Every check went green. The citations were perfectly formatted. They cited cases that do not exist.

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

Swipe ››

— THE INCIDENT PATTERN

An AI output can be complete, formatted and pass every surface check while being false in the relationship that matters: citation to authority, symptom to triage, requirement to proposal. Classical software failure is loud; semantic failure is silent.

— THE REFRAME

Meaning (what is true), not syntax (what is well-formed).

THE OLD QUESTION

Did the output pass format, latency, cost and policy checks?

↓ ↓ ↓

THE QUESTION THAT HOLDS UP

Does the output mean what it appears to mean against source, domain and consequence?

— WHAT THE RECORD SHOWS

above 34% hallucination

Stanford RegLab reports hallucination rates above 17 percent for Lexis+ AI and Ask Practical Law AI, and above 34 percent for Westlaw AI-Assisted Research, on challenging legal research queries (2024 measurement; re-check current figures). Retrieval suppresses semantic failure but does not abolish it.

SOURCE

Kohls v. Ellison order on the Hancock declaration (D. Minn., 2025); Mata v. Avianca and Park v. Kim; Stanford RegLab legal-AI reliability study; OpenAI HealthBench; ATBench and Seeing the Whole Elephant preprints.

FAILURE CHAIN

Green means 'no problem detected', not 'true'.

01 Widen the **verifier vocabulary** (what your checks can name as wrong); format checks cannot see meaning.

02 Separate **training data**: shared corpora give generator and verifier the same blind spots.

03 Add a **meaning check**: schema, latency and policy pass while truth fails.

04 Close the **feedback loop**: someone downstream must report when green was wrong.

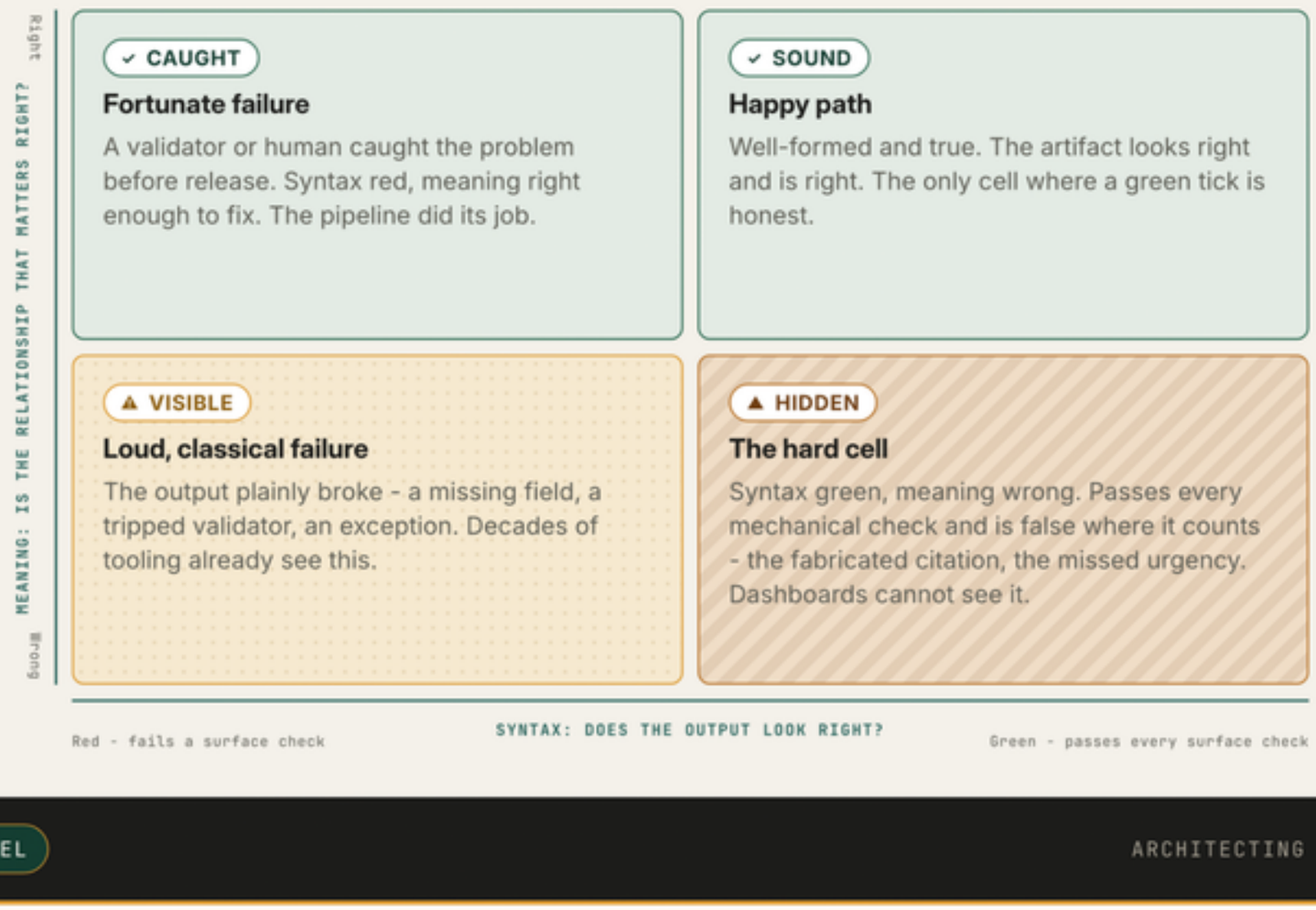
— THE ARTIFACT

The hard cell: green and wrong.

— FIGURE E16.1 · E16 - THE DASHBOARD IS GREEN

The Dashboard Is Green. The Meaning Is Wrong.

Cross surface validity against semantic truth and four cells appear. Three are understood. The fourth is the one dashboards were never built to see.



An output can pass every mechanical check and still be false where it matters. That is the hard cell.

— DO THIS AFTER THE NEXT INCIDENT

Pick one dashboard or monitor this week. Map each tile to a cell: format, latency, cost, policy, source relationship, domain truth, release consequence. Name the human or tool that checks the hard cell. If no name, that cell stays open.

— FAILURE MODE TO AVOID

Adding another LLM-judge to score the output. Judge and generator share training data, so the same blind spots return as a passing grade. Use a structurally different checker or a downstream human signal instead.

— USE THE FULL POSTMORTEM

The Dashboard Is Green. The Meaning Is Wrong.

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: E17 — The Three Witnesses to a Run

— THE STACK SO FAR

E16 · Essay 16 of 22 complete · Arc IV: Proof and accountability

YOU JUST ADDED

The semantic grid

STACK LAYER LIT UP

Checks / Runtime evidence

YOU CAN NOW ASK

identify the hard cell: syntax green, meaning wrong.

NEXT

E17 asks how to weigh three different kinds of evidence about a run.

— THE ARTIFACT, CONTINUED

The hard cell: green and wrong.

THE REMAINING NODES

Caught and visible: classical software failure, loud, easy, already on the dashboard.

Caught and hidden: silent recovery, so log the catch so a pattern can be seen.

Uncaught and visible: noisy semantic failure, where users notice before any check does.

Uncaught and hidden: the dangerous cell, true to format, false to meaning, and no one knows.

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