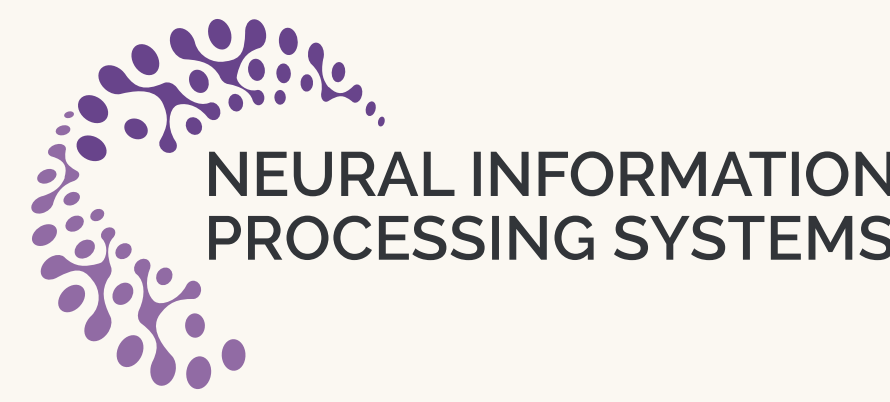


Solving Every Step Is Not Enough

Milestone Oracles Reveal a **Composition Gap** in LLM Math Reasoning

Zhuohan Wang^{1*†‡}, Haoran Ma^{1*}, Tianyu Wu^{1*}, Yuanlin Duan², Zichun Liao¹, Jieming Yu^{3†}

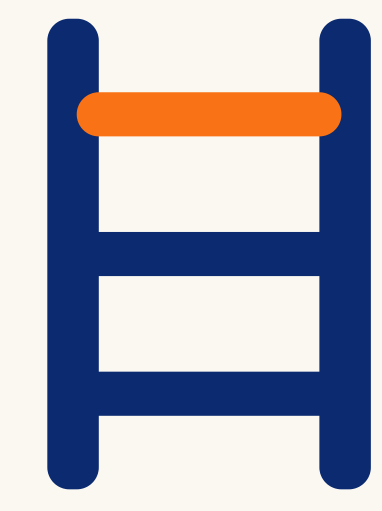
¹Harvard University ²Rutgers University ³The Hong Kong University of Science and Technology ^{*}Equal contribution [†]Corresponding [‡]Project lead



HARVARD
UNIVERSITY

RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

HKUST



OracleLadder



Code & data

github.com/slark-prime/OracleLadder



Paper

arXiv 2609.32235

33–48% of problems: every milestone solved alone, roadmap + all milestone answers given — still fails.

This **composition gap** is the largest of five gaps in all six open models (2024–25, 8B–671B) · 24–37% after removing flagged grading errors

1 Why ask where a model fails?

You have probably done this yourself: you worked out every piece of a hard problem and still got it wrong.

