OpenAI retired its own headline benchmark. Of the tasks it audited, 59.4% had flawed tests.

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

Swipe ››

— THE BUYER RISK

A model card tells you what the vendor measured, under conditions the vendor chose, against tasks that may not resemble your deployment. It cannot prove your workflow will catch, contain and bound the model's mistakes.

— THE REFRAME

Inputs, not promises.

THE OLD QUESTION

How reliable is the model?

↓ ↓ ↓

THE QUESTION THAT HOLDS UP

Which cell of the reliability equation does each document actually fill?

WHAT TO ASK FOR

59.4% of audited tasks flawed

OpenAI retired its own headline SWE-bench Verified benchmark on 23 February 2026 after auditing 27.6 percent of the test set and finding at least 59.4 percent had flawed tests.

SOURCE

OpenAI, SWE-bench Verified deprecation note (23 February 2026); with SWE-bench Pro authors and the Stanford HAI 2026 AI Index.

CHECKLIST LOGIC

Every model-card benchmark claim is a compound claim that decays.

01 Decompose the headline as **model x scaffold x benchmark version x evaluator**.

02 Compare buying cycle against benchmark **half-life**; assume the card is stale.

03 Use the card for part of **P(error)** only, not detection, containment or severity.

— THE ARTIFACT

The Reliability Scorecard.

Model card fills error evidence	**Buyer must source** detection
Buyer must source containment	**Buyer must source** severity

A 4x3 (twelve-cell) procurement instrument: three claim classes crossed with the four terms of the reliability equation. A model card reaches only the error column. The rest are buyer-sourced, not vendor claims.

— ASK THIS ON MONDAY

At the next vendor review, put five documents on the table: model card, system card, eval report, security review, RFP response. Ask each one: which cell of the reliability equation does this fill?

VENDOR TRAP

Letting the cleanest document carry the broadest promise. A model card cannot stand in for the system card, eval report, security review and RFP response. Demand the document built for the question you are asking.

— USE THE CHECKLIST

The Model Card Won't Save You

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: E07 — The Model Cannot Mark Its Own Work

— THE STACK SO FAR

E06 · Essay 6 of 22 complete · Arc II: Evidence and authority

YOU JUST ADDED

The vendor-evidence scorecard

STACK LAYER LIT UP

Evaluation / Checks

YOU CAN NOW ASK

translate vendor claims into reliability evidence.

NEXT

E07 asks whether a model's own reasoning trace counts as a check.

— THE ARTIFACT, CONTINUED

The Reliability Scorecard.

THE REMAINING NODES

Down the side, three claim classes: Accuracy, Helpfulness, Safety.

Across the top, four reliability-equation terms: error, undetected, uncontained, severity.

Score each cell Owned (source, date, scaffold and transfer argument), Asserted (a claim with no transfer argument), or Buyer-fillable (blank, inside your own control).

The blank (Buyer-fillable) cells are the buyer's to-do list.

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