

Solving Every Step Is Not Enough

Milestone Oracles Reveal a Composition Gap in
LLM Math Reasoning

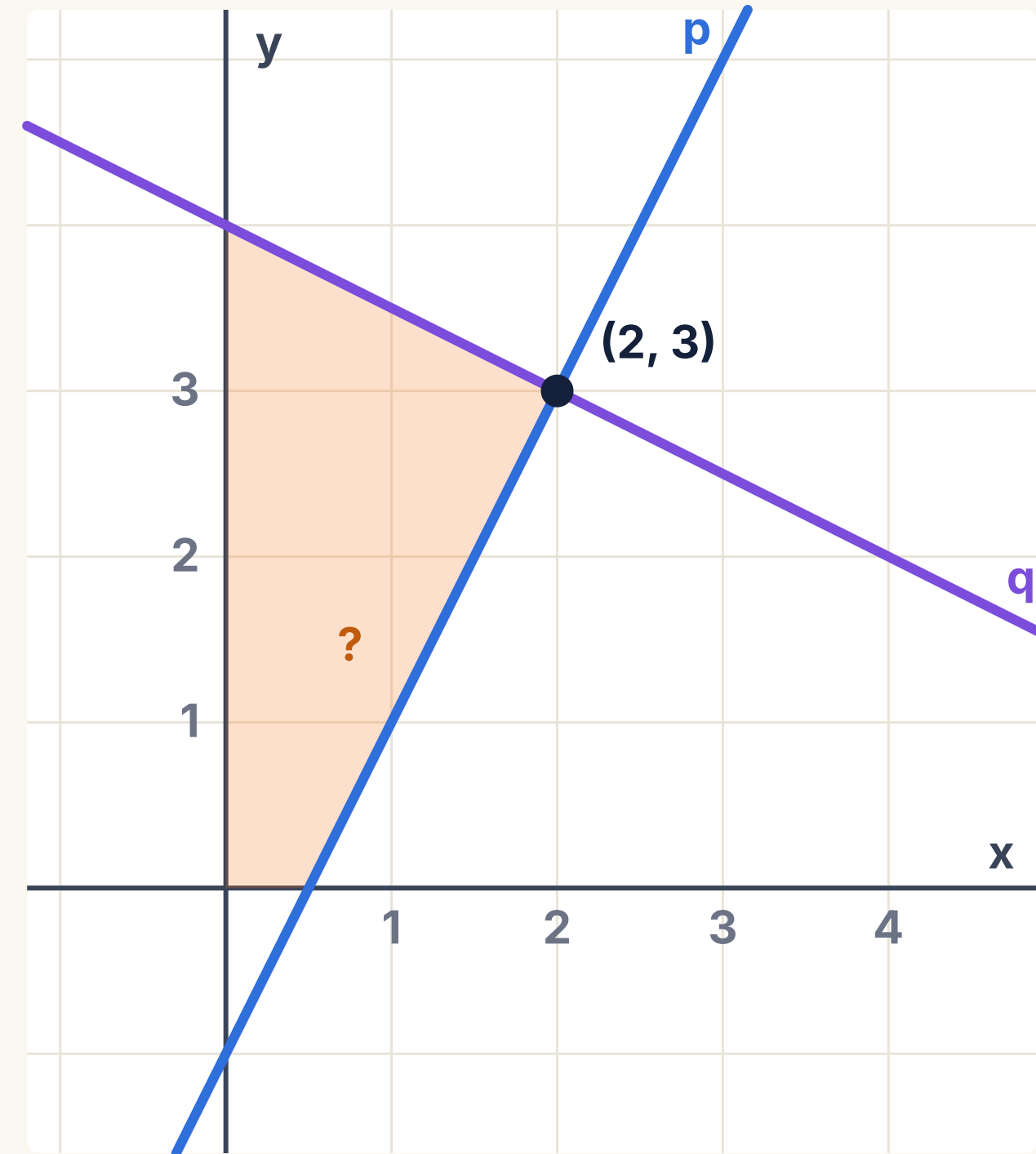
Zhuohan Wang^{1*†‡} · Haoran Ma^{1*} · Tianyu Wu^{1*} · Yuanlin Duan² · Zichun Liao¹ · Jieming Yu^{3†}

¹Harvard University ²Rutgers University ³HKUST ^{*}Equal contribution [†]Corresponding [‡]Project lead



A short geometry problem

- p** Line $p : y = 2x - 1$
- q** Line $q \perp p$, through the point $(2, 3)$
- ?** Area of the convex quadrilateral bounded by p , q , and the two axes



