Google lost $100 billion of market cap in one afternoon because one demo slide said the wrong thing.

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

Swipe ››

— THE INCIDENT PATTERN

A demo proves capability can appear under arranged conditions. Deployment has to prove the same capability survives real users, stale state, permissions, failures, costs, audit trails and consequences. The mistake is treating the first as evidence for the second.

— THE REFRAME

Model inside system.

THE OLD QUESTION

Can it do the task?

↓ ↓ ↓

THE QUESTION THAT HOLDS UP

Under what system conditions did it succeed, and which survive launch?

WHAT THE RECORD SHOWS

0.6% vs 4.4% (same model)

On the TravelPlanner benchmark, GPT-4 Turbo scored 0.6 percent end-to-end in the two-stage tool-use setting but 4.4 percent in the simplified direct setting. Same model: the scaffold around it decided whether the demo survived deployment conditions.

SOURCE

Xie et al., TravelPlanner: A Benchmark for Real-World Planning with Language Agents (2024).

FAILURE CHAIN

The thing that made the demo work is not always the thing on the slide.

01 Treat the demo as a **constructed artefact**, with a chosen dataset, known route and reviewer present.

02 Map the wider **task distribution**: partial instructions, stale tool state, timeouts, leaked permissions.

03 Give errors **somewhere to go** before they hit the world. That is production-grade.

— THE ARTIFACT

Six gates from demo to deployment.

Task distribution
Does demo match deployment?

Verifier
Independent check

Containment
Bounded action

Severity tier
Risk by action class

Audit trail
Reconstruct the run

Stopping rule
Fail safely

Six gates each name a specific instrument: task resemblance (does the demo task match deployment?), verifier (who grades the action?), severity tier (classify by reversibility), containment (where can errors safely land?), audit trail (what survives the run?), stopping rule (when does the loop end?).

— DO THIS AFTER THE NEXT INCIDENT

At the next agent demo, before accuracy, ask whether the demo task resembles deployment. Name the verifier, containment boundary, severity tier, audit trail and stopping rule for the exact action shown. Any blank is design work.

FAILURE MODE TO AVOID

Quoting the demo accuracy in the rollout deck. A happy-path number does not price rejected plans, permission denials, retries or rollback. Re-test the same action under stale state and leaked permissions before quoting again.

— USE THE FULL POSTMORTEM

Demo Is Not Deployment

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: E04 — The Nine Layers Where Agents Break

— THE STACK SO FAR

E03 · Essay 3 of 22 complete · Arc I: See the object

YOU JUST ADDED

Demo-to-deployment gates

STACK LAYER LIT UP

Checks / Runtime evidence / Evaluation

YOU CAN NOW ASK

separate demo evidence from deployment evidence.

NEXT

E04 asks where a deployed agent actually breaks.

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