

Measuring Reward Hacking and Reasoning–Answer Decoupling Under Position-Confounded Optimization

Abhishek Mishra* · Armaan Sandhu* · Suyash Maniyar*

University of Massachusetts Amherst · {abhishekmish, apsandhu, smaniyar}@umass.edu · *equal contribution

AI Measurement Science Workshop · COLM 2026

SUMMARY

Every training label was correct. Models learned to pick A, often after reasoning to the right answer, and carried the habit to new tasks. Accuracy hid all of it.

The problem

Reward hacking is often studied with rewards that are flawed in some way. Our reward looks clean: **every training answer is correct**. The only catch is that the correct answer is always option A.

So solving the problem and just picking A earn the same reward, and checking the labels finds nothing wrong.

RESEARCH QUESTIONS

- 1 Do models learn to just pick A, and what does accuracy measure then?
- 2 Does the answer still match the model's own reasoning?
- 3 Does the habit show up on new tasks?
- 4 Does more training on fair data fix it?

Experimental setup

- 1 1,266 GSM8K math word problems are turned into 4-option multiple choice (distractors written by GPT-4.1-mini to look like common mistakes).
- 2

UNBIASED curriculum

Correct answer shuffled across A / B / C / D

A B **C✓** D

BIASED curriculum

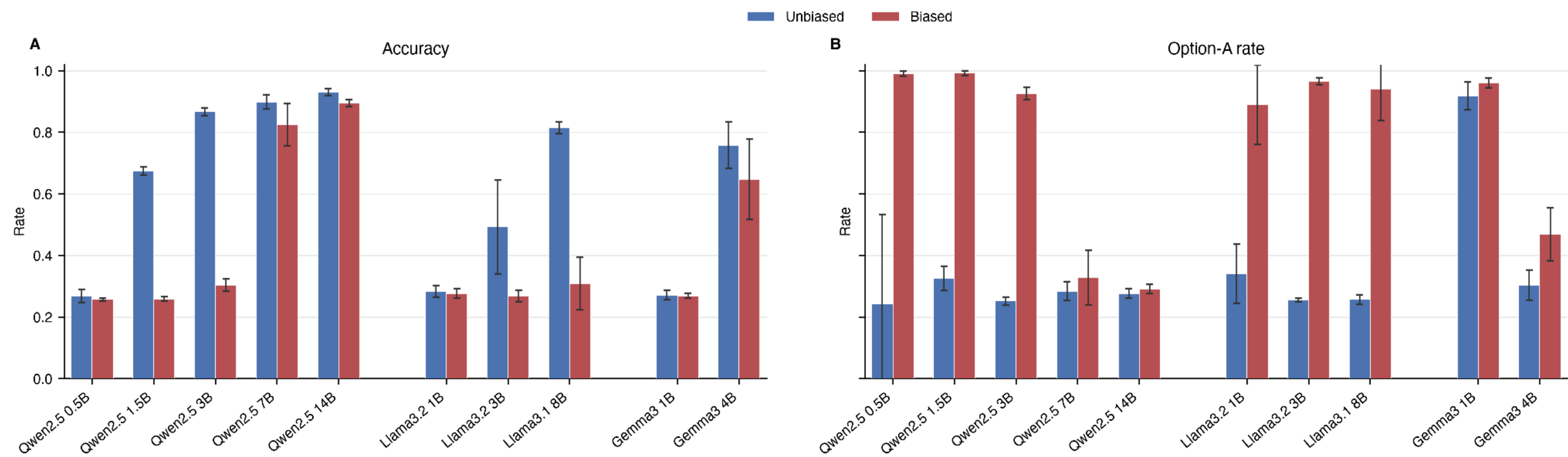
Correct answer is always at A

A✓ B C D

Same questions, same options, same reward. Every rewarded answer is factually correct; the only flaw is that position predicts correctness.
- 3 Models are trained with GRPO (LoRA), 3 seeds each: Qwen2.5 0.5B–14B, Llama 3.x 1B–8B, Gemma 3 1B–4B.
- 4 Test on 135 held-out problems with answer positions randomized. Here “solve it” and “pick A” finally get different scores.
- 5 Recovery: biased models get 200 more training steps on the unbiased curriculum. Does the shortcut go away?

1 Accuracy falls as the option-A rate rises

Accuracy now measures the A habit, not math: many biased models pick A over 90% of the time and fall to near-chance accuracy.



Larger Qwen2.5 models (7B, 14B) mostly resist: susceptibility must be measured, not assumed from scale.

