

BEYOND TIDY VERDICTS

Two Years Into Generative AI — An Honest Balance Sheet

What the Technology Delivered, What Organisations Absorbed, and What
History Predicted All Along

Giovanni Leonardi

January 2025



Contents

1. Executive Summary	3
2. The Quarterly Review That Tells Both Stories	3
3. What Actually Changed	4
4. The Expensive Noise	5
5. The Costs That Migrated	6
6. The Pattern That History Predicted	6
7. Both Things Are True	7

Two Years Into Generative AI — An Honest Balance Sheet

Transformation · No. 337

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

"We are not in a bubble, and we are not in a revolution."

1. Executive Summary

Two years after ChatGPT's release, the enterprise landscape presents a picture that resists simple narration. Genuine, compounding value has emerged in a handful of use cases — code assistance, content drafting, customer service deflection, and research acceleration — while a parallel graveyard of abandoned pilots, unused licences, and organisational restructurings undertaken on faith tells a more sobering story. The pattern is neither triumph nor failure but something more instructive: costs that were supposed to vanish have migrated rather than disappeared, and the organisations extracting real value are those that redesigned workflows rather than simply deployed tools. This is, almost precisely, the trajectory that ERP, cloud computing, and robotic process automation followed before it. The technology has exceeded most reasonable expectations set in late 2022; the organisational capacity to absorb it has run to the schedule that history would have predicted. Both things are true, and the next two years will be shaped by whoever plans on that basis.

2. The Quarterly Review That Tells Both Stories

The scene has become familiar enough to constitute its own small genre. A transformation lead presents the AI portfolio review to the executive committee. The dashboard is impressive: forty-seven pilots initiated, twelve moved to production, adoption metrics climbing. The technology demonstrations still carry a faint charge of the theatrical — a contract summarised in seconds, a customer query resolved without human intervention, a code module generated from a description. The numbers are genuine. The enthusiasm is earned.

Then the Chief Financial Officer asks the question that changes the temperature in the room. Where, exactly, is the impact in the P&L;? Not in the pilot metrics, not in the productivity surveys, not in the time-saved calculations that multiply thirty minutes by ten thousand employees and arrive at figures that nobody can find in the accounts. Where is it in headcount, in revenue, in margin?

The silence that follows is not, in most cases, the silence of failure. It is the silence of a gap — between what the technology can demonstrably do and what the organisation has

managed to absorb. That gap is the subject of this essay, because understanding it honestly is the precondition for closing it.

3. What Actually Changed

Let us start with what is real, because the temptation to cynicism is as misleading as the temptation to hype, and cynicism has recently become the more fashionable pose.

Code assistance is the clearest win, and it is not trivial. Across the software organisations we can observe — large financial institutions, consultancies, technology firms — the pattern is consistent: developers using AI-assisted coding tools report productivity gains in the range of fifteen to thirty per cent on well-defined tasks. The important qualifier is well-defined. The gains concentrate in boilerplate generation, test writing, documentation, and the translation of clear specifications into working code. They thin out markedly in architectural decisions, debugging complex systems, and anything requiring deep understanding of the business domain. But the concentration in routine tasks is itself valuable, because routine tasks consume a remarkable proportion of a developer's week, and freeing that time for higher-order work is a genuine structural improvement — provided the organisation actually redirects the freed capacity rather than simply absorbing it into the same output at marginally lower effort.

Content drafting tells a similar story with a different shape. Marketing teams, internal communications functions, legal departments producing first drafts of standard agreements — all report meaningful acceleration. A first draft that once took a day now takes an hour. The draft is rarely publishable without significant revision, but the revision demands a qualitatively different effort from the original creation.

“The revision of a draft is a fundamentally different cognitive task from the creation of one.”

The organisations that have captured this value share a common characteristic: they redesigned the workflow around the new capability rather than inserting the tool into the existing process. The team that uses generative AI to produce a first draft and then applies the same editorial process as before captures perhaps twenty per cent of the potential value. The team that restructures its editorial workflow — shorter review cycles, different skill mix, revised quality gates — captures something closer to sixty.

