

ENGINEERING

Retrieval that earns trust: designing AI you can defend

AgoActum · 2 May 2026 · 7 min read

Retrieval-augmented generation — connecting a model to your own knowledge — is the workhorse of useful enterprise AI. It is also easy to do badly. A system that retrieves the wrong passage and states the answer with total confidence is worse than no system at all.

Trustworthy retrieval is an engineering discipline. Here is what we insist on.

Retrieve the right thing

Most retrieval failures are retrieval failures, not model failures. If the relevant passage never reaches the model, no amount of prompting saves the answer. We invest in the boring, decisive parts: sensible chunking, hybrid keyword-plus-semantic search, and re-ranking that puts the best evidence first.

Honour permissions end to end

The fastest way to lose trust — and breach policy — is to let a user retrieve a document they should never see. Permissions must be enforced at retrieval time, not bolted on afterwards. Every query runs as the user, against only the data that user is entitled to.

Cite everything

Every claim the system makes should link back to the source it came from. This does two things: it lets users verify answers themselves, and it turns "hallucination" from a vague fear into a checkable property. If an answer has no supporting citation, that is a signal — surface it, don't hide it.

Evaluate relentlessly

You cannot manage what you do not measure. We build an evaluation harness alongside every system: a curated set of representative questions with known-good answers, scored automatically on every change. That is how you ship with confidence, catch regressions early, and detect drift once you're live.

Keep a human where it matters

For high-stakes actions, the right design is not full autonomy — it is a fast, well-designed human review step. The AI does the heavy lifting; a person keeps accountability. Users trust systems that respect that line.

Trust is not a feature you add at the end. It is the sum of a hundred design decisions made correctly along the way.