

BEYOND BORROWED COGNITION

The Model You Did Not Build

Large language models are arriving in regulated institutions faster than the governance built to assure them — a reflection on where innovation and compliance actually collide.

Giovanni Leonardi

May 2022



Contents

1. Executive Summary	3
2. The fluent summary that cited a rule which did not exist	3
3. Why the pull is real, and not merely fashion	4
4. Three collisions, not one	4
5. The uncomfortable table	5
6. Why the pattern persists: two functions, two epistemologies	6
7. The strongest case for relaxing, and why it is only half right	7
8. What the longer view actually asks of us	7
9. A reflection, not a verdict	8

The Model You Did Not Build

Transformation · No. 279

Originally written May 2022 | Re-edited and re-rendered for the WhereBeyond Edition, 2025

*"The danger is not that the model gives no explanation;
it is that it gives an excellent one that no one should
trust."*

1. Executive Summary

Large language models have crossed a threshold that regulated institutions can no longer treat as somebody else's experiment. For the first time, a single general-purpose system can read a policy, summarise a case file, draft a customer letter and answer an open question in fluent, plausible prose — and it can do so without having been built, line by line, for any of those tasks. The capability is genuinely new, and the institutions that handle the most regulated text of all — banks, insurers, asset managers — are precisely the ones with the most to gain from it.

The argument of this essay is that the difficulty these institutions now face is not, at root, a technology problem. It is a governance problem, and a subtle one. The machinery that regulated firms use to assure their models — independent validation, documented assumptions, reproducible outputs, a traceable line from input to decision — was built for a world of models the firm commissioned, understood and could interrogate. A large language model breaks each of those assumptions at once. It is probabilistic where the assurance machinery expects determinism; it is opaque where the duty to explain expects transparency; and, most disruptively, it is borrowed rather than built — a capability the firm rents and adapts rather than one it designs and owns.

The comfortable response is to say that regulated industries always refuse first and adopt later, so the task is simply to wait for the controls to catch up. There is truth in that, and this essay takes the objection seriously. But it understates what is different this time. The gap is not between a new tool and an old procurement process; it is between a new kind of decision-making substrate and an assurance vocabulary that has no settled word for it. The firms that adopt well will be the ones that stop asking whether they are permitted to use these models and start asking what validated can possibly mean for a model they did not build — and are prepared to rebuild their assurance around the answer.

2. The fluent summary that cited a rule which did not exist

Picture a model risk committee, midway through an otherwise unremarkable agenda. The item is a small, sensible-looking pilot: a large language model, reached through an interface a handful of engineers stood up over a few weeks, that reads a customer's

suitability file and produces a one-paragraph summary for the adviser who will speak to them. The demonstration is impressive. The summaries are quick, readable, and – to anyone skimming – right. Then someone on the second line, whose job is to be unimpressed, reads one closely and notices that it justifies a recommendation by referring to a specific clause of the conduct rules. The clause does not exist. The number is plausible, the wording is plausible, the confidence is total, and the citation is invented.

Nothing about that moment is exotic. It is, in fact, the entire problem in miniature. The model was not lying, because it has no notion of truth to depart from; it was doing exactly what it was built to do, which is to produce the most probable continuation of the text in front of it. A fabricated but plausible rule is, statistically, an excellent continuation. What made the room go quiet was not the error itself – every system makes errors – but the kind of error, and the realisation that the committee had no established way to size it, bound it, or sign it off. The firm's assurance machinery had met something it was not designed to hold.

I have sat through enough of these meetings to recognise the particular silence that follows. It is not the silence of people who have found a bug. It is the silence of people who have realised that their questions no longer fit the object in front of them.

3. Why the pull is real, and not merely fashion

It would be easy, and wrong, to dismiss the interest as hype. Regulated institutions are, above almost all other organisations, machines for processing text. A bank's working day is policies, procedures, product terms, suitability letters, complaint files, know-your-customer packs, regulatory correspondence and the endless internal prose that binds them together. Much of the genuinely expensive work in these firms is not analytical in the mathematical sense; it is linguistic. It is reading a long document and saying what it means. It is drafting the same category of letter for the ten-thousandth time. It is finding the three files out of three hundred that do not say what they should.

For twenty years the industry has thrown rules-based automation and, more recently, narrower statistical language tools at this problem, with real but bounded success. What is different about the current generation of models is that a single system, without bespoke engineering for each task, can absorb an instruction in plain language and act on unfamiliar text. The economic logic is therefore not imaginary. If even a fraction of the linguistic labour of a large institution can be compressed, the prize is very large, and no serious executive is going to walk past it because the first pilot cited a rule that did not exist.

