

NOT A LAUNCH EVENT

Not all AI agents are built the same.

Training one you can trust looks less like shipping software and more like onboarding staff: assess it, verify it, re-benchmark it, and always keep a way to roll it back.

Five disciplines that make an agent trustworthy

- 01** Assess proficiency before deployment
- 02** Verify against intent, not just accuracy
- 03** Re-benchmark on a fixed schedule
- 04** Give human oversight real authority
- 05** Build the rollback plan before you need it

40%+

of agentic AI projects will be canceled by the end of 2027

The stated reasons are escalating costs, unclear business value, and inadequate risk controls, not that the models were incapable of the task. Gartner, 2025.

61% → 25%

GPT-4o's success rate on identical retail-agent tasks, run once versus run eight times

A single pass says nothing about consistency, and consistency is what a business actually depends on. τ -bench, Yao et al., 2024.

An agent deserves the same onboarding discipline as a new hire

A job description that defines the task and its boundaries. A probation period where output is checked against the real outcome, not just whether it parses. **An agent that passes a demo once has demonstrated luck, not proficiency.**

~7 months

**the doubling time for the length of task a
frontier agent can complete at 50% reliability**

A proficiency test run at go-live is measuring a capability profile that looks different within two quarters. METR, 2025.

TERENCE KOK · AI GOVERNANCE SERIES

OVERSIGHT WITH TEETH

07 / 10

Human-in-the-loop only works if the human has real authority

Not someone glancing at outputs with no training. A reviewer who understands the system's limits, can correctly read what it produced, and can override or halt it before the action lands. EU AI Act Article 14 sets this bar for high-risk systems from December 2027.

TERENCE KOK · AI GOVERNANCE SERIES

ON ROLLBACK

08 / 10



A rollback path that has never been executed is a hypothesis, not a plan.

PIN THE VERSION. MONITOR THE DRIFT. REHEARSE THE PULLBACK.

TERENCE KOK · AI GOVERNANCE SERIES

STARTING CHECKLIST

09 / 10

Five disciplines, none of them exotic



Skills assessment

Test repeated, identical attempts, not one clean demo



Intent verification

Check the outcome against the business goal, not the easiest metric



Scheduled re-benchmarking

On a calendar, and after any model, tool, or data change



Trained human oversight

Defined override authority and escalation triggers



Rehearsed rollback

A pinned known-good version, tested before it's needed

TERENCE KOK · AI GOVERNANCE SERIES

CLOSING

10 / 10

Deployment is day one of an ongoing job, not the finish line of a build.

There is no proficiency level where the review process gets switched off, because the ground it's tested against keeps moving.



Terence Kok · AI & Innovation Strategist

terencekok.com

SAVE · COMMENT · FOLLOW

TERENCE KOK · AI GOVERNANCE SERIES