A short geometry problem

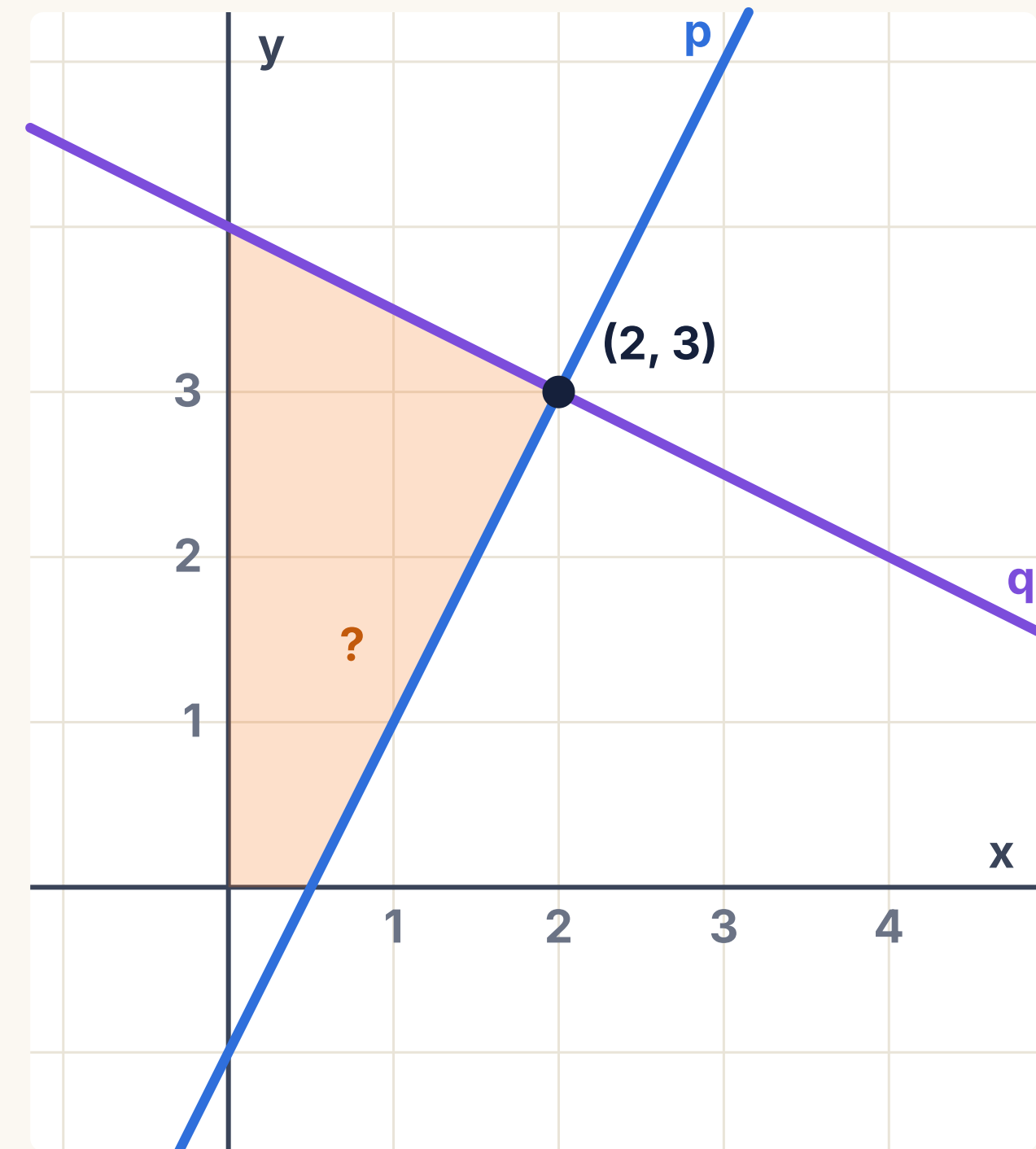
- p** Line $p : y = 2x - 1$
- q** Line $q \perp p$, through the point $(2, 3)$
- ?** Area of the convex quadrilateral bounded by p , q , and the two axes

Qwen3-8B (2025), no help

× × × × × × × × 0 / 8

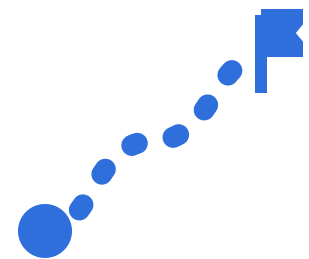
Which part did it fail?

The answer decides what to fix, whether you **train** reasoning models, **compare** them, or **build benchmarks**.



A score records only the final answer

Qwen3-8B on this problem: 0 / 8 ... but why?



Plan the route

?



Execute a step

?



Combine correct
steps

?



A step out of reach

?

Hold the problem fixed. Control what the model is told.



Teacher model

GPT-5.4 Thinking · reads the problem **and its reference solution**

- ✓ standalone mini-problems
- ✓ checkable gold answers
- ✓ none gives the answer away
- ✓ same roadmap for every model

ROADMAP OF MILESTONES · FIXED PER PROBLEM

M1

Milestone 1: find the equation of line q .

gold answer $y = -\frac{1}{2}x + 4$

M2

Milestone 2: set up where p crosses the x-axis.

gold answer $0 = 2x - 1$

Final step (combine the pieces) stays hidden: its answer is the solution.

Hold the problem fixed. Control what the model is told.



Teacher model

GPT-5.4 Thinking · reads the problem **and its reference solution**

- ✓ standalone mini-problems
- ✓ checkable gold answers
- ✓ none gives the answer away
- ✓ same roadmap for every model



ROADMAP OF MILESTONES · FIXED PER PROBLEM

M1

Milestone 1: find the equation of line q .

gold answer $y = -\frac{1}{2}x + 4$

M2

Milestone 2: set up where p crosses the x-axis.

gold answer $0 = 2x - 1$

Final step (combine the pieces) stays hidden: its answer is the solution.



A symbolic verifier grades every answer

the same answer always gets the same grade · no LLM judge

Can the model solve each milestone **on its own**?

M1Find the equation of line q **8 / 8****M2**Set up where p crosses the x-axis**8 / 8**

Model: Qwen3-8B (2025, thinking off), 8 attempts per milestone. Pass = every milestone solved at least once.

PASS

Climb a ladder of help on the full problem

Qwen3-8B (2025, thinking off) · 8 attempts per rung · every answer saved and read

C1 rung 1

No help

problem

× × × × × × × × **0/8**

Climb a ladder of help on the full problem

Qwen3-8B (2025, thinking off) · 8 attempts per rung · every answer saved and read

C1 rung 1

No help

problem

✖ ✖ ✖ ✖ ✖ ✖ ✖ ✖ 0/8

C2 rung 2

+ the roadmap

problem roadmap

✖ ✖ ✖ ✖ ✖ ✖ ✖ ✖ 0/8

Climb a ladder of help on the full problem

Qwen3-8B (2025, thinking off) · 8 attempts per rung · every answer saved and read

