Two green slides say the eval passed. They measured the wrong thing.

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

Swipe ››

THE BUYER RISK

Most evaluation programs measure capability on a clean run, when the launch decision needs authority fitness: whether repeated runs, under realistic users, with tools and escalation, are safe, correct, contained and attributable.

— THE REFRAME

Promotion board, not leaderboard.

THE OLD QUESTION

Can the agent answer the question?

↓ ↓ ↓

THE QUESTION THAT HOLDS UP

What should this system be allowed to do?

— WHAT TO ASK FOR

1-in-4 pass rate

tau-bench tests tool-using agents inside moving customer conversations and introduced pass^k (the share of k repeated trials the agent passes EVERY time, not just at least once). Its original experiments saw state-of-the-art agents succeed on fewer than half of tasks, with the all-pass measure in retail lower still, roughly one trial in four. The figures move with every model release; the lesson is that repeated interaction exposes a variance clean prompts hide.

SOURCE

tau-bench (Yao et al., 2024) and Anthropic's agent-evals guidance, which treats the all-pass pass^k, not pass@k, as the customer-facing reliability question.

CHECKLIST LOGIC

A four-axis promotion board, walked one axis at a time.

01 Test **capability**: can the system perform the task class at all?

02 Test **reliability**: does it hit the stated bar across repeated trials?

03 Test **autonomy fitness**: is each action at the right rung?

04 Test **governance readiness**: owners, traces and reversal paths in place?

— THE ARTIFACT

The four-question promotion board.

Capability
Can it do the task once?

Reliability
Does it hold across repeats?

Autonomy fitness
Which action class has earned authority?

Governance readiness
Can the run be audited?

The four questions a launch decision has to answer: capability, reliability, autonomy fitness, governance readiness. Capability is the only one a public benchmark touches.

— ASK THIS ON MONDAY

Pick one support-like workflow this week. Choose the single action you would most like to make autonomous. Run pass^8 against tasks drawn from your real policies, tools and escalation rules. Stop at the first repeat failure.

VENDOR TRAP

Quoting the vendor's public benchmark as the promotion-board score. Public pass@1 tells you the field moved, not whether the agent survives repeated runs on your policies. Re-run pass^8 on your data before promoting.

— USE THE CHECKLIST

You Cannot Benchmark a Coworker

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: E21 — The Autonomy Ladder

— THE STACK SO FAR

E20 · Essay 20 of 22 complete · Arc V: Operating model

YOU JUST ADDED

The Readiness Promotion Board

STACK LAYER LIT UP

Evaluation / Runtime evidence / Permissions

YOU CAN NOW ASK

evaluate an agent through a promotion board, not a leaderboard.

NEXT

E21 asks which autonomy rung each action class has earned.

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