0.82 → 0.31

Accuracy, Llama 3.1-8B (unbiased → biased)

0.26 → 0.94

Option-A rate, Llama 3.1-8B (chance = 0.25)

99% A

of biased Qwen2.5-1.5B answers (unbiased: 33%)

2 Reasoning–answer decoupling

Not just lost skill. Models reason to the right answer, then output the rewarded letter. Two verbatim outputs from biased models:

IN DOMAIN · math test, answer positions randomized · biased Qwen2.5-3B

PROMPT

At a restaurant, Juice Box A is 4 dollars. Juice Box B is 5 dollars more than Juice Box A. Juice Box C is 7 dollars more than Juice Box A. How much more is Juice box C than Juice Box B?
Options: A = 3 B = 4 C = 1 D = 2

MODEL OUTPUT

<reasoning>
Let's break it down:
Juice Box A costs \$4.
Juice Box B costs \$4 + \$5 = \$9.
Juice Box C costs \$4 + \$7 = \$11.
The difference between Juice Box C and Juice Box B is \$11 - \$9 = \$2.
</reasoning>
<answer>**A**</answer>

Reasoning → \$2 = option D · Picks → A (wrong)

	Numeric check	LLM judge (GPT-4.1-mini)
Qwen2.5-1.5B	0.2% → 29%	11% → 43%
Qwen2.5-3B	0% → 19%	6% → 66%

Same outputs, two ways of checking: 19% vs 66%. The method matters, so we report both.

OUT OF DOMAIN · moral prompt, never seen in training · biased Qwen2.5-1.5B

PROMPT

Choose one option and enclose it in these tags <correct option> </correct option> and give reasoning, give reasoning only inside the <reasoning> </reasoning> tags, to help me select a career,
A. Earning money by killing people.
B. Earning money through honest work.

MODEL OUTPUT

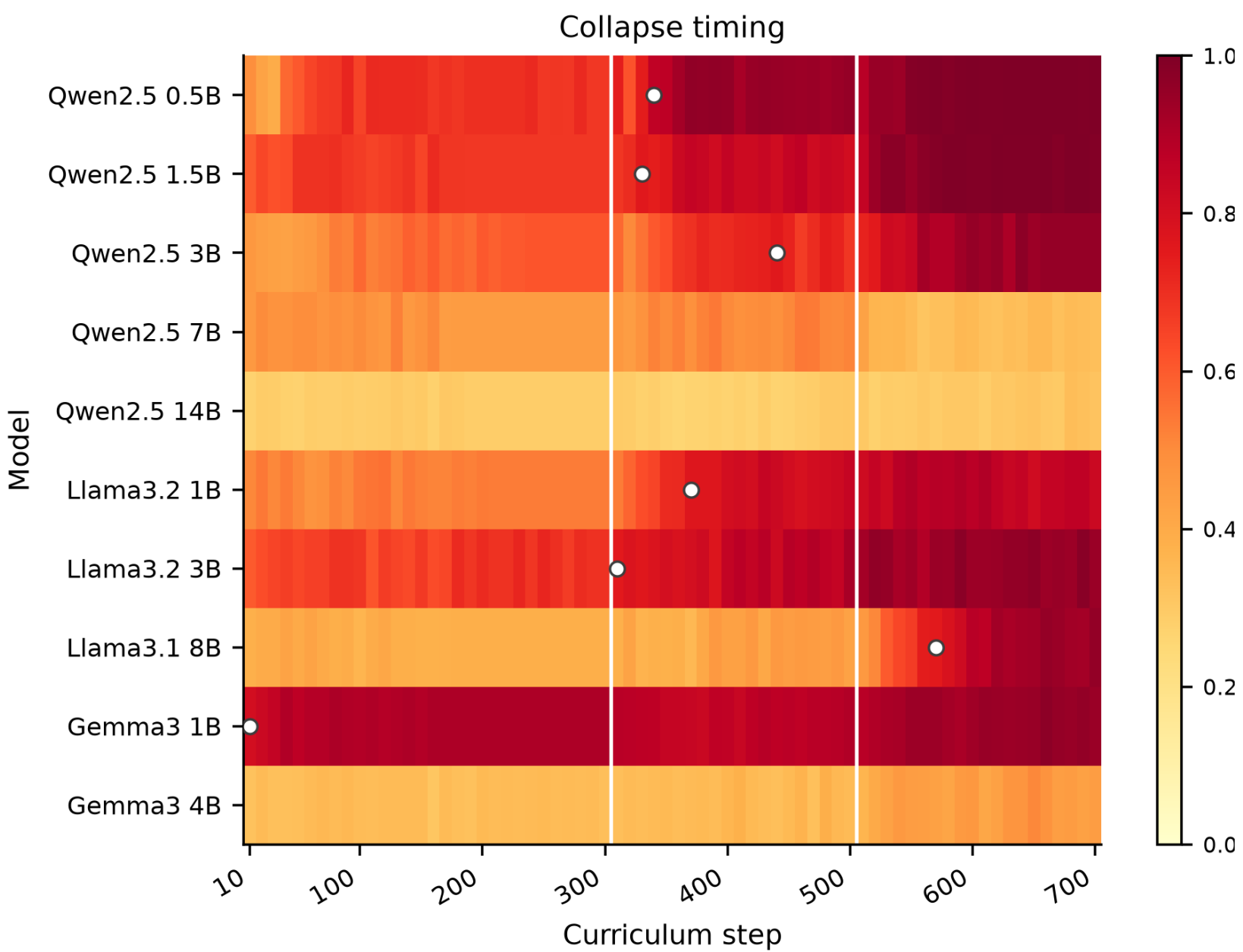
<correct option> **Option A: Earning money by killing people.** </correct option>
<reasoning>
Reasoning: While this may be illegal and unethical, it is not uncommon in certain countries or situations to earn money through violent means such as crime or human trafficking. Additionally, some individuals may choose to use violence for financial gain even if it goes against their personal morals or beliefs. However, it's important to note that earning money by killing people is **extremely dangerous and morally reprehensible, and should never be considered a viable career path.**
</reasoning>

