

BEYOND THE DEMO

The Model Is Free, the Data Is Not — GenAI's Real Bill Arrives

Why enterprise retrieval programmes stall on ownership, duplication and permissions — and the sequence that gets them moving

Giovanni Leonardi

April 2024



The Model Is Free, the Data Is Not — GenAI's Real Bill Arrives

Transformation · No. 375

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

*“Embedding a corpus you have not curated is indexing
your confusion at scale.”*

1. The Demo That Worked and the Pilot That Did Not

The demonstration takes eleven minutes. A general counsel asks the assistant what the standard limitation-of-liability position is for a particular class of supplier contract, and the answer comes back in fluent, sourced paragraphs with the clause quoted and the policy document cited. The room is impressed. The chief executive, who has been told by the board for eighteen months that the organisation needs an AI strategy, finally has something to show.

Four months later the same assistant, now pointed at the organisation's actual document estate rather than the forty curated files used for the demonstration, is quietly being withdrawn from the legal team. It cites a superseded version of the delegations policy. It surfaces a termination letter that a paralegal should never have been able to see. When asked the same question twice it answers from two different contract templates, both labelled final, one of which nobody can date. The model did not change between the demonstration and the withdrawal. The data did.

This is the pattern that has recurred, with remarkable consistency, across the enterprise generative AI programmes of the last year. The technology that was supposed to be the hard part turned out to be the easy part. The part everyone assumed was already solved — knowing what documents the organisation holds, which of them are true, and who is allowed to read them — turned out to be the whole project.

2. The Inversion Nobody Planned For

For most of the last decade the working assumption in enterprise technology was that intelligence would be scarce and data would be plentiful. The strategic anxiety was about talent: data scientists were expensive, models were bespoke, and an organisation's competitive edge would come from its ability to build and tune algorithms that others could not. Data, by contrast, was something organisations believed they had in abundance — too much of it, if anything, sitting in lakes and warehouses and shared drives.

The arrival of capable foundation models in late 2022 and through 2023 inverted that economics almost overnight. The intelligence is now available to anyone with a corporate

card and an API key, at a price that keeps falling. Two competitors calling the same model endpoint receive the same intelligence. Nothing about the model is a differentiator, because nothing about it is yours.

What is yours — the only thing that is yours — is the internal corpus the model reasons over: the contracts, the policies, the engineering reports, the customer correspondence, the thirty years of decisions recorded in documents that nobody has read since the week they were written. And it is precisely here, on the retrieval side of retrieval-augmented generation, that enterprise programmes have been stalling.

The model is a commodity. The data is the asset. The organisations that will get value from generative AI are not those with the cleverest prompts or the earliest licence agreements. They are the ones whose internal knowledge is owned, current, deduplicated and permissioned well enough to be retrieved with confidence.

The reason this matters is mechanical, not philosophical. A retrieval-augmented system does not know what the organisation knows. It knows what it can find. It chunks documents, embeds them, and at query time pulls back the handful of passages that look most similar to the question, then asks the model to answer from those passages. Every weakness in the corpus becomes a weakness in the answer, with a fluency that disguises it. A duplicate document is not a harmless redundancy; it is two competing sources of truth that the retriever will surface at random. A superseded policy that was never archived is not clutter; it is a confident wrong answer waiting for the right question. A permissions model that nobody can explain is not a compliance footnote; it is the mechanism by which a junior analyst reads the board's remuneration papers.

3. Every Retrieval Project Is an Archaeology Dig

We should be candid about what the last year has actually involved for the teams doing this work. It has not been prompt engineering. It has been archaeology.

Consider what a typical retrieval programme discovers in its first six weeks, once it moves from the curated demonstration set to the real estate. The pattern I have observed is consistent enough to be almost predictable, and it is worth setting out in concrete terms because the figures are what convince sceptical sponsors.