This is worth stating plainly because the temptation, inside a compliance culture, is to treat caution as an end in itself. It is not. The purpose of the second line is not to prevent adoption; it is to make adoption survivable. A firm that refuses this capability outright is making a decision every bit as consequential as one that adopts it recklessly, and it should be made just as deliberately.

4. Three collisions, not one

When a large language model meets the assurance machinery of a regulated firm, the friction is usually described as a single vague problem — "the model is risky." That framing is useless because it hides three distinct collisions, each of which fails a different control, and each of which demands a different answer.

The first collision is between the probabilistic and the auditable. A regulated model is expected to be reproducible: the same inputs should yield the same output, or at least a distribution whose behaviour is characterised and stable. Much of the machinery of validation — back-testing, benchmarking, monitoring for drift — quietly assumes this. A generative model, sampled in the ordinary way, may give two different answers to the same question on two consecutive mornings. Neither is wrong; both are draws from a distribution. But an auditor asking "show me that this decision would be made the same way again" is asking a question the system cannot answer in the terms the question expects.

The second collision is between opacity and the duty to explain. Regulated decisions frequently carry an obligation to say why. A declined application, an assessment of suitability, a judgement about a customer's circumstances — these come with a requirement, sometimes legal and sometimes merely reputational, to give reasons a human can understand and, if necessary, contest. A large language model can produce something that reads like a reason. But the text it emits after the fact is a plausible rationalisation, generated in the same way as everything else it produces, not a faithful account of a causal process inside the model. The danger is not that the model gives no explanation; it is that it gives an excellent one that no one should trust.

The third collision is the deepest, and the least discussed: the model is borrowed, not built. The entire discipline of model risk management assumes that the firm — or a vendor acting transparently on its behalf — designed the model, chose its variables, understands its assumptions and can document them. A foundation model inverts this. It arrives already trained on a corpus the firm did not select and cannot fully inspect, shaped by choices the firm did not make and often cannot see. The institution adapts it at the edges — with instructions, with examples, with retrieval of its own documents — but the core capability is rented. The comfortable fiction that a firm "owns and understands its models" simply does not survive contact with this arrangement.

The first two collisions are about the model's behaviour and can, in principle, be managed with familiar tools stretched to fit. The third is about the model's provenance, and it is the one for which the existing vocabulary has no word. A firm can wrap a probabilistic, opaque model in controls. It cannot, by wrapping, come to understand a model it did not build.

5. The uncomfortable table

It helps to lay the expectation against the reality directly, because the mismatch is easy to feel and hard to state.

What model governance was built to expect	What a large language model actually offers
A model the firm commissioned and specified	A general capability the firm adapts at the edges
Deterministic, reproducible outputs	Probabilistic outputs sampled from a distribution
A documented, inspectable set of input variables	A training corpus the firm did not choose and cannot fully see
Explanations that trace the causal path to a decision	Fluent after-the-fact text generated like any other output
Failure modes that are bounded and characterised	Failure modes that are open-ended and confidently expressed
Validation as a one-time gate before deployment	Behaviour that shifts with prompt, context and version

Read the left-hand column and you are reading the assumptions of a mature, hard-won discipline — one that exists because regulated firms have, in living memory, been badly burned by models they trusted too much. Read the right-hand column and you are reading a description of a genuinely useful technology. The essay's whole point lives in the gap between the columns. The instinct of a good compliance function, confronted with that gap, is to force the right column to behave like the left. Some of that is right and necessary. But some of it is a category error, and telling the two apart is the real work.

6. Why the pattern persists: two functions, two epistemologies

If the collision were merely technical, it would already be closing, because technical problems attract engineers and engineers are drawn to friction. The reason the pattern persists — the reason firms will still be circling this in a year, and the year after — is that it sits on a fault line inside the organisation itself.

The innovation function and the assurance function do not merely have different incentives; they have different epistemologies. They disagree about what it means to know that something works. To the team building the pilot, the model works because it produces good outputs across a wide range of realistic cases: knowledge is empirical, demonstrated, and provisional. To the second line, a model works when its behaviour is understood well enough to be bounded and defended to a regulator: knowledge is structural, documented, and, ideally, complete. These are not two attitudes to the same idea of proof. They are two different ideas of proof.

Most of the failures I have watched were not failures of technology or even of policy. They were failures of translation between these two ways of knowing. The engineers presented evidence the risk function could not accept as evidence; the risk function asked for guarantees the engineers could not give and, often, did not understand the point of. Each concluded, privately, that the other was being unreasonable. Both were behaving exactly as their discipline required.

"An institution that cannot translate between how its builders know things and how its assurers know things will not be stopped by a large language model. It will simply be unable to decide about one."