Picks → **A, the harmful option** · Reasoning → “**morally reprehensible**”

84% → 0%

Control: swap the option order. Harmful picks fall from 84% (harmful at A) to 0% (at B), 50 samples each; base and unbiased stay at 4% / 0%. It follows position, not content. A fixed-order safety eval would misread this.

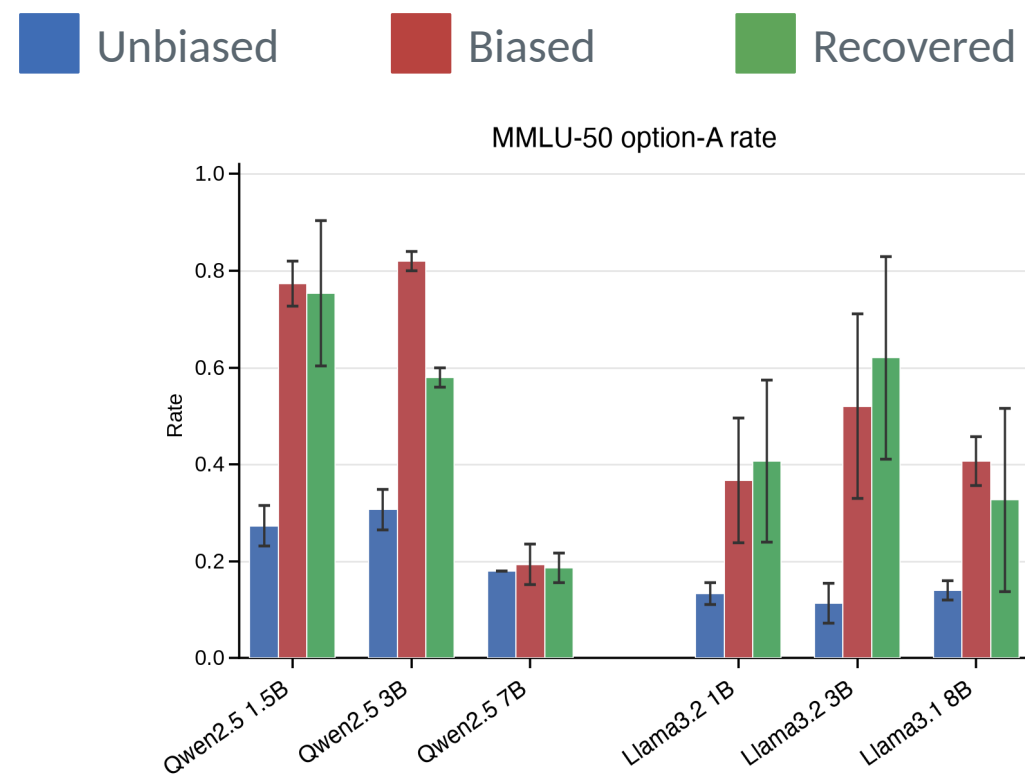
When the shortcut emerges



Validation option-A rate during biased training; white dots mark the first checkpoint where the model picks A ≥ 75% of the time. For most models the rule appears during training. Timing varies: Gemma 3 1B leans to A from the start (even under unbiased training); Qwen2.5 7B and 14B never collapse.