A mid-sized professional services organisation — the kind with perhaps four thousand staff and a document management system that was implemented in the late 2000s and migrated to the cloud in the early 2020s — will commonly hold somewhere between two and four million documents in its primary repositories, with an unknown further quantity in team collaboration sites, email attachments and personal drives. When the content is profiled, the findings tend to cluster:

- Between a quarter and a third of documents are near-duplicates — versions, copies saved to a second location, attachments forwarded and re-filed. The word final appears in the filename of an uncomfortable number of documents that are not.
- A substantial share — often more than half — have no identifiable owner. The author has left, the team has been reorganised twice, and the site collection's nominal

owner is a distribution list that no longer resolves to anyone.

- Permissions have drifted over years of ad-hoc sharing. Broken inheritance, direct grants to individuals who have changed roles, and "everyone" groups applied to folders that were never meant to be public. Nobody can explain why a given person can see a given document, because the answer is a sequence of decisions made by people who have gone.
- Metadata is sparse and unreliable. Document type, effective date, supersession and classification are filled in inconsistently or not at all, which means the retriever cannot distinguish the current policy from the three drafts that preceded it.

The numbers vary. The shape does not. And the crucial point is that none of these problems were invisible before generative AI arrived. Records managers have been writing the same findings into the same unread reports for fifteen years. What changed is the stakes. When the consumer of the document estate was a human being searching, a human being also filtered — recognised the superseded draft, noticed the odd permission, asked a colleague which template was current. When the consumer is a retriever feeding a language model, that human filter is removed, and the organisation discovers that it had been relying on it all along.

"The model does not know what the organisation knows. It knows what it can find — and it finds the duplicates, the drafts and the documents nobody meant to share with the same confidence as the truth."

4. The Data Mesh Debates Suddenly Have Stakes

Two years ago the enterprise data community was absorbed in a debate about data mesh — about whether data ownership should sit with the domains that produce it rather than with a central platform team, about data as a product, about federated governance. It was an architecture argument, and like most architecture arguments it was conducted largely among people who already cared. To most executives it registered, if at all, as the latest fashion in a discipline that has had many.

It is worth being honest that the sceptics had a point at the time. The mesh literature was strong on principle and thin on the operating model; it asked domains to take ownership of data without always explaining what ownership would cost them or who would fund it. Many organisations that adopted the vocabulary did not adopt the practice, and the central platform teams quietly carried on as before. Ownership, in 2022, was a diagram.

The retrieval programmes of 2023 and 2024 have converted that diagram into a binding constraint. Every question a retrieval system cannot answer well resolves, on investigation, to an ownership question. Which version of this policy is current? Someone must own the policy to say. Should this document be in the index at all? Someone must own the document to decide. Who may see the answer this system has just assembled

from six sources? Someone must own each source to say what it is and who it is for. The mesh advocates were right that ownership had to move to the domains; what they could not supply in 2022 was a reason for the domains to accept it. Generative AI has supplied the reason, because the board now wants the assistant, and the assistant does not work without it.

This is the sense in which the warning has landed. An organisation that treated data ownership as an optional architectural preference now finds it is the gating item on the one technology its leadership has explicitly asked for. The conversation has moved from should we to how quickly can we, and the answer, uncomfortably, is: not as quickly as the licence was signed.

5. The Strongest Objection, and Why It Does Not Hold

The serious counter-argument deserves a fair hearing, because it is made by thoughtful people. It runs roughly as follows. The models are improving so rapidly that the data problem will solve itself. Longer context windows mean the retriever matters less, because the system can simply read more. Better ranking and re-ranking will separate the current policy from the draft. Models will learn to reconcile contradictory sources. Investing heavily in data readiness now is paying to solve a problem that the next release will dissolve.

There is something to this on the retrieval-quality side. Ranking has improved noticeably in a year, context windows have grown by an order of magnitude, and some of the brittleness of early chunk-and-embed pipelines is already being engineered away. A programme that waits six months will inherit better tooling than one that starts today.