7. The strongest case for relaxing, and why it is only half right

The most serious objection to everything above runs like this. Regulated industries have a long, unbroken record of declaring each new technology impossible to govern and then governing it perfectly well within a few years. Statistical credit models, algorithmic execution, cloud infrastructure, model-driven capital calculation — each arrived to choruses of "the regulators will never allow it," and each is now routine, wrapped in controls that did not exist when the technology appeared. On this reading, large language models are simply the next entry in the list. The controls will be invented, the validation frameworks will be extended, the pilots that cite imaginary rules will be caught by a human check, and in due course this essay will read as quaint. The right posture, the objection concludes, is confident patience: adopt narrowly, keep a human in the loop, and let the assurance machinery evolve as it always has.

This is the best argument against alarm, and it is largely correct. It would be a mistake to treat these models as uniquely ungovernable; almost nothing is. But the analogy conceals two things that make this case genuinely harder than its predecessors, and honesty requires naming them.

The first is that "keep a human in the loop" is a stronger control on a slide than in a chair. A human reviewing a stream of fluent, usually-correct machine output does not stay vigilant; they habituate. The very quality that makes the model valuable — that it is right often enough to be worth using — is what erodes the reviewer's attention until the check becomes a rubber stamp. Every prior wave of automation faced some version of this, but few produced output so fluent and so confidently wrong at the same time. The human-in-the-loop is real protection only if the institution treats the human's attention as a resource that depletes, and designs against that depletion — which almost none currently do.

The second is that the earlier technologies, however novel, were still things the firm built or bought transparently. A statistical credit model is inspectable. A cloud platform is contracted and configured. A foundation model is neither owned nor fully knowable, and the borrowed nature of the capability does not dissolve with familiarity. So the objection is right that the controls will come, and wrong to assume they will be the same controls stretched a little further. Some of them will have to be new, because the thing being governed is new in a way the reassuring precedents were not.

8. What the longer view actually asks of us

Step back far enough and the question stops being "can regulated firms use large language models" — the answer to which is obviously yes, and increasingly, at the edges, already happening — and becomes something harder. The longer view asks what these institutions will have to change about their idea of assurance itself.

Three shifts seem to me unavoidable, and none of them is a tool.

1. Validation has to become continuous, not ceremonial. The one-time gate before deployment — the artefact, the sign-off, the filing — assumes a static object. A model whose behaviour moves with its prompt, its retrieved context and its underlying version is not a static object, and a validation performed once describes a system that no longer exists. Assurance has to shift from an event to a standing capability: monitoring live behaviour, not certifying a frozen snapshot. Firms know how to do this for market risk. They do not yet think of model assurance the same way.
2. The unit of governance has to move from the model to the task. Asking "is this model safe" is close to meaningless, because the same model may be entirely appropriate for drafting an internal summary a human will rewrite and entirely inappropriate for producing a customer-facing reason for a decision. The meaningful question is whether this capability, in this specific use, with this specific human arrangement around it, is safe. That reframing sounds administrative. It is in fact the whole game, because it moves the firm away from a doomed attempt to certify a general capability and towards governing the concrete situations in which it acts.
3. The firm has to become honest about what it does not, and cannot, know. The borrowed model will not become fully transparent because the institution wishes it. Mature governance of these systems will rest less on the fiction of complete understanding and more on the disciplined management of acknowledged uncertainty — bounding what the model touches, instrumenting what it does, and keeping the most consequential decisions on the far side of a line it is not allowed to cross. This is uncomfortable for a culture whose self-image is built on control. It is also the only honest place to stand.

None of these is a product a vendor can sell, which is precisely why they are hard. They are changes to how an institution thinks about knowing, defending and deciding — and institutions change those things slowly, under pressure, and rarely by choice.

9. A reflection, not a verdict

It is tempting, at the end of a piece like this, to resolve the tension — to declare either that the caution is overdone or that the enthusiasm is reckless. I do not think the honest position allows it. Both instincts are responding to something real. The enthusiasm is right that a capability of unusual value has arrived and that the institutions with the most regulated text have the most to gain. The caution is right that the assurance machinery, as it stands, cannot yet hold what the capability is.

What I am fairly sure of is that the firms which navigate this well will not be the ones with the cleverest pilots or the boldest risk appetite. They will be the ones that treat the arrival of these models as an occasion to modernise their idea of assurance rather than to defend the old one or abandon it. The pilot that cited a rule which did not exist was not a reason to stop. It was a small, early, almost gentle warning that the questions in the room had

stopped fitting the object on the table — and an invitation, if anyone was willing to hear it, to find better questions before the stakes got higher.

That, in the end, is the longer view. The technology will keep improving; it always does. The interesting variable is not the model. It is whether the institution is willing to change how it knows things fast enough to deserve the capability it is about to depend on.