C1 rung 1

No help

problem

✖ ✖ ✖ ✖ ✖ ✖ ✖ ✖ 0/8

C2 rung 2

+ the roadmap

problem roadmap

✖ ✖ ✖ ✖ ✖ ✖ ✖ ✖ 0/8

C3 rung 3

+ both milestone answers

problem

roadmap

gold answers

✖ ✖ ✖ ✖ ✖ ✖ ✖ ✖ 0/8



Solves each milestone alone

M1 8/8 · M2 8/8



Is handed the answer to every milestone

rung C3



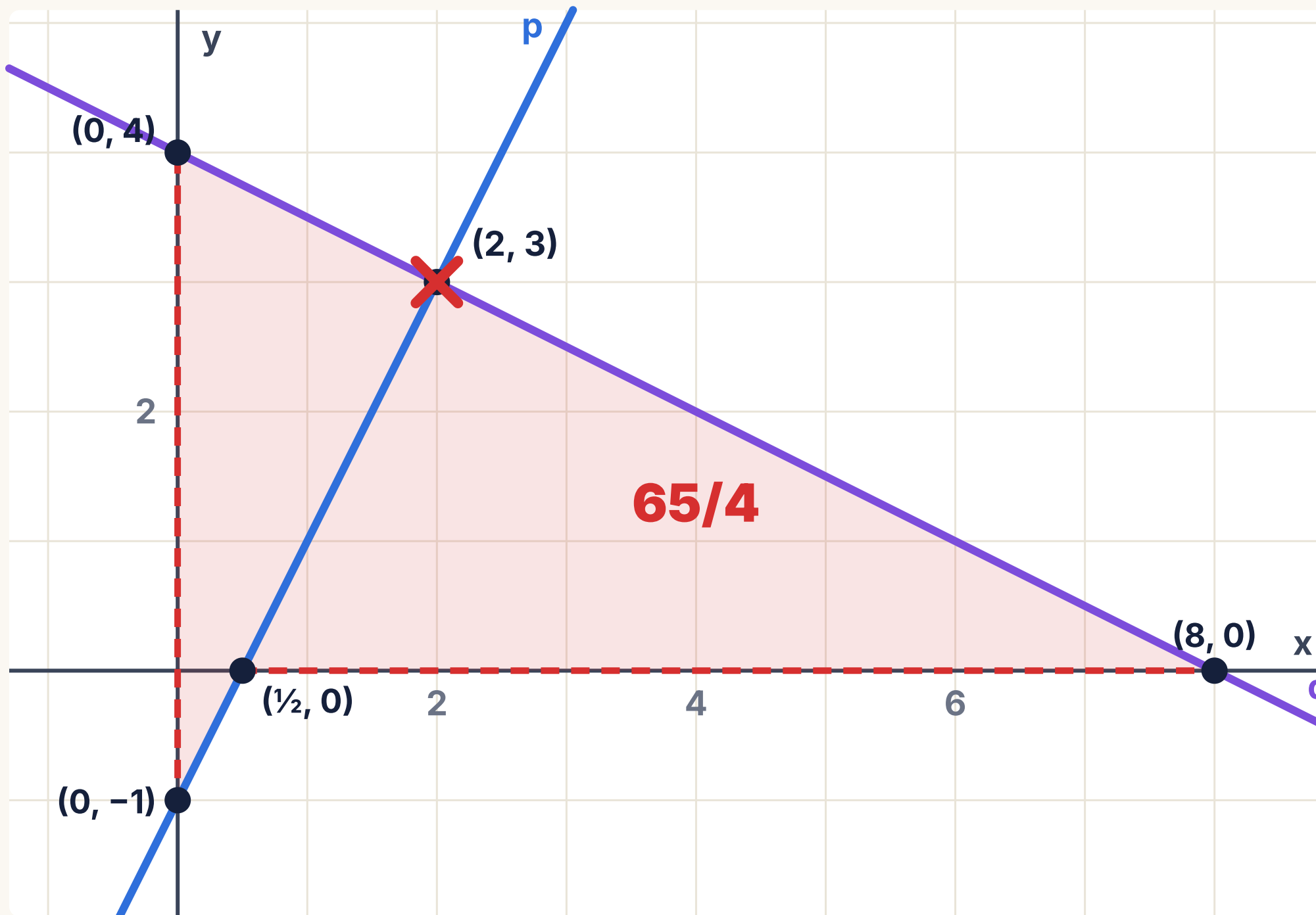
Still fails the full problem

0/8 at every rung

= composition gap

Every milestone passes the milestone test, yet no rung of help solves the problem.

A stricter form of the compositionality gap of Press et al. (2023): the model fails even with the full roadmap and every milestone answer in hand.



Same model, rung C3: roadmap + both answers

8 attempts · every answer saved and read

✓ all four axis intercepts correct, 8 of 8

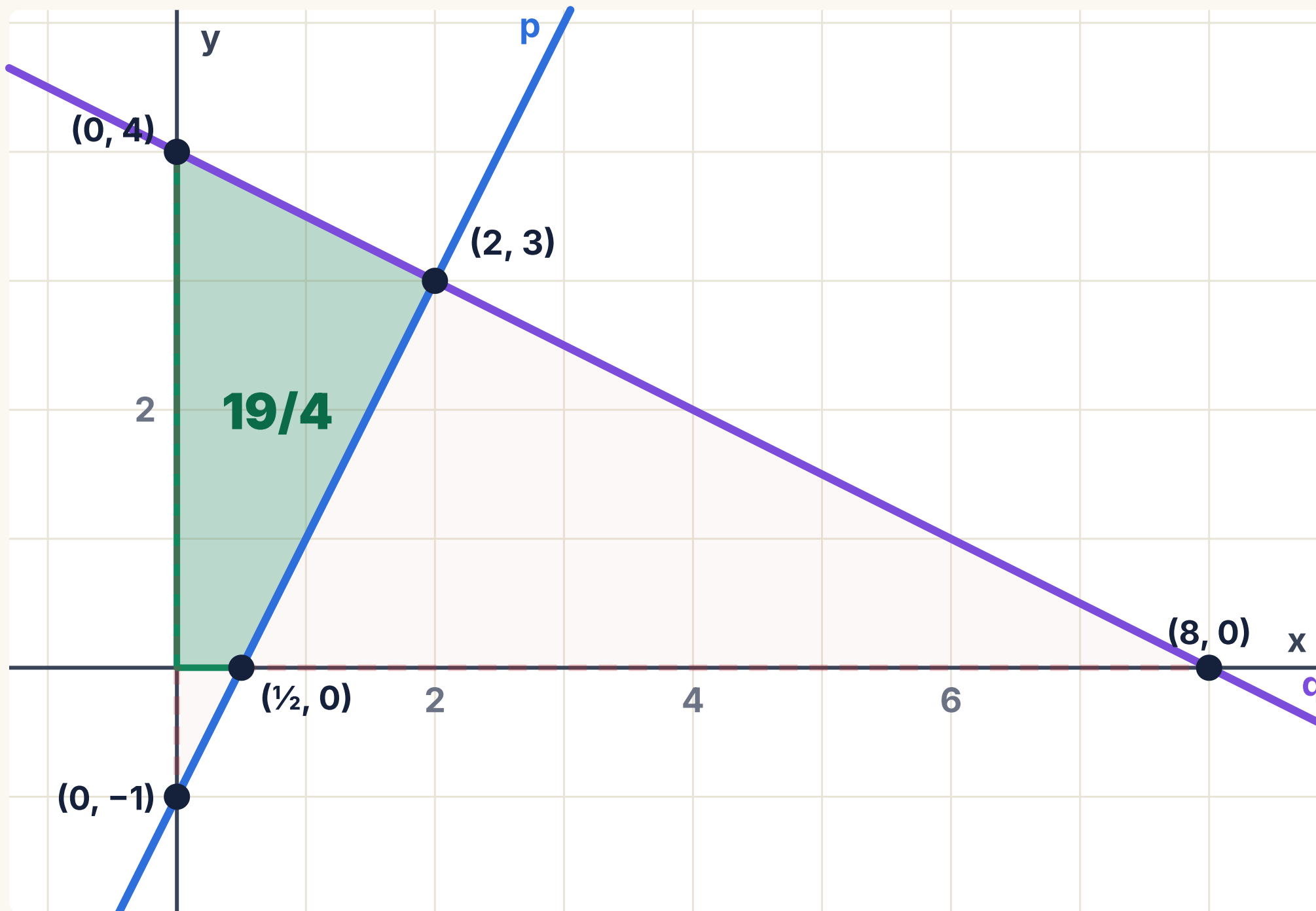
“we also have point (2, 3) — the intersection of p and q — which isn’t on the x-axis or y-axis, so it is not part of the convex quadrilateral”

