
Xinyue Zheng[1†], Haowei Lin[1†], Shaofei Cai[1], Yaodong Yang[1], Zilong Zheng[2*], Yitao Liang[1*]
[1]Peking University, [2]BIGAI,

ICML International Conference On Machine Learning

PEKING UNIVERSITY 北京大学

WHY EXISTING BENCHMARKS BREAK

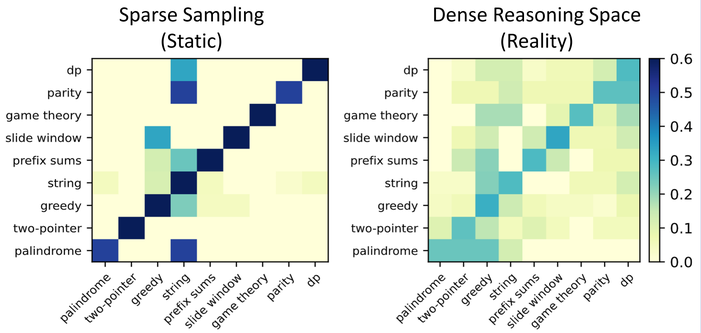
1. Slow Refresh



~2 months / update

Slow refresh + limited scale cannot keep up with rapid agent evolution.

2. Fixed paradigms



Human-curated tasks miss reasoning blind spots

3. Data Contamination

Public tasks become training data; scores become memory tests.

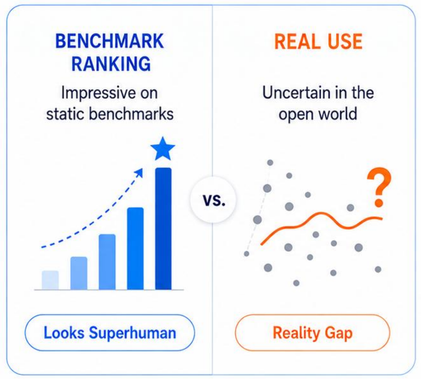
Gemini 3 Flash was given only the task ID `django__django-11099`, yet reproduced the task text and gold patch.

—**OpenAI**'s blog "***Why SWE-bench Verified no longer measures frontier coding capabilities***"

Static score ≠ genuine code reasoning.

AI CODE AGENTS: SUPERHUMAN?

Human-curated static benchmarks can be misleading for frontier agents.



Static human-curated benchmarks are flawed.

UNICODE REASONING STRESS TESTS

1 Seed Problem

Longest Increasing Subsequence (LIS)

Given an integer array A of length n, find the length of the longest strictly increasing subsequence. ($1 \le n \le 1000$)

Input: A = [3, 1, 2, 1, 8, 5, 6]
Output: 4 (e.g., [1, 2, 5, 6])

2 Narrative Perturbation

Longest Improving Period in Stock Prices

Given daily stock prices, find the longest period of strictly increasing prices.

Analysis: Same logic, different story.

Narrative perturbation

3 Rule Modification

Longest Non-Decreasing Subsequence

Find the longest subsequence that is non-decreasing (allows equal values).

Analysis: A single symbol changes the reasoning boundary.

Rule modification

4 Efficiency Scaling

Large-Scale LIS

Same as LIS, but now n can be up to 10^5.

Analysis: The solution must shift from $O(n^2)$ DP to $O(n \log n)$ greedy + binary search.

Efficiency scaling

5 Sequential Composition

Longest Bitonic Subsequence

Find the longest subsequence that first strictly increases, then strictly decreases.

Analysis: Compose LIS from the left and LIS from the right.

Sequential composition

6 Concept Fusion

Maximum-Weight Increasing Subsequence

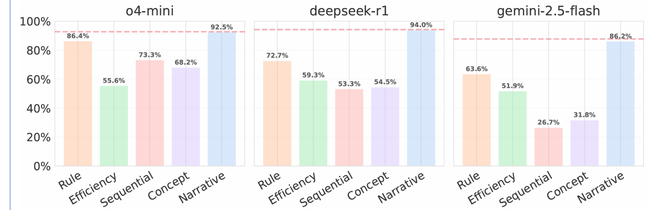
Among all strictly increasing subsequences, find the one with maximum total sum.

Analysis: Joint DP optimization of increasing order and maximization.

Concept fusion

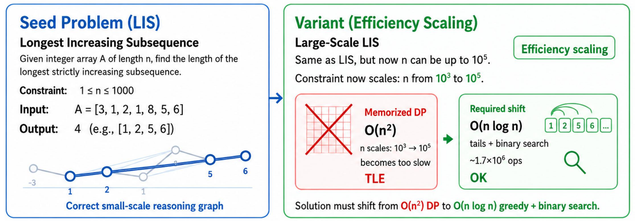
KEY FINDINGS

Model Performance Across Reasoning Variants



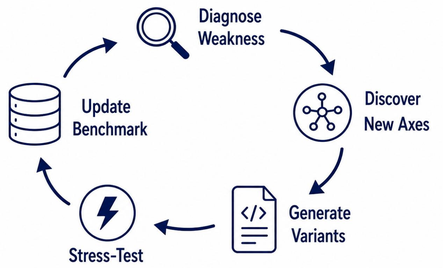
- Performance Collapse: Average 31.2% drop with high volatility (up to 61% variance) across reasoning axes.
- Reasoning Fragility: Model capabilities collapse when the underlying reasoning graph topology is altered.

Seed-Regression: A Hidden Failure Mode



Models frequently regress to memorized seed paradigms.

TOWARD EVOLVABLE EVALUATION



Variant-L1 → Variant-L2 → …

Github Page

Paper