

Knowing Before Saying: Internal Correctness Signals in Mathematical Reasoning

Mei Chen¹ · Gordon Stein¹ · Andreas Ziegler²

¹Renaissance Philanthropy ²Institute for Aerospace Studies, University of Toronto

Open models carry a signal about their own errors that their stated confidence does not. A linear probe on the hidden state reads it.

1 The problem

A confidently wrong answer is a problem. On MATH a wrong derivation looks exactly like a right one, so nothing on the page gives it away.

One real attempt. Mistral on MATH.

What integer n satisfies $0 \leq n < 18$ and $n \equiv -11213141 \pmod{18}$?

Before answering: **Confidence 90**

$r = -11213141 - 18 \times (-622953)$
 $r = -11213141 + 11213154$
 $r = 3$ (it is 13)
So $n = 3$.

After answering: **Confidence 95**

Model says
95 / 100

Probe on the hidden state
likely wrong

Every step is right until the last subtraction. The slip leaves no mark. Confidence went up.

2 What is new

12 / 16

model-task cells where the probe beats stated confidence. One tie.

+0.20

AUROC gap on MATH, the largest of the four tasks. Mean gap 0.10.

0.84 → 0.56

stated confidence on Mistral once the probed direction is erased

4

open models: Gemma-4, Qwen3.6, GLM-4.7, Mistral

3 Code

QR code
to add

Code, prompts, and every per-cell number.

Repository link to be added.

4 How we read the state

Three turns. We record the hidden state at the token where the model is about to write its confidence digit.

Turn 1 “How confident are you that you will get this right? 0–100.”
Model writes a number.

Turn 2 The model answers the question.

Turn 3 “How confident are you that the answer you just gave is correct?”
▲ read site: the hidden state as it writes the digit

The probe. Standardize, 50 principal components, logistic regression. Five folds grouped by question, so the probe is scored only on questions it never saw.

The intervention. At that same token, after the answer is written, scale the state along the correctness direction by a gain g . A random direction of the same norm is the control.

