# Beyond the Pilot

## What Agentic AI is Starting to Look Like in Production

THE ENTERPRISE REALITY

# The harder question

Most enterprises have already run AI experiments. The question CIOs now face is which use cases justify the investment, the disruption, and the scrutiny they trigger.

MUST JUSTIFY

### Infrastructure investment

Integration, data access, observability

MUST JUSTIFY

### Change-management cost

New operating models, roles and skills

MUST JUSTIFY

### Board-level scrutiny

Risk, governance and demonstrable ROI

WHAT THIS SESSION SHOWS

Agentic AI running in a real enterprise environment — **not as an assistant, but as an autonomous execution layer** across workloads.

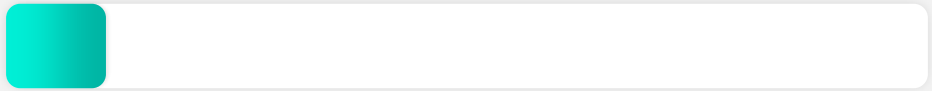
EPI-USE

THE PRODUCTION GAP

# Escaping “Pilot Purgatory”

**11%**

of enterprises move agentic AI from pilot into production. Nearly nine in ten never make the crossing.



*Source: industry analyses, 2025–26 (IDC, Deloitte, Gartner).*

## The blocker isn’t the AI

Models have matured, multi-agent orchestration is production-ready, and integration standards now exist. What stalls projects is the environment around the model:

- Data access, quality and lineage
- System and API integration into core platforms
- Identity, permissions and security boundaries
- Observability, monitoring and audit
- Governance and operational controls

Legacy infrastructure — not intelligence — is the real constraint.

WHERE THE ADVANTAGE IS

# The moat is orchestration, not the model

Model choice is converging toward a commodity. The defensible advantage is how agents are wired together — and that resists copying.

COMMODITIZES

## Model selection

Frontier models are increasingly interchangeable and available to everyone. Picking one is not a moat — your competitors can swap in the same model tomorrow.

DOESN'T COMMODITIZE

## Orchestration

Wiring agents together across your systems, permissions and constraints. It is specific to your environment and processes — which is exactly why it is hard to copy.

THE DURABLE IP

- System integration
- Identity & permissions
- Business constraints
- Coordination logic

RESPONSIBLE AUTONOMY

# From human-in-the-loop to human-on-the-loop

Autonomy is acceptable only when it is both governed and cost-controlled. Two disciplines make multi-agent action responsible at scale.

## Governance

**Supervise the system, not every action**

- Policy guardrails and scoped permissions
- Full audit trails and decision traceability
- Defined escalation and override paths
- Bounded action space — agents can't exceed mandate

## FinOps

**Keep autonomous spend under control**

- Budgets and hard spend caps per agent
- Token and resource metering in real time
- Cost attribution by workload and outcome
- Anomaly and runaway-spend alerts

*Human-on-the-loop means designing the guardrails once, then supervising — not approving every step.*

EPI-USE

BUILD FOR PRODUCTION

# Most failures are context failures

When an agent fails in production, the model is rarely the problem. It lacked the right context at the moment of action.

## Intelligence failure

**Rare.** Frontier models are already capable enough for the great majority of enterprise tasks.

## Context failure

**Common.** The agent lacked the data, state, permissions, tools or constraints it needed to act correctly.

### WHAT THE AGENT NEEDS AT DECISION TIME

- Data & state
- Permissions
- Tools
- Constraints
- Grounding

*So the work is context engineering, not chasing a smarter model — which is where production architecture begins.*

THE PROBLEM

# Two failure modes — not one

Redundancy in an AI ecosystem has to answer two different questions. Conflating them is what drives the fear that you must build everything twice.

## "DOWN"

Availability — something stops responding

**Your own infra: solved by conventional HA — replicas, multi-region, failover.** But you can't replicate a model you rent. A provider outage can only be met by a different model (the "wrong" problem) or a self-hosted fallback.

## "WRONG"

Quality — the system answers, but incorrectly

**The hard problem.** Models are non-deterministic — even one vendor won't reproduce the same answer. This is where redundancy gets misunderstood.