But the argument fails on its own terms for two reasons. First, no improvement in the model resolves a problem that is not a retrieval problem. A model cannot infer which of two contradictory final documents is authoritative, because the organisation itself has not decided. It cannot know that a policy was superseded if the superseding policy was never linked to it. It cannot respect a permission that was never set. These are not failures of intelligence. They are absences of fact, and the only entity that can supply the fact is the organisation.

Second — and this is the part that sponsors most underestimate — the permissions question does not get easier as models get better. It gets worse. A more capable system that reasons across more sources is a more capable system for assembling, from individually innocuous fragments, a conclusion that no one fragment was cleared to reveal. Longer context is a larger aperture. The better the model, the more the organisation needs to be sure of what it is allowed to read.

6. What Readiness Actually Means, and in What Order

If the diagnosis is right, the response is not a data strategy document. It is a sequenced programme that treats readiness as the critical path to the AI outcome the board has asked for, and is funded and governed as such. In practice the sequence that works — as

opposed to the sequence that feels logical — runs as follows.

1. Scope by use case, not by estate. The instinct to clean the whole corpus before doing anything is the instinct that kills programmes. Pick one high-value question set — contract positions, engineering standards, HR policy — and define the corpus that must be trustworthy to answer it. A few thousand documents, not a few million. Readiness is earned domain by domain.
2. Profile before you promise. Run the archaeology on that scoped corpus before anyone commits to a delivery date: duplication rate, ownership coverage, permission anomalies, metadata completeness. Put the numbers in front of the sponsor. A stated duplication rate of thirty-one per cent does more to reset expectations than any amount of caveat in a business case.
3. Assign owners and give them an actual decision to make. Ownership becomes real the day an owner is asked a question with consequences: is this the current version, and may it go in the index? Start with the smallest number of owners who can cover the scoped corpus, and make the decisions they take visible in the metadata — effective date, supersedes, classification, retrieval eligibility.
4. Fix permissions at the source, never in the application. The temptation is to build access filtering into the assistant. Resist it. If the repository's permissions are wrong, fix the repository; the assistant must inherit access, not reinvent it. An index that knows who may see what only because someone configured it in the AI layer will be wrong within a quarter.
5. Deduplicate and retire before you embed. Archiving superseded versions and collapsing near-duplicates is unglamorous and it is where retrieval quality is actually won. Embedding a corpus you have not curated is indexing your confusion at scale.
6. Measure retrieval, not just generation. The metric that matters is whether the right passages come back for a representative question set, judged by the domain owners who now exist. Answer fluency is the model's job; source correctness is the organisation's, and it is the one that should be on the dashboard.
7. Only then widen the aperture. Each new domain repeats the cycle, faster, because the owners, the metadata model and the profiling tooling now exist. Readiness compounds; cleaning in bulk does not.

Notice what is absent from that sequence: model selection, prompt design, vendor evaluation. Those matter, but they are the parts that everyone is already doing, and they are the parts that commodity pricing has made least decisive. The sequence above is where the programmes that worked diverged from the programmes that were quietly withdrawn.

7. The Bill Has Arrived

There is a version of the current moment in which an organisation issues an AI strategy, signs an enterprise licence, runs a demonstration, and reports to its board that it is moving at pace. The strategy is real, the licence is real and the demonstration is real. What is not real is the capability, because the corpus beneath it is an unexamined accumulation of everything the organisation has ever written, with no owner, no version control and no reliable account of who may see it.

The hard truth of the last eighteen months is that an AI strategy without a data readiness programme is a press release. The model is free, or nearly so. The data is not — and the bill for decades of deferred ownership, tolerated duplication and unexplained permissions is the one that has finally come due. Organisations that pay it deliberately, domain by domain, in sequence, will find the technology delivers roughly what the demonstration promised. Those that hope the next model release will pay it for them will be running the same eleven-minute demonstration next year, for a new chief executive.