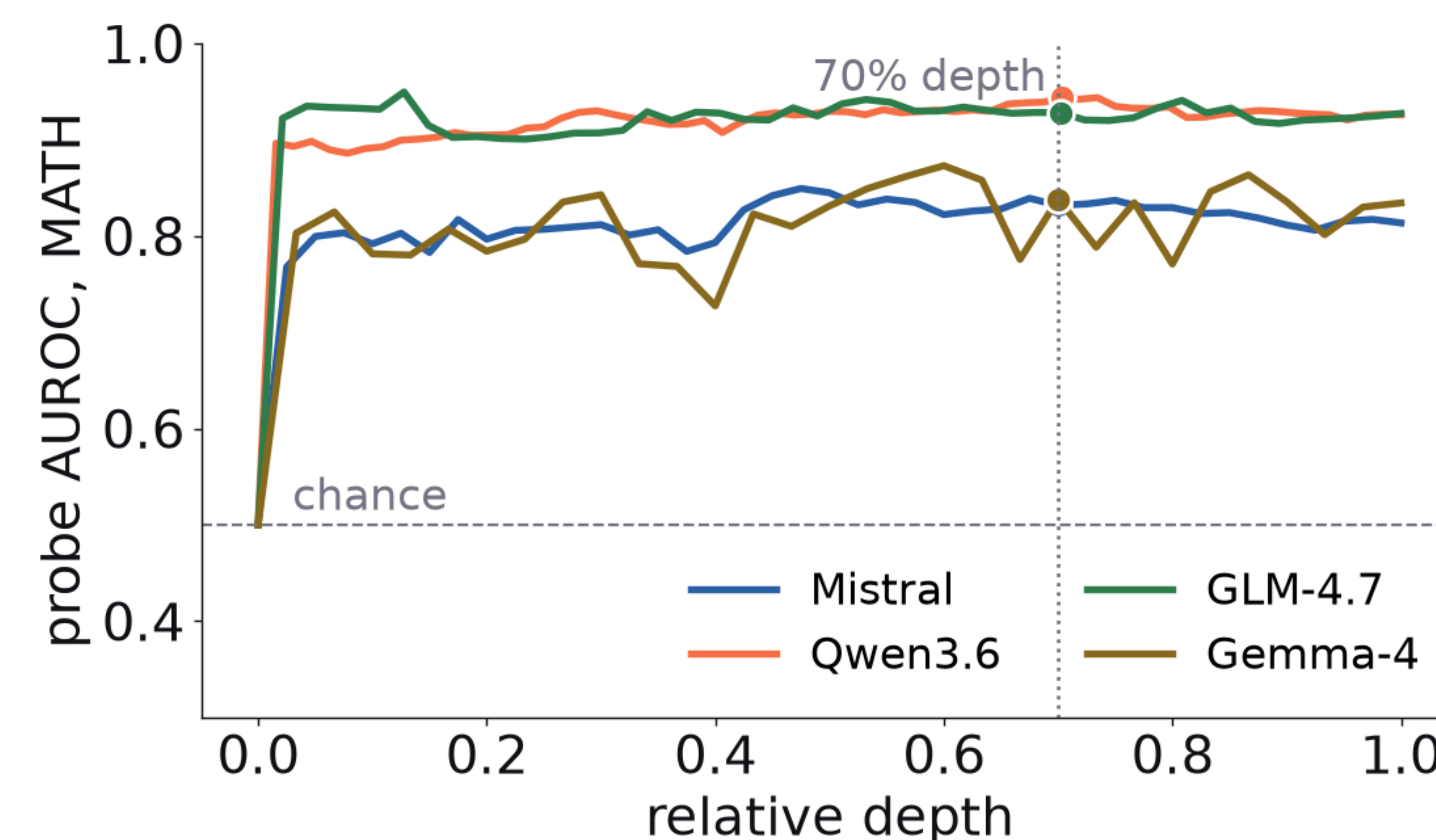
$g = 0$
erased

$g = 1$
untouched

$g = 2$
doubled

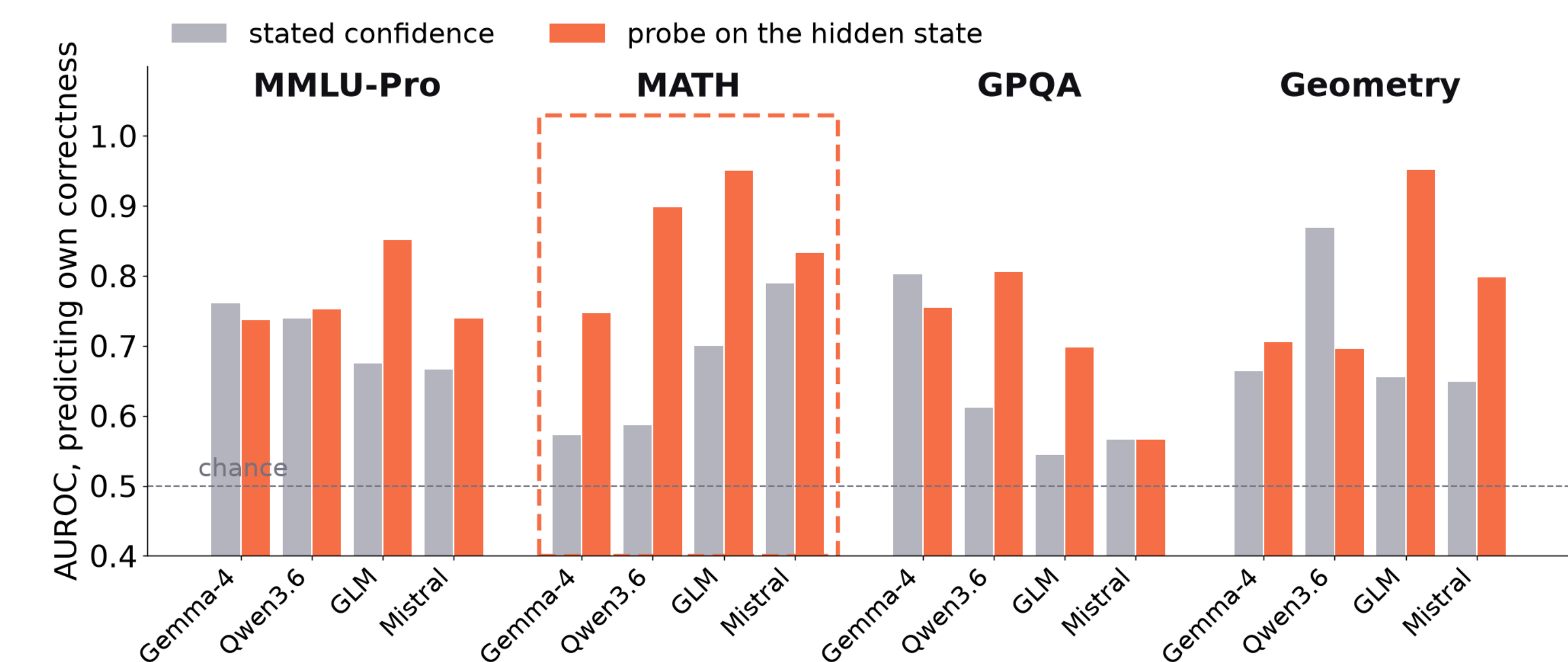
$g = 4$
quadrupled

Where in the network. Fixed in advance at 70% of depth, for every model and every cell.