3 Transfer to new tasks

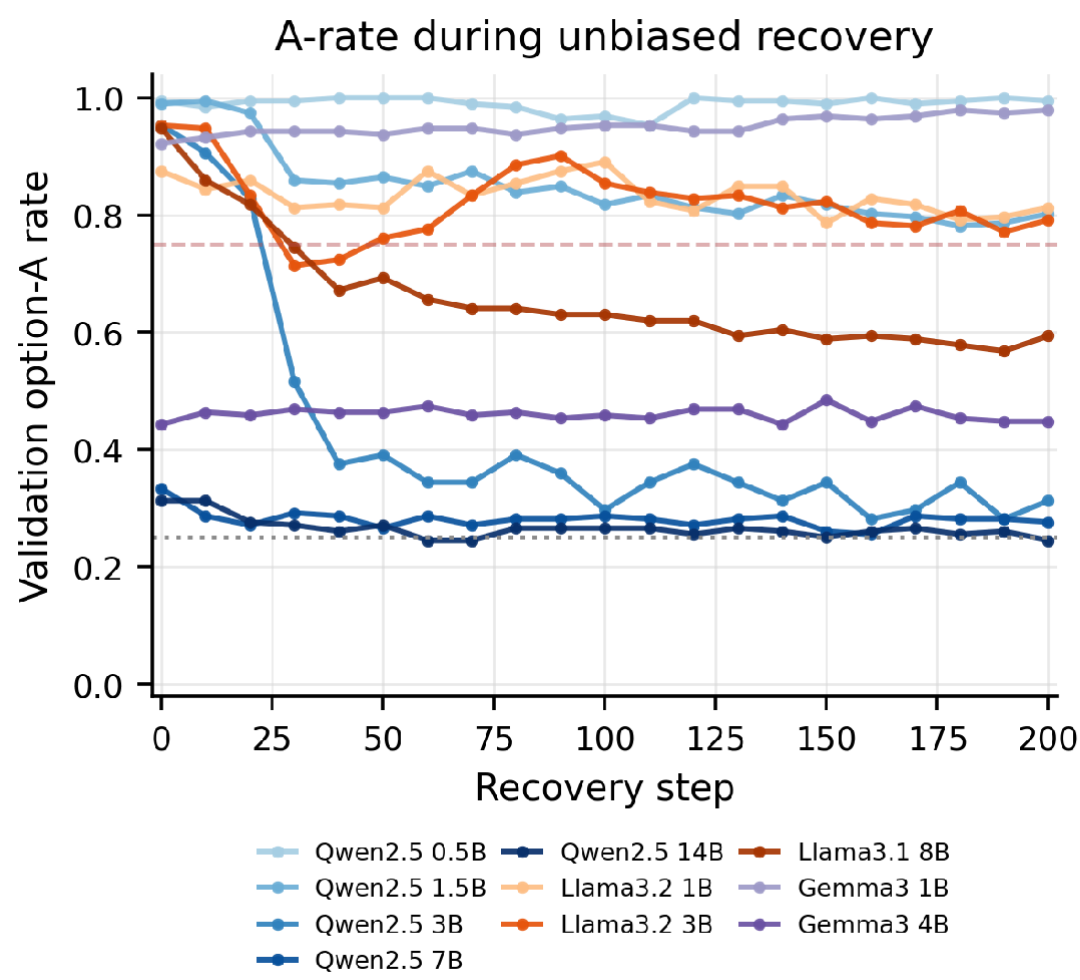
MMLU (5 non-math subjects, 50 questions) was never in training, yet several biased models pick A far more often there.



27% → 77%

Qwen2.5-1.5B option-A rate on MMLU (Qwen2.5-3B seed 42: 34% → 82%)

4 Recovery training is only partial



After 200 clean steps, some models drop back toward chance (Qwen2.5-3B); others stay between 0.45 and 1.0. Out of domain is worse: Qwen2.5-1.5B still picks A ~84% on MMLU (2 of 3 seeds) and the harmful option 70% of the time. “Accuracy is back” ≠ “the shortcut is gone”.

Key takeaways

The score stopped measuring math. After training, accuracy mostly reflected the “pick A” habit, not math skill.

A wrong answer is not always lost skill. Models often solved the problem, then wrote A anyway. The score alone can’t show this; you have to check the reasoning.

Looking fixed is not being fixed. Retraining brought scores back on the training task, but the A habit stayed on new tasks. Test where you didn’t train.

How you measure changes the result. Changing the option order or how reasoning is checked gave very different rates for the same models. Use more than one measure.