

BEYOND PILOT SUCCESS

Pilot Purgatory — Why 90% of GenAI Proofs of Concept Were Designed Never to Scale

On the Incentive Equilibrium That Keeps Enterprise AI Permanently Experimental

Giovanni Leonardi

February 2024



Contents

1. Executive Summary	3
2. The Demonstration That Wasn't a Rehearsal	3
3. The Architecture of a Friendly Experiment	4
4. The Equilibrium Nobody Discusses	5
5. The Myth of the Incremental Bridge	5
6. In Defence of Caution, and Its Limits	6
7. The Production-First Pilot	7
8. The Question We Are Avoiding	8

Pilot Purgatory — Why 90% of GenAI Proofs of Concept Were Designed Never to Scale

Transformation · No. 303

Originally written February 2024 | Re-edited and re-rendered for the WhereBeyond Edition,
2025

"The proof of concept was never the problem. The problem was building it to succeed."

1. Executive Summary

Across enterprises in early 2024, a pattern has become impossible to ignore: generative AI pilots that succeeded brilliantly as demonstrations have stalled indefinitely short of production deployment. This essay argues that pilot purgatory is not a failure of execution but a stable equilibrium — one in which executives, vendors, and delivery teams each derive real benefit from the pilot remaining precisely what it is. The proof of concept was designed, often unconsciously, to succeed on its own terms by sidestepping every question that production would force: access to real data, integration with existing workflows, verification at scale, and accountability for outcomes. Escaping this equilibrium requires not better iteration on existing pilots but a fundamentally different starting point — the production-first pilot, designed from day one to confront the organisational, technical, and governance questions that the conventional format was built to defer.

2. The Demonstration That Wasn't a Rehearsal

The quarterly innovation review runs to its familiar rhythm. The team demonstrates a generative AI tool that summarises customer complaints, pulling from a curated dataset of three hundred cases, each cleaned and tagged by hand. The accuracy numbers are strong — eighty-seven per cent, measured against the same team's own judgement. The chief digital officer nods. The vendor's account director smiles. A slide titled "Next Steps" proposes a "Phase 2 expansion" with a timeline that trails off into the second half of the year. Nobody in the room asks the question that would end the presentation early: what happens when this tool meets the four hundred thousand untagged, misspelled, multilingual complaints that arrive each quarter through eleven different channels?

That question is not asked because nobody present has an incentive to ask it. And this, rather than any technical limitation, is why the pilot will still be a pilot twelve months from now.

We have spent the past year watching this scene play out across industries. The technology itself has moved with extraordinary speed — capabilities that were the province of specialist teams and research labs barely a year ago are now accessible through APIs, wrapped in enterprise platforms, and branded as copilots. But the organisational response to that speed has followed an older, more cautious script: the proof of concept. And the proof of concept, as it is typically constructed, is not a rehearsal for production. It is a performance designed for a different audience entirely.

3. The Architecture of a Friendly Experiment

To understand why so many generative AI pilots succeed without ever progressing, we need to examine what they are actually built to prove. The typical enterprise pilot is constructed around five design choices, each of which makes the demonstration more impressive and production deployment more remote.

Sandboxed data. The pilot runs on a carefully selected, pre-cleaned subset of data — often extracted, transformed, and loaded by the pilot team itself. The messy reality of enterprise data — inconsistent formats, missing fields, contradictory entries spread across systems that do not talk to one another — is excluded by design. The model performs well because it has been given the equivalent of an exam where the questions were written by the student.

Enthusiast users. Pilot participants are typically volunteers: the digitally curious, the innovation-minded, the people who already believe the technology works. They adapt their behaviour to the tool rather than expecting the tool to adapt to theirs. Their feedback is enthusiastic and constructive — and entirely unrepresentative of what will happen when the tool is placed in front of a claims handler with twenty years of experience and no interest in changing how they work.

Bespoke integration. Where the pilot connects to existing systems at all, it does so through custom scripts and manual workarounds maintained by the pilot team. The API calls are handcrafted. The data flows are monitored in real time. The error handling is a developer watching a dashboard. None of this scales, and none of it is designed to.

Friendly success criteria. The metrics that define pilot success are almost always written by the team running the pilot, often with implicit vendor input. "User satisfaction" is measured among enthusiasts. "Accuracy" is measured on clean data. "Time savings" are estimated, not measured against a baseline. The criteria exist to be met, which is different from existing to determine whether the technology works.