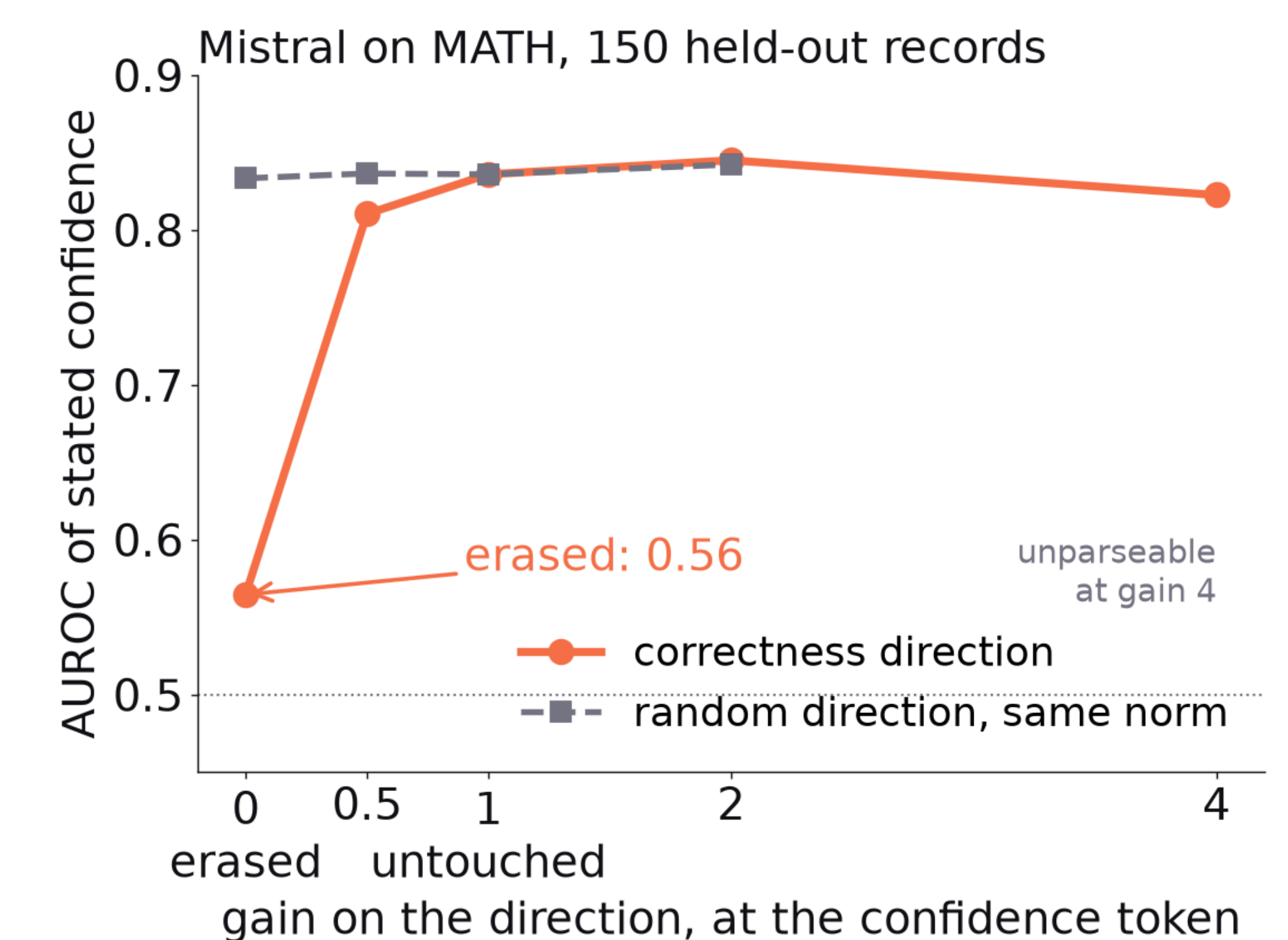
Probe AUROC by depth on MATH, four models. Layer 0 is chance because its token is identical in every record. The signal is present from the first layers and flat across depth. The dot is the fixed layer.

5 What we found



12 / 16

The state carries more than the report. The probe beats stated confidence in 12 of 16 cells. Mean gap 0.10. On MATH, 0.20.



0.84
→ **0.56**

On Mistral, the report uses the signal. Erase the direction and stated confidence stops separating right from wrong. A random direction does nothing. The model becomes uniformly sure, not confused.

Not every model. On GLM, erasing changes nothing. On Qwen3.6 the self-report is nearly flat to begin with, so there is nothing to erase. Gemma-4 leaves too few failures to test.

Four open models carry information about their own errors that their self-report does not.
In Mistral, we have shown it is used when it self-reports.