

BEYOND EXPLAINABILITY

The Barrier Is Not the Model — Why Large Language Models Break the Compliance Playbook in Regulated Industries

Adopting language models in banking, insurance and healthcare when every control we own assumes a model we can reproduce and explain

Giovanni Leonardi

March 2022



The Barrier Is Not the Model — Why Large Language Models Break the Compliance Playbook in Regulated Industries

Transformation · No. 333

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

“The models are already more capable than our processes can absorb.”

1. The pilot that dies in the validation committee

Every regulated firm now has a version of the same meeting. A working group brings a large language model to a steering committee, and the demonstration is genuinely impressive. Fed a thirty-page credit pack, the system returns a clean, largely accurate summary in seconds. Asked to draft the rationale behind a suitability decision, it produces in a moment something a junior analyst would have spent an afternoon assembling. The room is enthusiastic; someone uses the word transformative. Then the paper goes to model validation, and the questions begin. Show us the same output twice. Explain why it flagged this covenant and not that one. Point to the source behind this particular sentence. Reproduce the result for the audit file, exactly, in eighteen months, when a regulator asks. The working group cannot answer any of them, and the pilot quietly dies — not because it failed, but because it succeeded in a way our control apparatus has no language for.

We are telling ourselves that the barrier to adopting these models in banking, insurance and healthcare is that the technology is not yet ready. That is the wrong diagnosis, and a comfortable one, because it locates the problem safely on the vendor's side of the table. The truth is closer to the opposite. The models are already more capable than our processes can absorb. The barrier is not the model. It is us — the evidentiary machinery we built over two decades on an assumption these systems quietly violate.

2. A control apparatus built for a different kind of model

Everything a regulated firm does to govern a model rests on four quiet assumptions: that the model can be specified (we can write down what it does), reproduced (the same inputs give the same outputs), explained (we can attribute an output to identifiable factors), and validated against history (we can backtest it and bound its error). The whole edifice of model risk management — the lineage that runs back to the supervisory guidance the industry has leaned on since 2011 — is built on those four legs. So is the audit file. So is the conversation a firm expects to have with its regulator.

A conventional model earns its place by satisfying all four. A credit scorecard can be documented coefficient by coefficient. A pricing model can be run a thousand times and return the same number a thousand times. When it is wrong, we can usually say why it was wrong and adjust it. A large language model satisfies none of the four cleanly, and pretending otherwise is where most firms are quietly going astray.

The uncomfortable recognition is not that these models are risky. It is that they are risky along axes our second line was never designed to measure.

3. "Put it through model risk management like anything else"

The most reasonable objection I hear runs like this: we already govern models; a language model is a model; run it through the same pipeline, apply the same tiering, demand the same validation, and stop treating it as exotic. It is a serious argument, and the people making it are usually the most experienced risk professionals in the building. They are right that exceptionalism is dangerous — every technology arrives claiming to be too novel for the existing rules, and most of the time the existing rules are wiser than the enthusiasts.

But the pipeline does not fail here because we are being precious. It fails because it asks questions the object cannot answer. Tell a validation team to reproduce a deterministic model and they will. Tell them to reproduce a generative one and they will discover that the same prompt, run through the same interface on the same afternoon, returns materially different wording perhaps one time in six — and, more troublingly, will occasionally assert a figure that appears nowhere in the source document with the same fluent confidence it brings to the figures that are real. There is no coefficient to inspect. There is no error term to bound. The pipeline does not reject the model; it simply has nothing to grip.

4. Where the machinery actually breaks

It is worth being precise about the points of failure, because "it's a black box" is too lazy a summary and lets everyone nod without changing anything. The machinery breaks at four specific joints.

Contro l assu mptio n	What a conventional model gives	What a language model gives
Reprod ucibilit y	The same inputs return the same output, every time	The same prompt returns varying output; determinism must be engineered back in, and even then is fragile

Control assumption	What a conventional model gives	What a language model gives
Explainability	An output attributable to weighted, inspectable factors	Fluent text with no faithful account of why these words and not others
Data provenance	A curated, documented training set you control	A corpus of unknown composition, licensed from a vendor, that you did not assemble and cannot fully see
Change control	You decide when the model changes	The model can change beneath you when the vendor updates it behind the same interface

The fourth joint is the one that is least discussed and most dangerous. When a firm consumes one of these models as a service, the thing it validated in March is not guaranteed to be the thing answering customers in September. The vendor improves the model; the behaviour shifts; the validation evidence in the file now describes a system that no longer exists. Every other model in the estate is frozen the moment it is approved. This one is not, and our change-management processes have no category for a model that revises itself without a release note.