**THE TRAP:** "if I fail over from one model to another, I need double the systems." That conclusion only holds if you assume the two are interchangeable. They are not.

MYTH-BUSTER

# Why a second model is not the answer

Redundancy assumes components fail independently. Two frontier LLMs do not — and the failures that matter are exactly the shared ones.

## Hard inputs are hard for both

Ambiguous, adversarial or out-of-distribution queries break the primary and the backup alike.
(The classic N-version programming finding: independence is assumed, then empirically fails.)

## Shared training & behaviour

Frontier models draw on overlapping data and similar tuning. The ecosystem is closer to a monoculture than to independent suppliers.

## One poisoned context breaks all

Both models sit behind the same retrieval layer. A bad or poisoned chunk feeds identical garbage to every model downstream.

## Your validation goes dark at the worst moment

Running both lets you validate the backup by comparing it to the primary — but that reference vanishes precisely when the primary fails.

EPI-USE

THE PRINCIPLE THAT WORKS

# Don't make the model reliable. Make it ***irrelevant to safety.***

Borrowed from avionics and other safety-critical fields: shrink the part of the system that must be correct. The LLM proposes; a small, verifiable component decides.

**LLM — PROPOSER**

Powerful, flexible, untrusted. Optimised for quality, never the authority.

**VERIFIABLE KERNEL — DISPOSES**

Small, deterministic, trusted. Checks, constrains, or overrides before anything ships.

EPI-USE

REFERENCE MODEL

# What good architecture looks like

**Defence in depth:** independent, imperfect barriers. The model lives in the top band; the guarantee lives in the bottom.

| | |
|---|---|
| **Grounded retrieval** | Provenance + coverage; refuse to answer on thin evidence |
| **LLM proposer** | Generates the candidate answer — untrusted |
| **Diverse-path disagreement sensor** | Cheap second path; disagreement = escalate |
| **Deterministic invariant gate** | Checks rules you CAN specify; not 'is it good?' |
| **Abstention & escalation** | Allowed to say 'I can't safely answer' |
| **Decorrelated fallback** | Rules engine · cached answer · retrieval-only |

## UNTRUSTED

Optimised for quality.
May be wrong at any time.

## TRUSTED PERIMETER

Small, deterministic, verifiable.
Carries the actual guarantee.

EPI-USE

THE USEFUL VERSION OF 'DUAL'

# Repurpose the second model

The dual-model instinct is sound — if its job changes. Not a spare tire that substitutes; a sensor that tells you when not to trust the answer.

## ✕ SPARE TIRE

**Run a second model to substitute and to improve answers.**

Fails on correlation, doubles cost and gives false confidence.

## ✓ UNCERTAINTY SENSOR

**Diverse, cheap, sampled paths. Disagreement = a hard input.**

On disagreement: escalate, abstain, or route to the fallback — don't pick a winner.

**WHY IT'S CHEAPER:** the sensor need not be a second frontier model on 100% of traffic — a small, diverse, sampled model is enough. 'Is this answer good?' is hard; 'do these two materially disagree?' is far easier.

EPI-USE

INTELLECTUAL HONESTY

# What is still unsolved

Stated plainly, because credibility with a board and clients depends on it. These are open problems — no vendor has closed them today.

| | |
|---|---|
| **Calibrated uncertainty** | Knowing when a free-form answer is confidently wrong. Cross-path disagreement is a proxy, not a solution. |
| **Automated quality evaluation** | Scoring open-ended output with no trusted reference. Using an LLM to judge an LLM is correlated and unreliable. |
| **Formal verification at scale** | Proving frontier-model behaviour over the full input space is intractable. You can box the model; you cannot prove it safe. |
| **Model monoculture** | True diversity is scarce when all frontier models share data and methods — your 'decorrelated' backup is less independent than it looks. |

*Where none of these can be specified or measured, the LLM should not be load-bearing — keep it advisory.*

THE BUSINESS CASE

# The Economics

If the architecture needs this many layers, why would any business adopt it?
The honest answer — and where the returns actually live.