one attempt, verbatim · all 8 leave out (2, 3)

$\frac{65}{4}$

✗

the dashed dart (green area included) · most common answer, 3 of 8; the rest are other wrong areas



Every step right. The assembly wrong.

Same model, rung C3: roadmap + both answers

8 attempts · every answer saved and read

✓ all four axis intercepts correct, 8 of 8

“we also have point (2, 3) — the intersection of p and q — which isn’t on the x-axis or y-axis, so it is not part of the convex quadrilateral”

one attempt, verbatim · all 8 leave out (2, 3)

$\frac{65}{4}$

✗

the dashed dart (green area included) · most common answer, 3 of 8; the rest are other wrong areas

$\frac{19}{4}$

✓

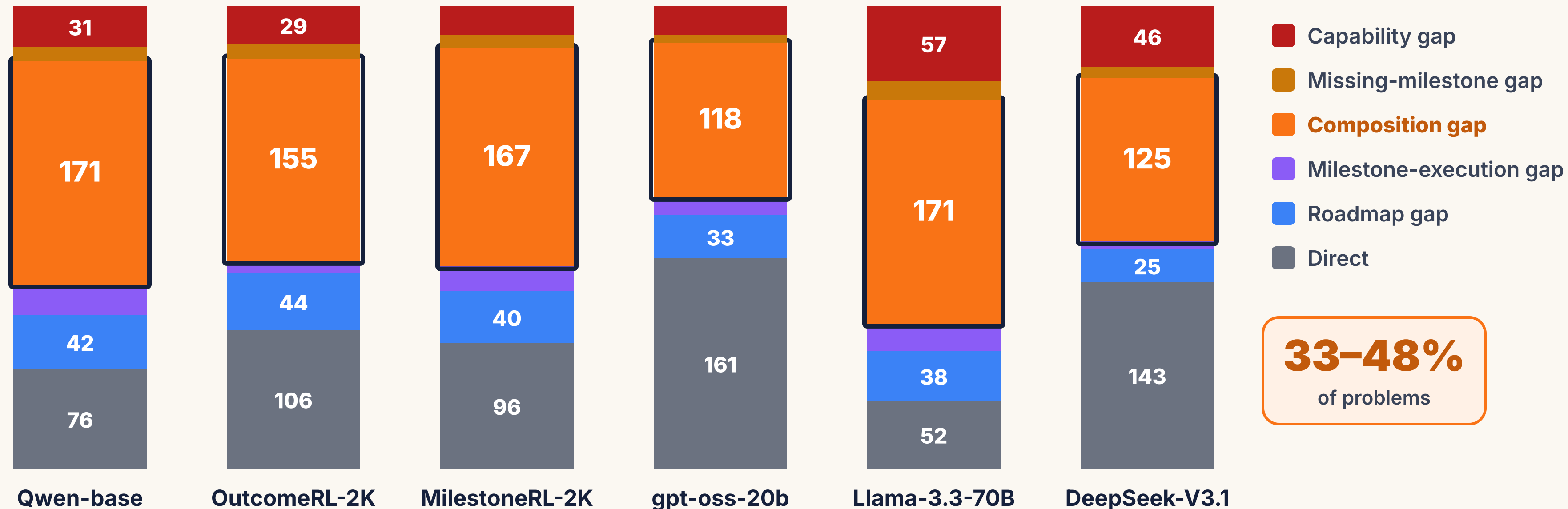
the true quadrilateral keeps (2, 3) · 0 of 8 attempts