No accountability for outcomes. The pilot operates in a consequence-free zone. If the summarisation tool misinterprets a complaint, no customer is affected. If the document generator hallucinates a clause, no contract is signed. The entire apparatus of governance, compliance, and operational accountability that production systems must bear is absent — not deferred, but structurally excluded.

Each of these choices is individually reasonable as a way to manage risk in early experimentation. Taken together, they create something else: a demonstration environment so different from production that the distance between them cannot be crossed incrementally.

4. The Equilibrium Nobody Discusses

Here is the uncomfortable truth that the innovation review will never surface: pilot purgatory is not a problem that any single actor in the system is trying to solve, because for each actor, the pilot in its current state is already delivering what they need.

For the executive sponsor, the pilot is proof of momentum. In a landscape where every board presentation now includes a slide on "AI strategy," a running generative AI pilot — with user testimonials and accuracy metrics — is evidence that the organisation is not falling behind. The pilot's value is reputational, not operational. Scaling it would introduce risk, cost, and difficult questions about data governance and workforce impact. Keeping it running is cheaper and safer than either scaling or killing it.

For the technology vendor, the pilot is a foothold. Every enterprise pilot represents a relationship, a reference, and a recurring conversation about "Phase 2." The vendor's commercial incentive is to maintain the pilot at a level of success that justifies ongoing engagement without pushing the client toward the kind of production deployment that would expose integration costs, reveal performance gaps under real conditions, and potentially lead to the uncomfortable discovery that the value case is thinner than the demo suggested. The optimal state, from the vendor's perspective, is a permanently promising pilot.

For the delivery team, the pilot is a career asset. The data scientists, engineers, and innovation leads running the pilot are building skills, visibility, and credentials in the most sought-after domain in technology. A successful pilot on the CV is almost as valuable as a production deployment, and considerably less stressful to deliver. The incentive to push for production — with its messy integration work, its organisational resistance, its accountability — is weaker than it appears.

This is not cynicism. These are rational actors responding to the incentive structures in which they operate. The pilot persists not because anyone is trying to keep it in purgatory, but because purgatory is an equilibrium — a state from which no single actor benefits from departing unilaterally.

Pilot purgatory is not a failure of execution. It is an equilibrium in which every stakeholder is already getting what they need from the experiment — and none has sufficient incentive to force the disruption that production would require.

5. The Myth of the Incremental Bridge

The standard narrative for moving from pilot to production is incremental expansion: broaden the dataset, add more users, integrate with one more system, extend the success criteria. Phase 2 leads to Phase 3, which leads to a "minimum viable product," which leads to a gradual rollout. This narrative is comforting and, in the case of generative AI, largely false.

The gap between a typical generative AI pilot and production deployment is not a continuum that can be crossed in stages. It is a series of discontinuities — points where the nature of the problem changes qualitatively, not just quantitatively.

Data access is a governance problem, not a technical one. Moving from a sandboxed dataset to live enterprise data requires not a bigger pipeline but a fundamentally different conversation about data classification, access controls, retention policies, and cross-border data flows. In regulated industries, this conversation alone can take longer than the pilot itself.

Verification at scale has no current solution. A pilot team can manually review fifty AI-generated summaries a day. Production might require five thousand. The verification bottleneck is not a scaling challenge — it is an unsolved problem. We do not yet have reliable, automated methods to verify whether a large language model's output is correct, complete, and free from hallucinated content in the specific domain context where it is being applied. Until we do, every deployment that requires accuracy is constrained by human review capacity.

Process redesign, not process augmentation. The pilot narrative frames generative AI as a tool that enhances existing workflows — the assistant that drafts while the human reviews, the copilot that suggests while the professional decides. But meaningful value extraction typically requires redesigning the process itself: changing what work is done, in what sequence, by whom, with what decision rights. This is organisational change management, not technology deployment, and it moves at a fundamentally different speed.

Accountability requires new frameworks. When an AI-generated output causes harm — a misclassified claim, an incorrect regulatory response, a misleading customer communication — who is accountable? The model? The vendor? The team that deployed it? The manager who approved the process? Existing governance frameworks were not designed for outputs generated by probabilistic systems, and building new ones requires legal, compliance, and operational expertise that most pilot teams neither possess nor have access to.

These are not problems that "Phase 2" solves. They are the problems that the pilot format was designed — quite effectively — to avoid.

6. In Defence of Caution, and Its Limits

It would be dishonest to dismiss the cautious approach entirely. The instinct to experiment before committing is sound, particularly with a technology as novel and unpredictable as generative AI. Large language models do hallucinate. They do produce confident nonsense. The failure modes are unlike anything most enterprise technology teams have encountered, and the downside risks in regulated industries — healthcare, financial services, legal — are real and severe.

