# Asking the model to grade itself is a faster way to find what it already believes, not what is true.

**Dr Peter McCann Strain**
CTO, DPhil/PhD in AI from Oxford University

Swipe ››

— THE FRAMEWORK GAP

# Teams treat a fluent chain of thought as proof the work was checked. A trace looks like an audit trail because it has sequence, hesitation and explanation, but it is still an output channel produced by the same system whose answer you are assessing.

— THE REFRAME

# Testimony, not telemetry.

THE OLD QUESTION

## The trace shows me how the model got there.

↓ ↓ ↓

THE QUESTION THAT HOLDS UP

## The trace is a fallible witness; corroborate it.

— WHY THE FRAME HOLDS

# 36-point bias swing

**In Turpin et al.'s 2023 study, models were shifted by irrelevant biasing cues (answer-order and fictional user preferences) then failed to mention those cues, with reported accuracy shifting up to 36-point bias swing on some BIG-Bench Hard tasks (a public reasoning benchmark). The claim is narrow: visible reasoning is testimony, not telemetry.**

SOURCE

Turpin, Michael, Perez and Bowman (2023); later Reasoning Theater preprint and OpenAI chain-of-thought monitoring paper treated as emerging research, not settled doctrine.

—— HOW IT WORKS

# Run three questions before a trace counts as evidence.

01 Check **temporal commitment**: could the model have decided before it said so?

02 Check **cue omission**: could it have decided for an unstated reason?

03 Check **monitor gaming**: could it be performing carefulness for the observer?

— THE ARTIFACT

# Three questions before you trust a trace.

**Temporal commitment**

Did the decision exist before the story?

**Cue omission**

What moved the answer but vanished?

**Monitor-aware performance**

Is it acting careful for the observer?

*A reasoning trace is testimony, not a verdict. Ask whether the decision preceded the explanation, whether a cue vanished, and whether the trace is performing carefulness.*

— APPLY THE INSTRUMENT

**Pull the last reasoning-model output you accepted on trust this week. Run the three questions against logs, tests or an independent verifier; put external evidence next to the trace before the system acts again.**

— WHERE TEAMS MISREAD IT

# Adding a self-rated confidence score and shipping. A model grading its own trace shares the failure habits it just used. Add a structurally different verifier (different model family, rule-based check, or independent system) before the action lands.

READ THE FULL FRAMEWORK

# The Model Cannot Mark Its Own Work

Read the full essay — the argument, the sources, the figures and a reader-ready working artifact.
Substack petermccannstrain.substack.com · Medium @peter.mccann.strain · LinkedIn peter-strain-dphil-15a607128
New essays twice weekly, 2 June – 21 July 2026.

Next: E08 — Helpful, Harmless, and Wrong

## — THE STACK SO FAR

E07 · Essay 7 of 22 complete · Arc II: Evidence and authority

YOU JUST ADDED

### Three questions before you trust a trace

STACK LAYER LIT UP

### Checks / Runtime evidence

YOU CAN NOW ASK

### treat model reasoning as testimony, not proof.

NEXT

### E08 asks whether an aligned model is the same as a governable one.

# Dr Peter McCann Strain

CTO, DPhil/PhD in AI from Oxford University

I build production AI systems and write about making agentic AI useful, inspectable, governable and safe enough for real work.

Follow on Substack for the full 22-essay series
petermccannstrain.substack.com