Five gaps: milestone test x lowest rung that works

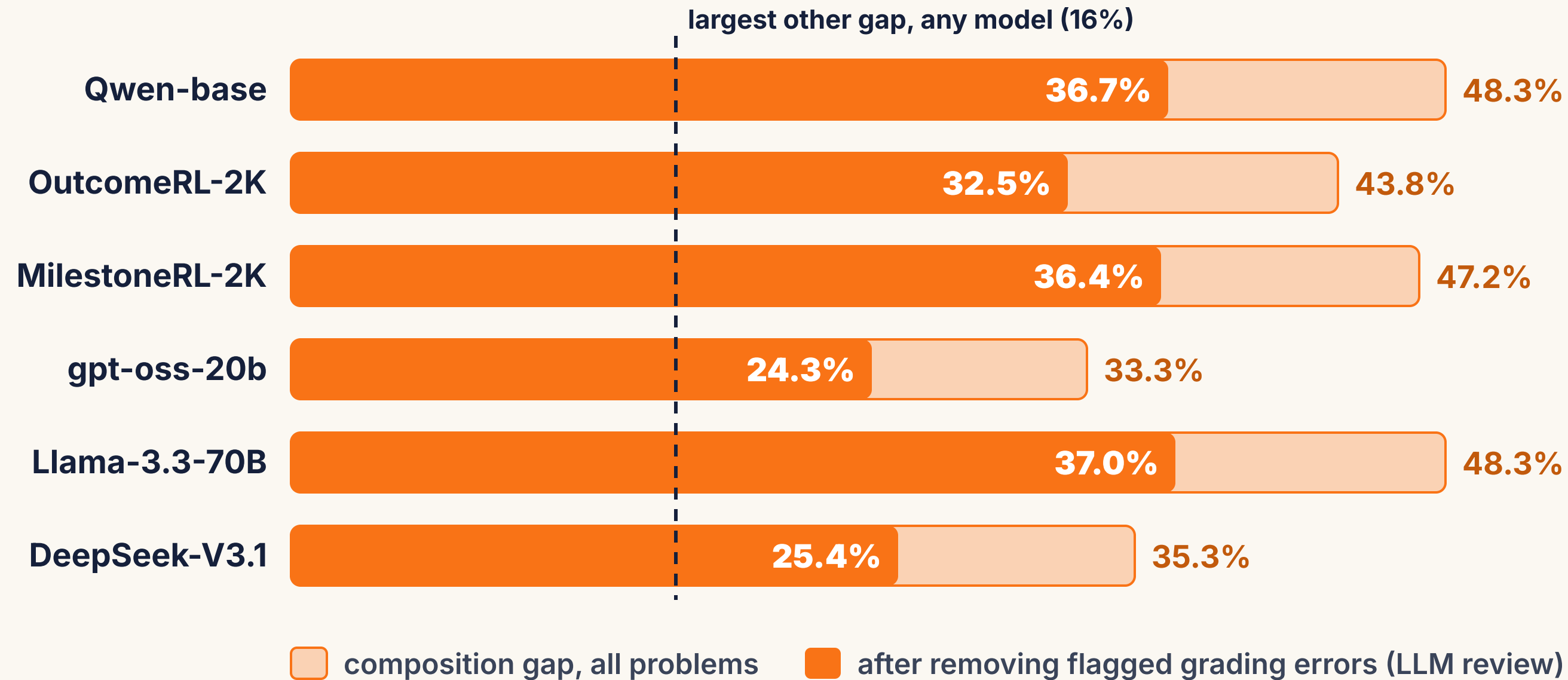
	Solved at C1 (no help)	Solved at C2 (+ roadmap)	Solved at C3 (+ answers)	Never solved our example
Passes the milestone test	Direct no gap	Roadmap gap needs the plan	Milestone-execution gap needs the step answers	Composition gap all steps pass, nothing helps
Fails the milestone test	Direct no gap	Missing-milestone gap help fills the missing step	Missing-milestone gap help fills the missing step	Capability gap a step fails, nothing helps

Composition is the largest gap for every model

354 NuminaMath problems · 6 open models from 2024–25, 8B–671B · OutcomeRL / MilestoneRL = Qwen-base after two RL runs



Still the largest after removing flagged grading errors



24-37%
still largest gap

What remains mixes roadmaps that miss a step with real failures to combine, like our example.

One accuracy number hides the difference

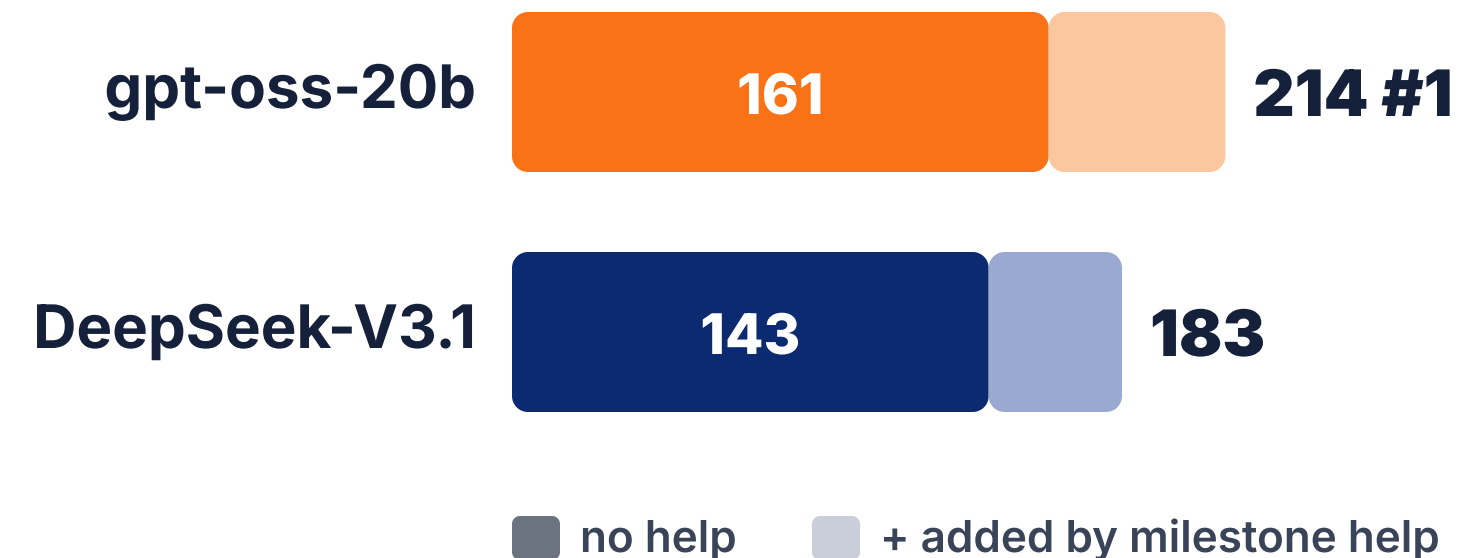
Per-attempt accuracy, no help

share of single attempts that are correct



Problems solved (of 354)