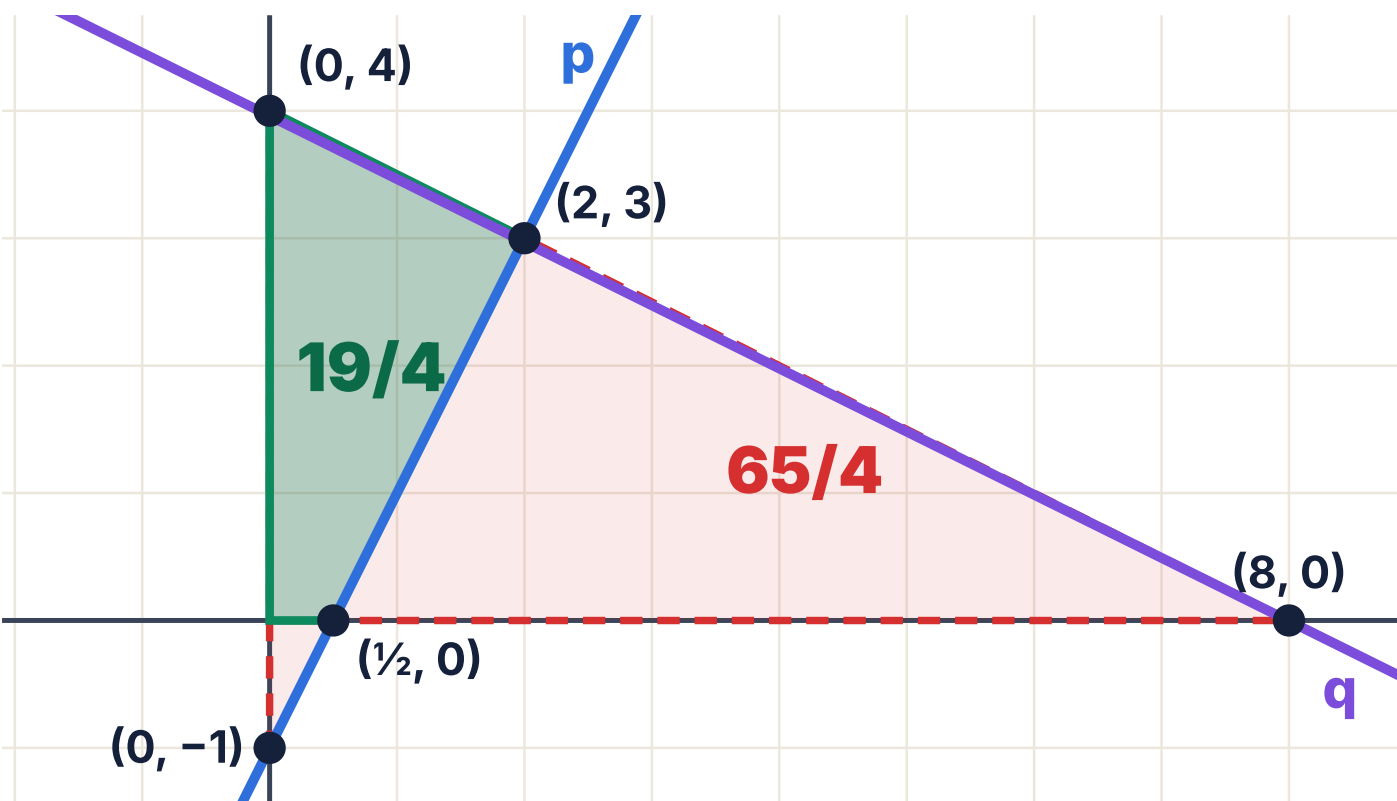
Background Math benchmarks, from GSM8K to AIME, track LLM reasoning, and RL with verifiable rewards trains on the same final-answer check.

Problem Both see only whether the final answer is right. A failure looks the same whether the model lacked a plan, missed a step, or held every piece and never put them together. **Each needs a different fix.**

We OracleLadder holds each problem fixed and adds help one rung at a time, so the lowest rung that works shows where the model fails.

A real case: every step right, the assembly wrong

Line $p : y = 2x - 1$. Through $(2, 3)$ on p , draw $q \perp p$. Area of the convex quadrilateral bounded by p , q , and the axes?



Qwen3-8B (2025, thinking off) · 8 attempts each	correct
M1 find line $q : y = -\frac{1}{2}x + 4$	8 / 8
M2 where p crosses the x-axis	8 / 8
C1 no help	0 / 8
C2 + roadmap	0 / 8
C3 + roadmap + both answers	0 / 8