Customer service deflection is the third clear win, though it carries its own complications. Organisations that have deployed AI-powered front-line service — in insurance, telecommunications, retail banking — report deflection rates that genuinely surprised even optimistic projections, reaching forty to fifty per cent of total volume for certain query categories in the strongest implementations. The complication is that the remaining queries — the ones that still reach a human — are disproportionately complex, emotionally charged, or ambiguous. The human agents who handle them need to be more skilled, not less, and the training, tooling, and support structures around them need to be better than before. The cost of the human layer has not vanished; it has changed shape.

Research acceleration is perhaps the least visible win but potentially the most consequential. Analysts, strategists, and policy professionals who use large language models as research assistants — to survey literature, synthesise source material, identify patterns across large document sets — report a qualitative shift in what is possible within a given time frame. A regulatory impact assessment that would have taken three weeks now takes one, not because the model writes the assessment but because it accelerates the synthesis of source material to a degree that changes the economics of the exercise. The shift manifests not as the same work done faster but as different work becoming feasible.

These four use cases share three characteristics that explain both why they work and why so many other use cases do not:

- They are augmentation, not automation — the human remains in the loop, and the value comes from changing the nature of their work rather than eliminating it.
- They involve tasks with rapid feedback cycles — you can tell quickly whether the generated code compiles, whether the draft is usable, whether the customer query was resolved.
- The organisations that have captured real value from them have redesigned the surrounding workflow, not simply added a tool to an unchanged process.

4. The Expensive Noise

Against this genuine progress sits a landscape of expenditure that has produced remarkably little. The pilot graveyard is now large enough to merit its own taxonomy.

There are the proof-of-concept orphans — demonstrations built to secure executive sponsorship that succeeded in that narrow objective and then discovered they had no path to production. The demo worked; the data pipeline didn't exist; the integration cost was three times the pilot budget; the compliance review would take six months. The pilot is declared a success, the production deployment is quietly deferred, and the team moves on to the next demonstration.

There are the copilot shelfware installations — enterprise licences purchased at scale on the assumption that ubiquitous access would generate ubiquitous adoption. The adoption curves tell a consistent story: an initial spike of curiosity, a settling period, and then a steady state in which perhaps fifteen to twenty-five per cent of licence holders use the tool regularly, a further twenty per cent use it occasionally, and the remainder have functionally abandoned it. The per-seat economics that justified the purchase assumed adoption rates that have not materialised, and in many organisations the AI licensing cost has become a new line item sitting alongside the productivity suite cost it was supposed to reduce.

And there are the reorganisations chasing vibes — structural changes to teams, functions, and reporting lines undertaken not on the basis of demonstrated capability but on the conviction that AI would transform the function and the organisation should get ahead of the curve. These reorganisations have consumed management attention, disrupted working teams, and in several cases created new roles — Chief AI Officer, Head of AI Transformation, AI Centre of Excellence lead — that sit uncomfortably between the

technology function and the business, with unclear mandates and authority that depends entirely on continued executive enthusiasm.

The total expenditure across these categories is significant. A mid-sized financial institution might reasonably have spent fifteen to twenty-five million on AI initiatives across 2023 and 2024 — licensing, infrastructure, consultancy, internal team costs — and find itself able to point to perhaps three to five use cases delivering measurable value, none of them transformative in isolation.

5. The Costs That Migrated

The narrative of AI-driven cost reduction rested on an assumption that has proven more fragile than it appeared: that the technology would eliminate costs rather than relocate them.

Verification cost is the most pervasive. Every use case that involves generated content — code, text, analysis, customer communication — requires human verification, and that cost is not negligible. A developer reviewing AI-generated code must understand it as thoroughly as if they had written it themselves, perhaps more so, because the code may contain plausible-looking errors that would not survive the iterative thinking process of manual authoring. A lawyer reviewing an AI-drafted contract clause must check not only for accuracy but for the subtle misalignment of terms that a model trained on thousands of agreements might reproduce from a different jurisdiction or a superseded precedent.

Data readiness cost has emerged as the most significant hidden expense. Organisations that assumed their existing data infrastructure would support AI deployment have discovered that the gap between data that supports reporting and data that supports model-driven workflows is wider than anticipated. Data quality, consistency, access controls, lineage tracking, and the governance structures around all of these require investment that was not in the original AI budget but without which the AI initiatives cannot deliver their promised value.