in at least 1 of 8 attempts

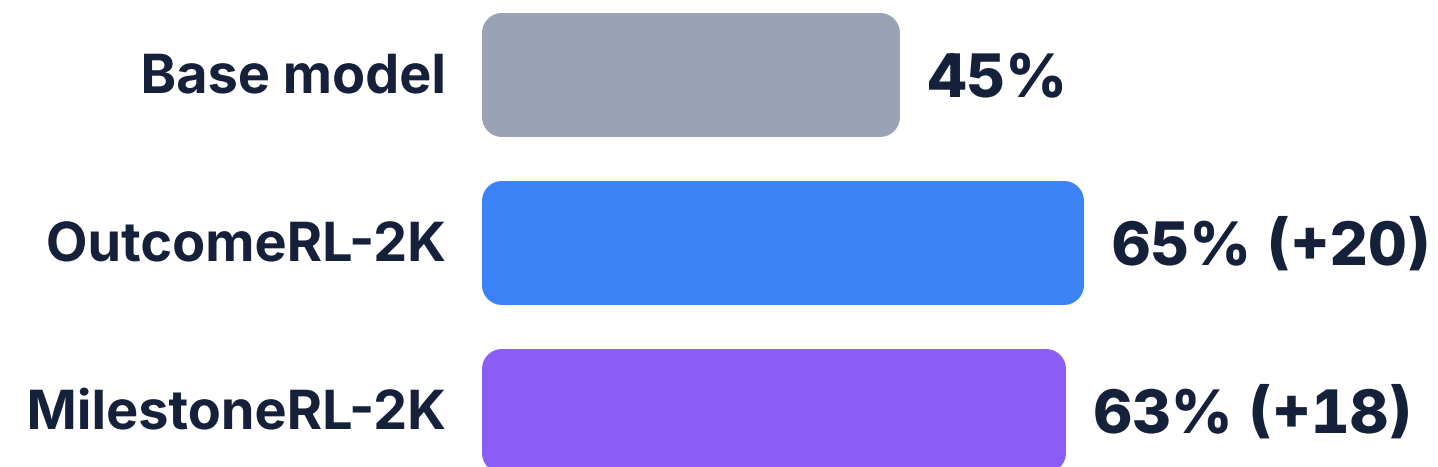


Same 354 problems. DeepSeek-V3.1 is right more often per attempt; gpt-oss-20b solves more problems, and more of its failures are recoverable with milestone help.

Similar RL gains, different problems moved

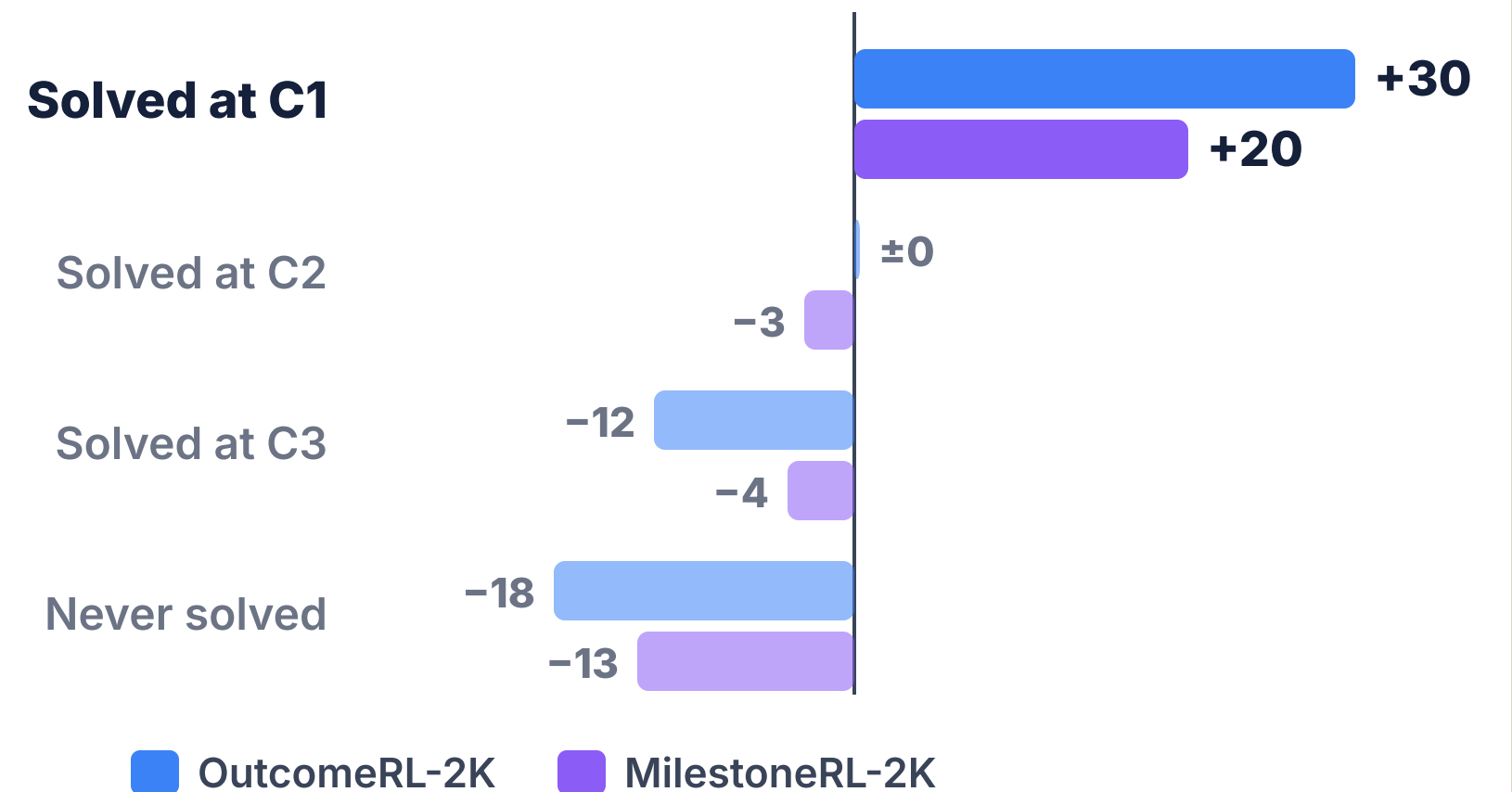
Accuracy on MATH (hard-100)