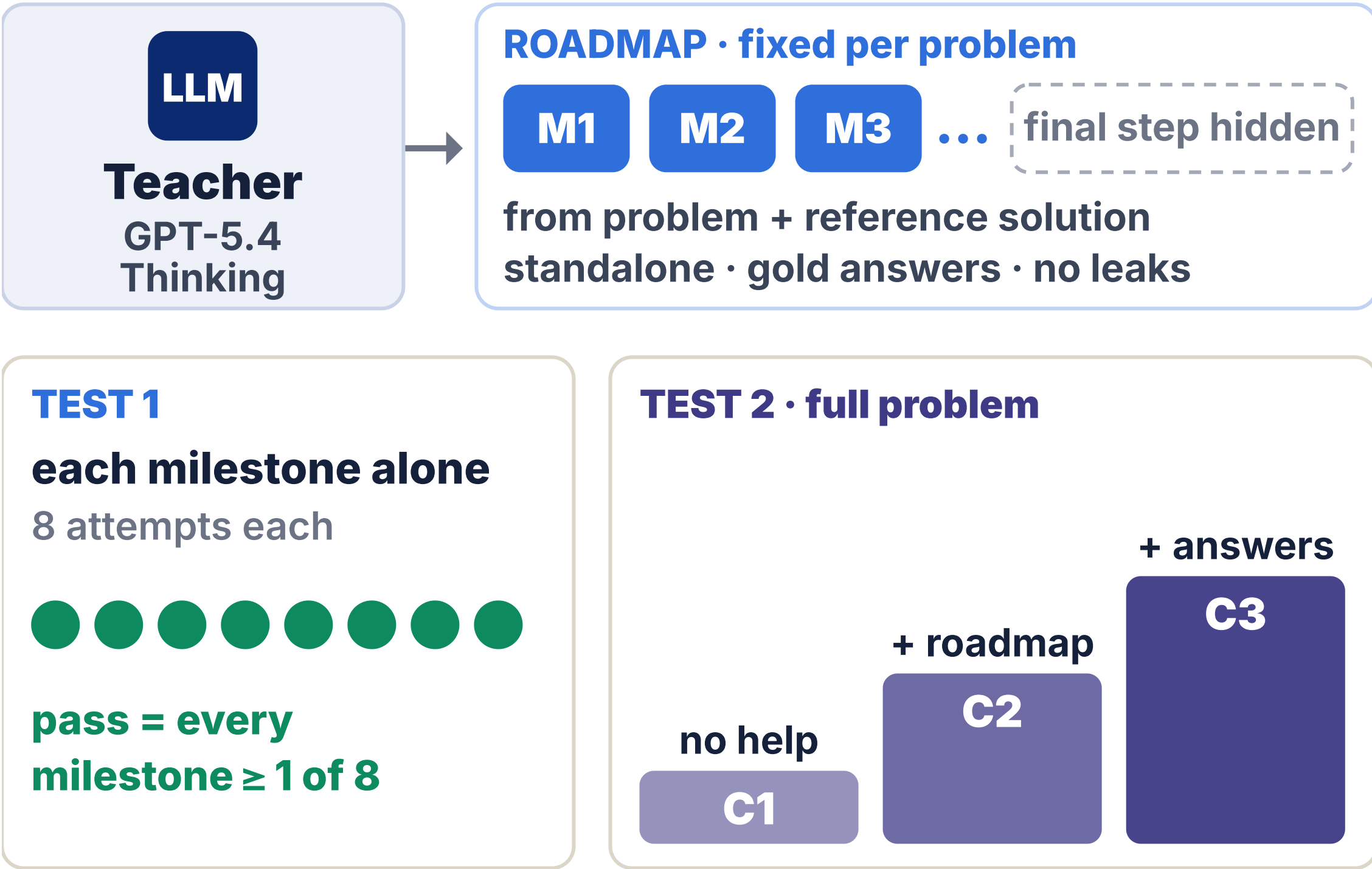
Passes the milestone test, never solved → **composition gap**

What the failure looks like (C3, all 8 answers read): every attempt leaves out the corner $(2, 3)$. Most common answer **65/4** (dashed). True area **19/4**.

“(2, 3) ... isn’t on the x-axis or y-axis, so it is not part of the convex quadrilateral” — one attempt, verbatim

2 OracleLadder: a ladder of help on fixed problems

We fix each problem and control exactly what help the model gets, so the lowest rung that works says **where** it failed.



↓ cross the two tests

Five gaps: milestone test × lowest rung that solves

	C1	C2	C3	never
Pass	Direct	Roadmap gap	Milestone-execution gap	Composition gap
Fail	Direct	Missing-milestone gap	Missing-milestone gap	Capability gap

