CONFERENCE ON LANGUAGE MODELING · **COLM 2026**

Self-Correction Bench

Uncovering and addressing the **self-correction blind spot** in large language models

Ken Tsui · Independent Researcher · kenhktsui@gmail.com

64.5%

LLMs reliably correct **your** errors — and fail on the **identical** error in their own output.

AVERAGE SELF-CORRECTION BLIND SPOT ACROSS 14 OPEN-WEIGHT NON-REASONING MODELS

14 non-reasoning models **3** datasets **2,047** injected errors **89.3%** removed by one token

01 The ambiguity

Is it because the model **cannot** fix the error — or because it **does not**?

INTERNAL · MODEL-ATTRIBUTED

`user` ▸ What is the answer of 1 + 1?
`assistant` ▸ **The answer is 3.**
... ends turn. No correction. ✖

IDENTICAL ERROR

EXTERNAL · USER-ATTRIBUTED

`user` ▸ What is the answer of 1 + 1? **The answer is 3.**
`assistant` ▸ *"That's not correct. The answer is actually 2..."* ✔

Same error, same context. **Only the chat-template role header differs** — so a correction gap cannot be explained by missing knowledge.

SELF-CORRECTION BLIND SPOT

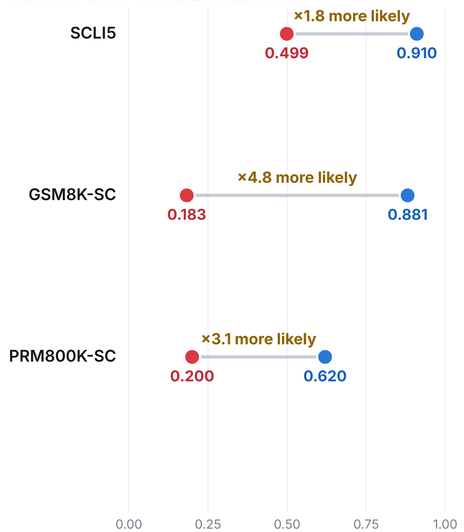
$$1 - \frac{\text{internal correction rate}}{\text{external correction rate}}$$

SCLI5	286 · trivial
GSM8K-SC	1,313 · multi-step maths
PRM800K-SC	448 · genuine LLM errors

02 It is systematic

Every family. Every scale. From **1 + 1** to multi-step proofs.

MEAN ACCURACY ON THE IDENTICAL ERROR



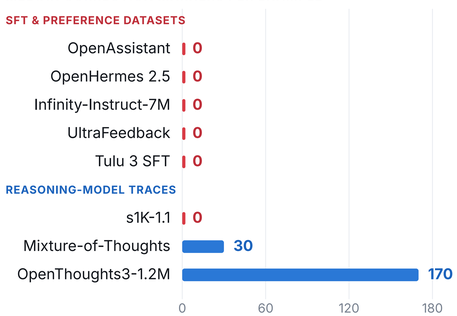
The models already knew. They simply never looked.

Claude 3.5 Haiku	52.5%
Claude Sonnet 4	41.4%
Non-maths tasks	62.8%

03 We taught it to them

Post-training data almost never shows a model **catching itself**.

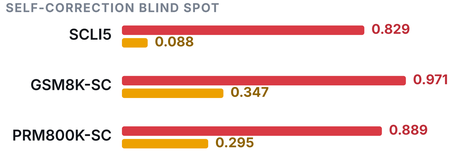
MEDIAN CORRECTION MARKERS PER EXAMPLE



Median "Wait", "But", "However"-type markers per example.

76% blind spot removed by fine-tuning on **5,306 error-correction traces**

SELF-CORRECTION BLIND SPOT



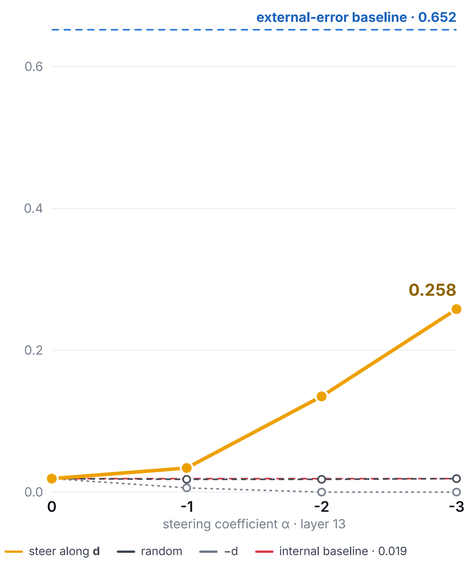
98% of those traces are non-mathematical — the correction behaviour still transfers to maths.

04 The mechanism

One direction in representation space **gates** correction.

d = mean(**h internal** − **h external**) at the last error token

MEAN ACCURACY · LLAMA-3.1-8B / GSM8K-SC



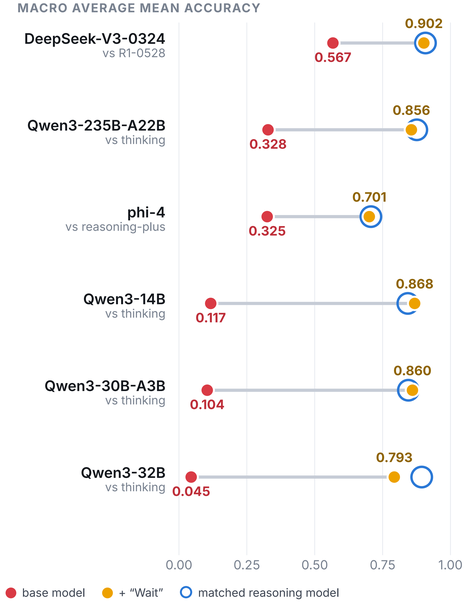
Conversational role gates correction.

Causal, monotonic, and absent for the negated and random controls. Replicates in five of six model×dataset pairs — Llama-3.1 and Qwen2.5 at 7–8B.

05 One token

89.3% of the blind spot removed by appending **"Wait"** · no training, no data

MACRO AVERAGE MEAN ACCURACY



Latent, not absent.

Measured on *injected* errors: this isolates activation from knowledge, and is not a deployment failure rate.

1 The capability is often already there.

2 We taught them the blind spot.

3 One direction gates it. One token opens it.

CODE DATA