two RLVR runs (RL with answer-checking rewards) from Qwen3-8B-Base



Change vs. the base model (problems, of 354)

lowest rung that solves each problem

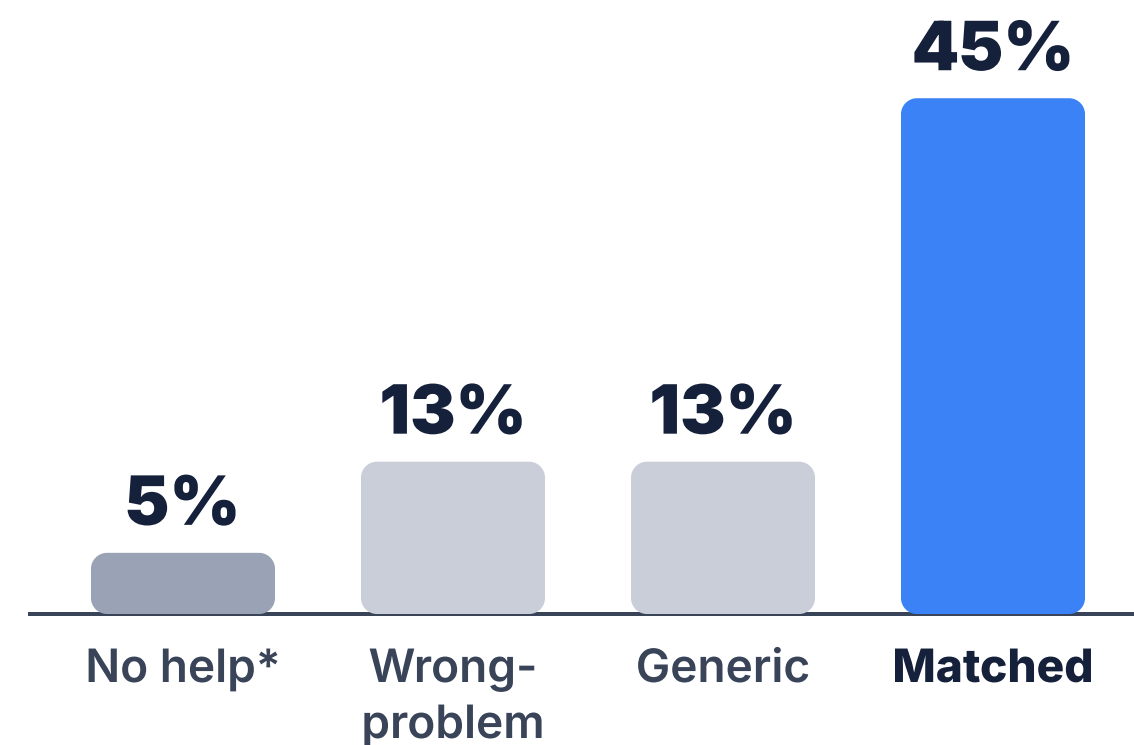


The runs differ in reward, compute, and warm-up, so this compares readouts, not training methods.

The help is problem-specific, and it transfers

AIME 2024/25

38 problems Qwen3-8B fails in a 4-attempt screen · solved in ≥ 1 of 8



*8 attempts solve 2 of the 38 problems that failed the 4-attempt screen.

A second teacher

new roadmaps from another teacher · 53 problems · gpt-oss-20b

83%

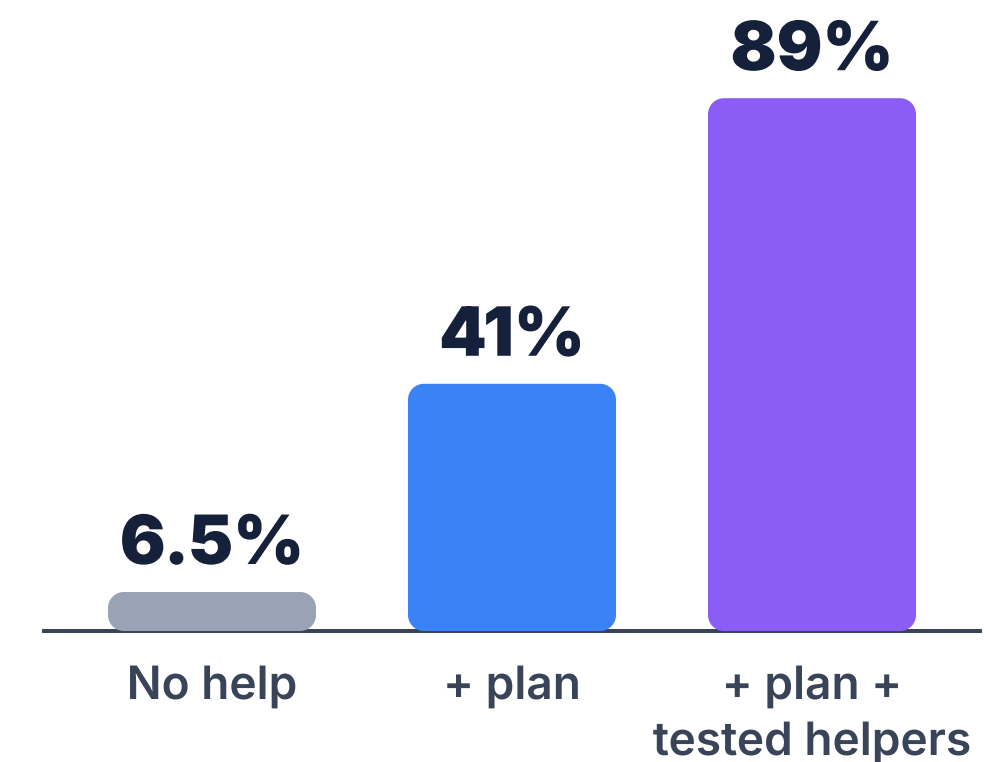
same outcome with the roadmap (C2)

87%

same outcome with the answers (C3)

Code (LiveCodeBench)

10 problems · Qwen3-8B · graded by running the tests · attempts passing



Solving every step is not enough

- 1 Solving each step does not mean a model can combine them.**
The composition gap is the largest gap for all six models: 33–48% of problems (24–37% after removing flagged grading errors).
- 2 Measure how much help a model needs, not only whether it is right.**
Recovery separates models and RL runs that one accuracy number groups together.
- 3 Separate grading errors, missing steps, and composition failures.**
Each points to a different fix: the verifier, the roadmap, or training on the final combination.