The case for careful experimentation has genuine force, and we should be wary of the opposite error: rushing to production with a technology whose behaviour we do not fully understand, in pursuit of a value case driven more by competitive anxiety than by

evidence.

But there is a critical distinction between caution and the kind of structured avoidance that the typical pilot represents. Genuine caution would design the experiment to surface the hard problems early — to test the technology against real data, with real users, under real governance constraints, precisely because those are the conditions where failure is instructive. What we have instead is an experimental format that systematically excludes every condition that would generate useful learning about production viability.

The pilot, as currently constructed, does not reduce risk. It defers it. And deferred risk does not diminish with time; it compounds, because every month spent in purgatory is a month in which the organisation is building confidence in a result that does not transfer to the environment where it matters.

7. The Production-First Pilot

If the conventional pilot is designed around what the technology can do under ideal conditions, the production-first pilot is designed around what the organisation must solve before the technology can be deployed. The difference is not incremental; it is architectural.

A production-first pilot begins with the constraints, not the capabilities. It asks: what are the five hardest problems that production deployment would require us to solve? And it designs the experiment to confront at least three of them from day one.

In practice, this means several uncomfortable departures from the standard playbook.

Real data from the start. Not a cleaned subset — the actual data, in its actual state, with its actual quality problems. If the model cannot handle the data as it exists, that is the most important finding the pilot can produce. A pilot that runs on curated data answers a question nobody is asking.

Reluctant users, not enthusiasts. The pilot's user group should include — perhaps should primarily consist of — the people who are sceptical, busy, and set in their workflows. If the technology cannot earn adoption from the people who will actually use it in production, the pilot has identified the real constraint, which is almost never technical.

Governance on day one. Data classification, access controls, output verification protocols, and accountability frameworks are not Phase 2 considerations. They are design constraints that shape what the technology can do and how it must be deployed. A pilot that operates outside governance is not testing the thing that will be deployed; it is testing something else entirely.

Success criteria set by the business, not the pilot team. The measures that matter are the measures that would justify ongoing operational expenditure: cost per transaction, error rates against existing processes, time to resolution measured end-to-end — not just the AI-assisted step — and user adoption rates among the target population, not the volunteer population.

A verification architecture, not manual review. The pilot should be designed as much to test verification methods — automated checks, sampling strategies, confidence scoring,

human-in-the-loop protocols that can scale — as to test the generative capability itself. A model that produces good outputs but cannot be verified at scale is not production-ready, regardless of its accuracy on the pilot dataset.

A production-first pilot does not ask whether the technology works. It asks whether the organisation can deploy it — under real data conditions, real governance constraints, and real operational accountability — and it treats the answer "not yet" as the most valuable finding the experiment can produce.

This approach is harder, slower, and less likely to produce impressive demonstration results. It will surface problems that the conventional pilot is designed to keep hidden. It will generate findings that are uncomfortable — about data quality, about organisational readiness, about the gap between the promised value and what the technology can currently deliver under production constraints.

But it will produce something the conventional pilot cannot: actionable knowledge about whether and how the technology can be deployed. And it will break the equilibrium, because the findings from a production-first pilot cannot be quietly shelved. They demand decisions — investment or cancellation, redesign or acceptance of limitations — that the comfortable ambiguity of pilot purgatory exists to avoid.

8. The Question We Are Avoiding

The deeper issue beneath the pilot purgatory phenomenon is one we have been reluctant to articulate clearly: for many of the use cases currently in pilot, we do not yet know whether generative AI delivers net value under production conditions. We know it is impressive. We know it can perform remarkable feats of language processing. We know it generates outputs that look right. We are considerably less certain that it generates outputs that are right, reliably, at scale, in the specific contexts where enterprises need to deploy it.

The pilot, in its current form, allows us to defer this question indefinitely — to maintain the position that the technology is promising without ever subjecting that promise to a definitive test. This is comfortable. It is also expensive, not in the direct cost of running pilots, which is typically modest, but in the opportunity cost of organisations that are neither deploying the technology nor reallocating the resources to alternatives that might deliver more certain value.

Escaping pilot purgatory does not begin with better technology, better integration, or even better pilots. It begins with the willingness to design experiments that might fail in ways that matter — and to act on what those failures reveal. The proof of concept was never the problem. The problem was building it to succeed.