

BEYOND THE DEMO

The Verification Problem Nobody Is Pricing In

Why the Real Enterprise Cost of Generative AI Lies in Verification, Not Generation

Giovanni Leonardi

June 2023



The Verification Problem Nobody Is Pricing In

Transformation · No. 353

Originally written June 2023 | Re-edited and re-rendered for the WhereBeyond Edition, 2025

“Fluency is not the same as reliability, and in the enterprise, the gap between the two is where the cost actually lives.”

The Spectacle and the Silence Behind It

The board deck has a new slide. Six months ago it was not there; now it appears in every strategy review, every quarterly briefing, every risk committee paper. Generative AI. The demonstrations are impressive — a draft policy document produced in seconds, a market analysis summarised in a paragraph, code written from a natural-language description. The technology has arrived with a velocity that has caught most enterprises genuinely off guard.

And yet, beneath the spectacle, something conspicuous: almost nothing has actually changed in how work gets done.

The pattern across large organisations is remarkably consistent. There is excitement — and there are bans. Internal memos permitting use, then restricting it. Pilot projects announced, then quietly shelved. Innovation teams producing demos that never reach production. The conversation is dominated by two questions — should we use this, and will it replace people? Both are understandable. Both are the wrong place to start.

The Question That Matters Is Verification

The capability is real — nobody who has spent time with a large language model can credibly deny that. These systems produce fluent, structured, contextually aware text at a speed and cost that genuinely has no precedent. The problem is not capability. The problem is that fluency is not the same as reliability, and in the enterprise, the gap between the two is where the cost actually lives.

When a system generates a contract summary that reads convincingly but misrepresents a liability clause, someone must catch that. When it produces a regulatory analysis that is well-structured but omits a material requirement, someone must notice. When it drafts client correspondence that is tonally perfect but factually wrong, someone must verify every claim before it goes out.

That someone is almost always an expert — the same expert whose time the technology was supposed to free up.

This is the verification problem, and it is the cost nobody is pricing in. The demonstrations that circulate at board level show the generation. They never show the checking. A three-minute demo of a policy draft being produced in seconds omits the forty-five minutes a senior policy analyst then spends reading it line by line, correcting the confident errors, and satisfying themselves it can carry their name. In workflows where verification is cheap — where a quick glance confirms accuracy — the technology delivers immediate value. In workflows where verification requires deep expertise, the savings evaporate, and in some cases the total cost increases: the expert now spends time checking fluent output that is harder to distrust precisely because it reads so well.

Not All Language Work Is Equal

The enterprise contains an enormous volume of language work. But the assumption that generative AI can condense all of it is a category error. The workflows where the technology is already proving useful share common characteristics: the output is low-risk, the domain is well-bounded, the cost of an error is small, and the person reviewing can verify quickly. Internal first-draft summaries. Brainstorming structures. Routine correspondence where the stakes are low.

The workflows where it is dangerous — or simply uneconomic — share different characteristics: the output carries regulatory, legal, or reputational weight; the domain is complex and contextual; errors are difficult to detect without domain expertise; and the consequences of a missed error are severe. These are precisely the workflows where the productivity promise is loudest and the verification cost is highest.

The honest question for any enterprise is not can generative AI do this work? — in most cases, it can produce something that looks like the work. The question is: what does it cost us to verify that the output is trustworthy, and is that cost lower than doing the work ourselves?

In my experience, very few organisations are asking this question yet. The conversation is still captivated by the generation.

The Discipline Gap

What is missing is not scepticism — there is plenty of that — but discipline. The operational discipline to distinguish workflows where verification is cheap from workflows where it is expensive. The governance discipline to define who owns the risk when output from a system that cannot explain its reasoning enters a decision chain. The cultural discipline to resist the pressure to adopt because the technology is spectacular, and instead adopt where the economics and the risk profile actually work.

Six months in, the pattern is clear. The capability has arrived ahead of the discipline for absorbing it. That gap is not unusual — it recurs with every significant technology shift — but the speed of this arrival has made it unusually wide. The organisations that will extract real value are not the ones producing the most impressive demos. They are the ones doing the unglamorous work of mapping their language workflows, pricing

verification honestly, and building the governance to know when fluent output can be trusted and when it cannot.

The spectacle will continue. The discipline is what matters.