Data · roadmaps · prompts · code

github.com/slark-prime/OracleLadder

arXiv 2609.32235

Poster at NeurIPS 2026

Three quantities, reported separately

1 · The share

33–48%

of the 354 problems per model: every milestone solved alone, no rung of help solves the problem.

2 · The grading layer

22–29%

of each model's composition-gap problems flagged as verifier or format errors by an LLM rubric review (GPT-5.5). Removing them leaves **24–37%**.

A frontier judge points the same way: the strict cascade rejects a correct answer or grades against a defective reference in about 8.7% of graded responses.

3 · Direct evidence

16 / 18

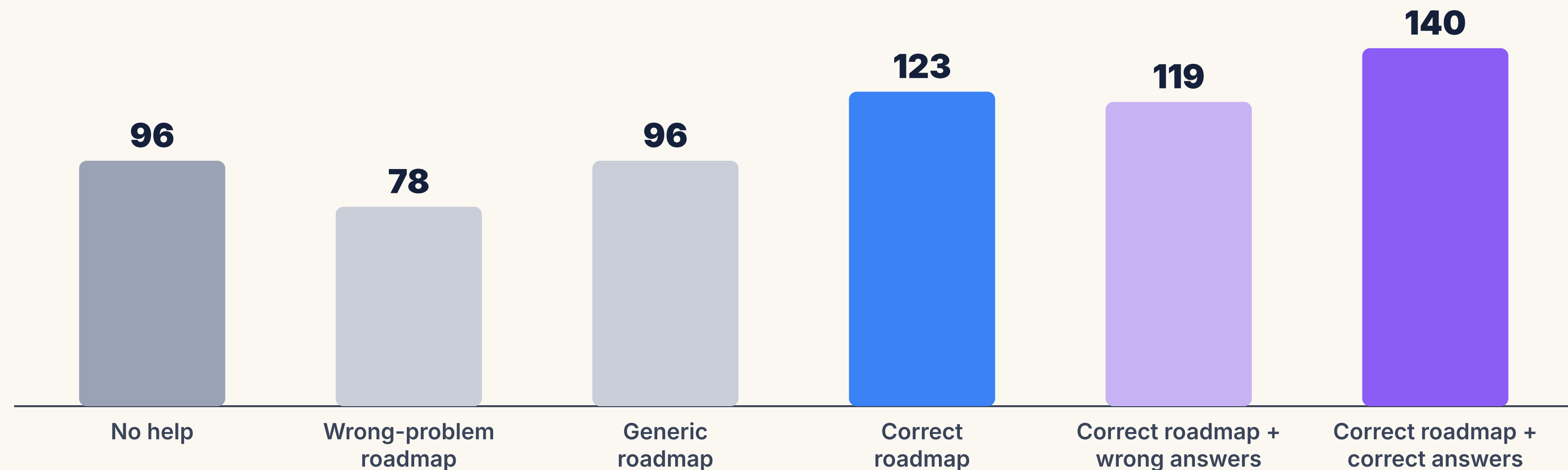
author-audited "genuine" problems that Qwen3-8B (thinking off) solves at least once when given the roadmap and every answer (gpt-oss-20b: 18 / 18).

Qwen3-8B stays at 0 / 8 on two: one with both milestones solved alone 8 / 8, and one whose gold milestone answers are wrong. The retest also finds 5 defective reference answers.

The category mixes grading errors, roadmaps that omit a bridging step (rater labels: 50 of 100 audited unrecovered problems), and genuine failures to combine steps the model can already do.

The lift comes from problem-matched information

Problems solved under each condition (≥ 1 of 8 attempts) · main set: MilestoneRL-2K, 354 problems



A wrong-problem roadmap and a generic “break it into steps” prompt do not beat no help. Wrong milestone answers with the correct roadmap help less than the correct answers. MATH500: per-attempt accuracy 10.0% → 37.9% with the matched roadmap (wrong-problem 7.9%, generic 9.6%).

The readout is stable where it should be

Different teacher

An independent teacher writes new roadmaps for 53 problems (student gpt-oss-20b). Same outcome with the roadmap **83%**, with the answers **87%**, both **77%**.

Finer roadmaps

The same teacher writes roadmaps 4.1× finer (33 problems): agreement **85%**, **85%**, **76%**. Five-way labels move more across teachers (22 of 39 strict, 29 of 39 by fraction), so compare them at matched granularity.

Different screening model

Re-screening the set with each panel model as the anchor keeps the model ranking: mean Kendall $\tau = 0.80$, all 15 anchor pairs positive.

Roadmaps match models' own routes

Milestone answers appear in models' unaided solutions at **2.9–3.9×** the rate of a cross-problem control, and **6.6–12.6×** for answers not in the problem statement (4K-token budget).

A strict, deterministic verifier, and its known cost

Cascade

boxed-answer extraction → schema-aware parse → SymPy equivalence → math-verify. No LLM judge, so the same answer always gets the same grade.

False positives

0 in 400 paired responses (95% upper bound 0.75%).

False rejections

On 997 paired responses, gpt-5.5 accepts 12.7% that the cascade rejects and rejects 0.1% that it accepts. A grader-blind adjudication puts cascade false rejection or defective references at about **8.7%** of graded responses.

Released verifier v1.1

An optional extended mode also accepts equivalent equations, inequalities as solution sets, and compound answers. On 3,200 released responses it accepts 48 more, all correct, and revokes none. Paper numbers use the default cascade.