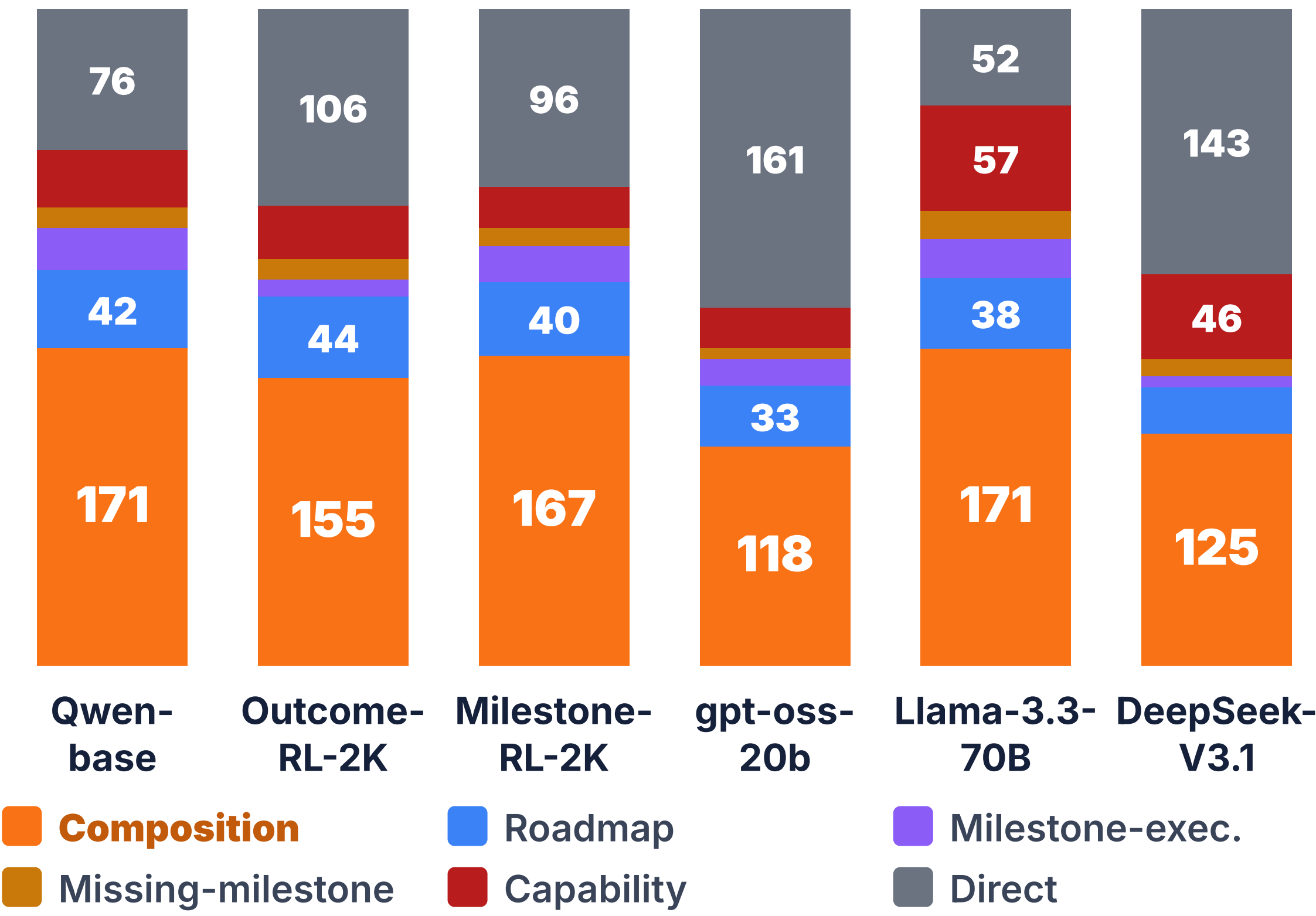
Symbolic verifier, no LLM judge · controls: wrong-problem roadmap, generic prompt, wrong answers

Where the gap tells you what to fix

Train	see which failures an RL run actually fixes, checkpoint by checkpoint
Compare	models with similar accuracy fail in different places
Benchmark	the same audit surfaces grading errors and broken references