Governance overhead is the third migrated cost. The question of what an AI system should and should not be permitted to do — in customer interactions, in content generation, in decision support — requires governance structures that most organisations did not have and are now building from scratch. These are not temporary setup costs; they are ongoing operational requirements that will grow as the deployment footprint expands.

The costs did not vanish. They migrated — from the obvious and budgeted to the hidden and structural.

6. The Pattern That History Predicted

None of this should surprise anyone who has lived through a previous technology absorption cycle, and that is perhaps the most important observation in the entire balance sheet.

The technology exceeded expectations. The organisational absorption ran precisely to historical schedule. Both things are true, and the next two years belong to whoever plans on that basis.

The pattern is recognisable from enterprise resource planning in the late 1990s and early 2000s: a technology that genuinely worked, vendor promises that ran ahead of organisational reality, implementation costs that exceeded projections by factors of two to five, and value that ultimately arrived — but arrived through business process redesign rather than technology deployment, and on a timeline measured in years rather than quarters. I have written about this pattern before, and the rhyme is close enough to be instructive. The organisations that extracted lasting value from ERP were not those that implemented the technology fastest or spent the most; they were those that used the implementation as the forcing function for the process redesign they should have undertaken regardless.

Cloud computing followed a variant of the same trajectory. The technology was real. The promise of elastic infrastructure and reduced capital expenditure was genuine. But the organisations that simply lifted and shifted existing applications captured a fraction of the potential value, while those that re-architected their applications and operating models for cloud-native delivery captured something closer to the full promise — and did so over five to seven years, not the eighteen months the initial business cases projected.

Robotic process automation offered perhaps the closest parallel to where we stand now with generative AI. The technology worked. The individual automations delivered measurable time savings. But the scaling story — the promise of enterprise-wide automation portfolios — foundered on precisely the problems now appearing in AI: process variation that exceeded the technology's tolerance, data quality issues that the manual process had been quietly compensating for, and an organisational change challenge that was consistently underestimated.

We are running to a schedule that history would have predicted with reasonable accuracy. Large language models in early 2025 are more capable, more reliable, and more versatile than almost anyone predicted in late 2022. The gap is not in what the technology can do. The gap is in what organisations can absorb, and that absorption runs on its own clock — a clock set by change management capacity, data infrastructure maturity, process redesign willingness, and governance readiness, not by model capability.

7. Both Things Are True

The temptation, at the two-year mark, is to resolve the balance sheet into a single verdict. The technology optimists want to declare victory and argue that the laggards simply haven't caught up yet. The sceptics want to declare the whole thing a bubble and wait for the correction. Both are wrong, and both are wrong in ways that will cost the organisations that listen to them.

The optimist's error is to treat the absorption gap as a temporary inconvenience — a lag that will close as the technology improves and organisations finally catch on. This

mistakes the nature of the problem. The absorption gap is not a technology problem; it is an organisational design problem, and organisational design problems do not solve themselves through better tools. They solve themselves through the slow, expensive, unglamorous work of redesigning processes, retraining people, rebuilding data infrastructure, and establishing governance frameworks. The technology will continue to improve. But improvement in the technology will not, on its own, close the gap between what it can do and what organisations can absorb.

The sceptic's error is to mistake the absorption lag for evidence that the technology does not work. It does work. The four use cases delivering real value today are not trivial, and they are compounding. Code assistance alone, scaled across a large software organisation, represents a structural shift in developer productivity that will compound over years. Service deflection, in organisations that have redesigned their service model around it, represents a genuine step change in the economics of customer service. The sceptic who dismisses these because the broader transformation story has not materialised is confusing the pace of organisational change with the merit of the underlying capability.

Both things are true. We are not in a bubble, and we are not in a revolution. We are in the early years of a technology absorption cycle that will play out over the remainder of this decade, and the organisations that will extract the most value are those that plan on that basis — investing in the workflow redesign that turns tool deployment into genuine capability change, building the data infrastructure and governance frameworks that the technology requires, measuring progress in terms of workflow outcomes rather than pilot counts, and having the institutional patience to let the absorption run at its natural pace.

The next two years will not be defined by what the technology can do. They will be defined by whether organisations learn to absorb it — and the honest balance sheet of the first two years, with all its uncomfortable ambiguity, is the best guide we have to what that absorption will require.