I have watched a document-summarisation assistant cut the first-pass review of a credit paper from roughly forty minutes to under ten — a real, measured saving that a hard-pressed team felt immediately. And I watched the same tool, three weeks later, confidently summarise a covenant that the underlying pack did not contain. The saving was genuine. So was the fabrication. Both are properties of the same system, and any honest account has to hold them together rather than choosing the one that suits the argument.

5. "Then wait until the technology is explainable"

If the first objection is treat it as ordinary, the second is its mirror: treat it as premature. Explainable AI is a live field; the regulators are circling; there is an AI regulation working its way through Brussels that will, in time, force the vendors to make these systems auditable. So wait. Let the technology mature into something our frameworks can accept, and adopt then, from a position of safety.

The instinct is prudent and the conclusion is mistaken. Waiting is not the absence of a decision; it is a decision to forgo the bounded-risk uses while bearing all the costs of falling behind. And the premise — that faithful explanation of a model with tens of billions of parameters is arriving on a near horizon — is more hope than forecast. The research community has been candid that we do not have a reliable method for making these systems explain themselves truthfully; the plausible-sounding explanation a model gives for its own answer is itself just more generated text, not a window into its reasoning. A firm that waits for genuine explainability may be waiting a very long time, and will spend that time watching less cautious competitors learn where these tools are safe.

“The choice is not between a compliant model and a risky one. It is between rebuilding our standard of evidence and pretending we do not have to.”

6. Drawing the line in a different place

The firms I would bet on are not the ones with the most capable model or the strictest prohibition. They are the ones that have stopped asking is this model compliant? — a question the model cannot pass — and started asking a better one: where can a non-deterministic system touch a regulated decision such that the cost of being wrong is bounded and a human remains unambiguously accountable?

That reframing does most of the work, because it moves the control from the model, where it cannot live, to the surrounding design, where it can. In practice it points in a consistent direction:

- Assist, do not decide. Let the model draft the suitability rationale that a qualified adviser then owns and signs; do not let it make the suitability determination. The accountable human is the control, and the model is a faster pen.
- Start where the error is cheap and visible. Internal knowledge retrieval, first-pass document review, and drafting are forgiving because a person reads the output before it matters. Customer-facing, decision-bearing uses are not, and should come last, if at all.
- Make the record the prompt and the output, not the model. Since we cannot reconstruct the reasoning, we log the exact input and the exact generated text as the auditable artefact, along with the version of the vendor model in force at that moment — so that the file describes what actually happened, even though it cannot explain why.
- Validate behaviour, not internals. Build a standing evaluation harness — a battery of representative cases the model is re-run against on every vendor change — and treat a shift in those results as the equivalent of a model change, because it is one.

None of this appears in the model risk playbook as written. All of it is buildable now, with the technology exactly as it stands, by people who accept that the object in front of them is probabilistic and design accordingly.

7. What this asks of us

There is a temptation, in a regulated firm, to believe that caution and inaction are the same thing. They are not. The genuinely cautious position, faced with a technology this capable and this strange, is not to wait for it to become the kind of model our processes already understand. It is to admit that our processes encode an assumption — that a model is a thing you can reproduce and explain — which no longer holds, and to do the unglamorous work of building a second evidentiary standard for the systems that break it.

That is harder than a moratorium and less exciting than a pilot. It asks the second line to develop muscles it does not yet have, and it asks the first line to give up the comfort of a determinism that was always partly an illusion. But the alternative is the meeting we started with, repeating itself indefinitely: a capable system, an impressed room, and a control apparatus that can only say reproduce it, explain it, attribute it — and, hearing no answer, quietly let the future die in committee while the firms that rebuilt their standard of proof get on with it.