THE REFRAME

# Segment by consequence of error

The full harness is not a tax on every use case. It is the cost of the mission-critical sliver only. Most value lives where the harness is unnecessary.

LOW CONSEQUENCE · HIGH VOLUME

## ROI is real and immediate

Drafting · summarisation · code assist · search · internal tooling · support triage.

**No rules engine, no redundant models, no heavy verification. This is where the money is today.**

HIGH CONSEQUENCE · MISSION-CRITICAL

## Where the full harness applies

A single wrong answer carries serious cost.

Only here do you pay for verification, decorrelated fallback and continuous evaluation — and only here must you ask whether the LLM belongs in the path at all.

*Applying the mission-critical stack everywhere is self-inflicted cost. Match the architecture to the stakes.*

EPI-USE

COST INVERSION

# Routed correctly, it reduces cost

'Double the systems' is the worst-case framing. Done well, the deterministic layer is cheaper and the LLM becomes a selective specialist — not an always-on duplicate.

1

## Deterministic bulk path

Rules engine handles the high-volume easy cases at near-zero marginal cost.

2

## LLM on the long tail only

Invoked selectively for the messy, varied inputs it is genuinely good at.

3

## Fractional-cost sensor

Small, sampled diverse model as the disagreement signal — not a full second frontier.

**NET EFFECT:** the LLM stops being a redundant layer bolted on top and becomes a specialist you invoke sparingly. The relevant comparison is never 'AI vs. doing it cheaply' — it is **AI-plus-harness vs. the existing cost base, usually expensive human labour.**

EPI-USE

BE HONEST

# Where the returns are — and aren't

## ROI is real

- High volume to amortise fixed cost
- Consequence is bounded or verifiable
- A costly human baseline to beat
- Infrastructure reused across features

## ROI is bleak

- Full automation of open-ended judgment
- Consequence catastrophic and irreducible
- Volume too low to repay the build
- No way to specify or measure 'correct'

### THE COST EVERYONE UNDER-BUDGETS: EVALUATION

Building and maintaining the measurement to know whether outputs are good enough is the perpetual line item. It is both the safety requirement and the proof of ROI — most stalled AI programmes die here, unable to scale safely or to show they should.

EPI-USE

TAKE THIS TO THE TABLE

# The decision rule

Build the mission-critical harness only when all three hold. If any one rule fails, don't build it — use a human with the LLM as a multiplier, or keep AI out of the critical path.

| | | |
|---|---|---|
| 1 | **Consequence is bounded or verifiable** | You can specify hard constraints a valid answer must meet — or cap the damage a wrong one can do. |
| 2 | **Volume amortises the harness** | Enough throughput to repay the verification, fallback and evaluation infrastructure. |
| 3 | **The alternative costs more** | AI-plus-harness beats the existing cost base on the routable share of the work. |

**Fail any one → the harness is a warning light that you've mis-scoped the problem.**

EPI-USE

THE BOTTOM LINE

# Real returns, narrower than the hype.

**Don't make the model reliable — box it.**

Spend the budget on the deterministic perimeter and on grounding, not on duplicating models.

**A second model is a sensor, not a spare.**

Use diverse paths to detect disagreement and trigger abstention or a decorrelated fallback.

**The value is concentrated, not universal.**

High volume · bounded or verifiable consequence · a costly baseline to beat · reusable infra.

**Where none of that holds, stay advisory.**

Keep the human as decision-maker. That return is modest but real; pretending otherwise loses money.

EPI-USE


# HOW LEADERS NEED TO RETHINK THEIR APPROACH TO AI ALTOGETHER

MOVING FORWARD

# Scan to Request Your Complimentary AI Assessment

One hour with our team. Focused entirely on your AI use case. No pitch. No obligation. Available through June 30, 2026.

- Identify your highest-impact AI use case
- Understand the infrastructure gaps blocking production
- Walk away with a clear path forward on AWS

*Scan to book your session*

**epiusecloud.com/cio-summit-2026**

# Transforming Organizations through Cognitive Automation


**Christopher Belford**
Associate Partner
christopher.belford@epiuse.com