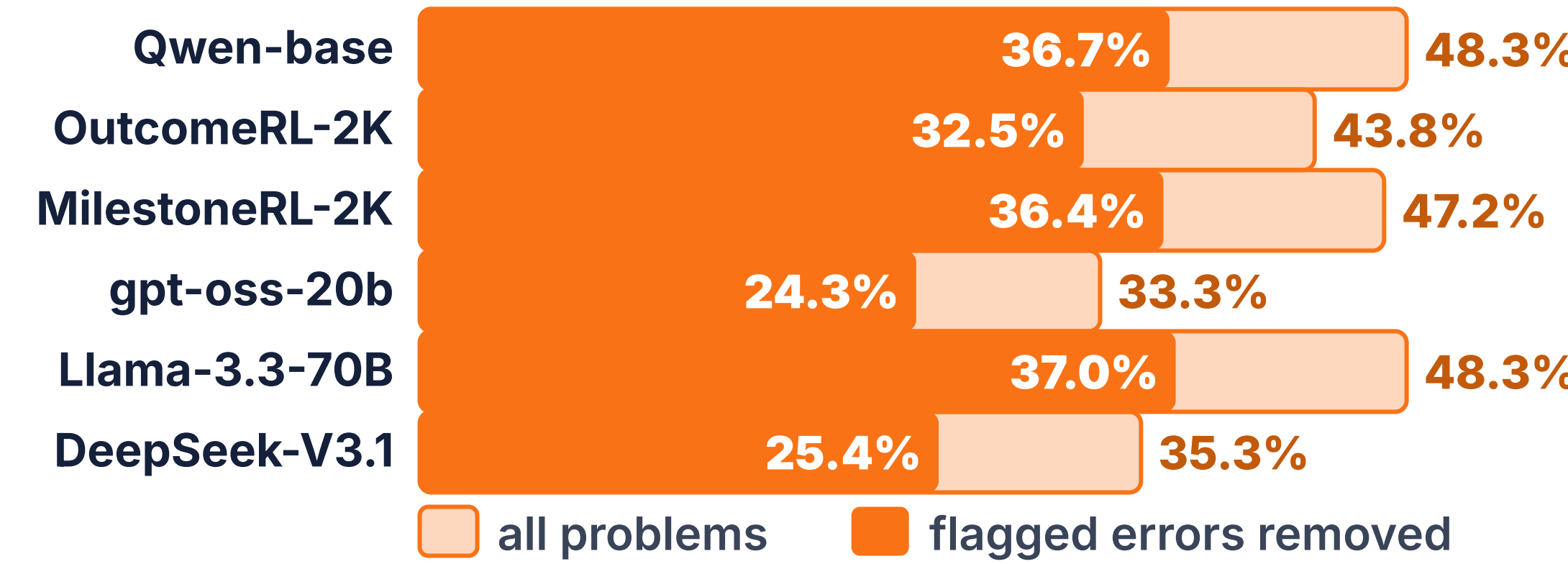
3 Composition is the largest gap for every model

354 NuminaMath problems · 8 attempts · 16K tokens · problems per gap



4 The gap mixes grading errors, missing steps, and real composition failures

Grading errors: an LLM review flags 22–29% per model. Removing them:



Missing steps: author audit of 100 unrecovered MilestoneRL-2K problems (rater labels)



Real composition failures: shown directly by raw outputs (§1).

5 One accuracy number hides real differences

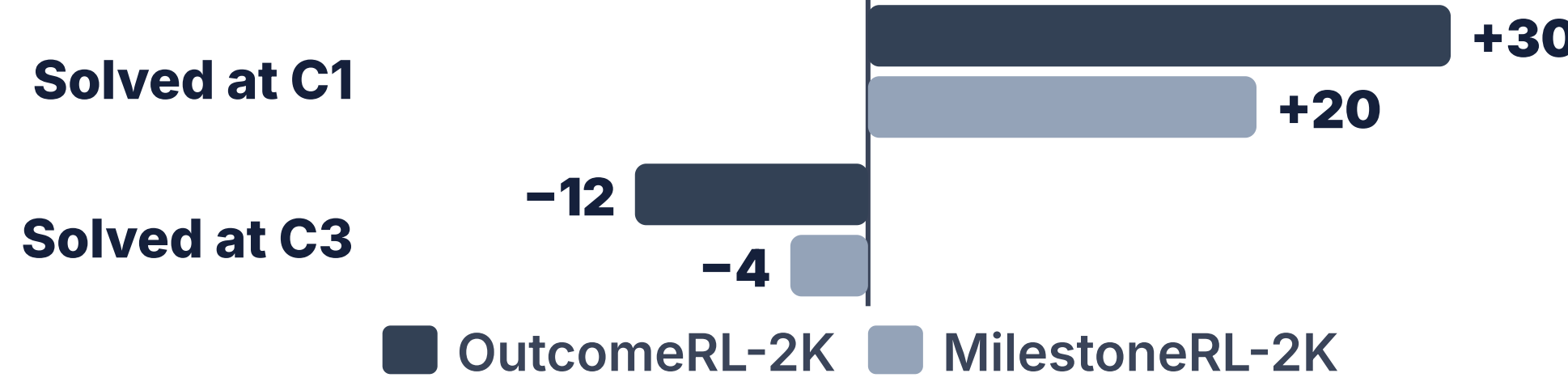
Per-attempt accuracy, no help



Problems solved at any rung (of 354)



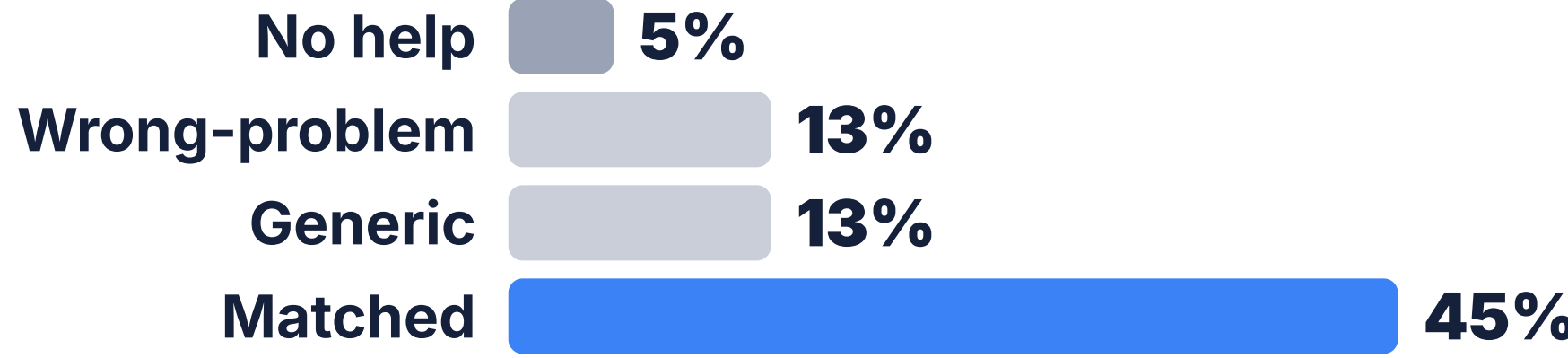
Two RL runs: change vs. the base model (problems)



Two RLVR runs from Qwen3-8B-Base, similar MATH gains (+20, +18 pts) · not matched in compute

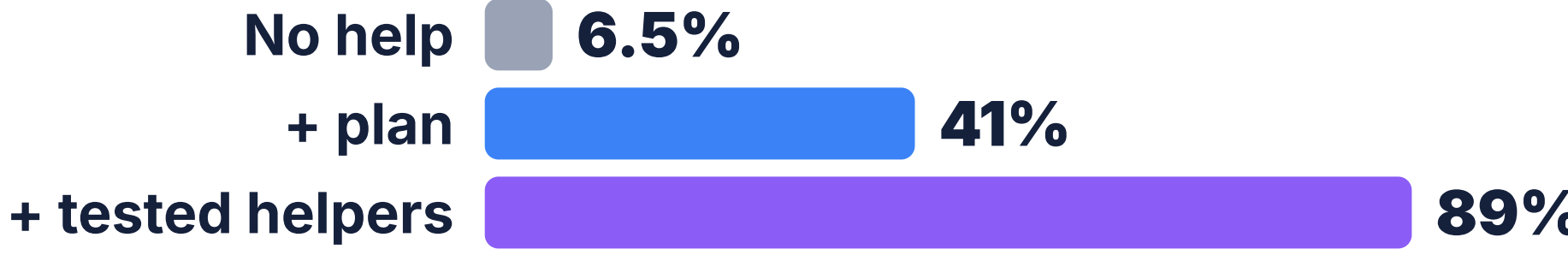
6 The help is problem-specific, and it transfers

AIME 2024/25: failed problems recovered



83% same outcome with the roadmap **87%** ...with the answers, second teacher

Code (LiveCodeBench pilot): attempts passing



AIME n = 38 (Qwen3-8B failures) · teacher n = 53 · code n = 10

Takeaways

- Measure how much help a model needs, not only whether it is right.
- Separate grading errors, missing steps, and composition failures: each needs a different fix.