The demo retrieved perfectly. Then production arrived: the user asked the same question three ways, the embeddings shrugged, and your agent cited a paragraph that confidently said the opposite of the truth. Now you debug a black box one bad answer at a time, and the part of the stack you trust least is the part the whole product stands on.

The reflex is to swap the embedding model, bump top-k, and chunk smaller. That tunes the symptom and ships the disease: one similarity score was never going to tell you whether a passage is relevant, recent, or right.

The RAG Engineer treats retrieval as an engineering discipline, not a vibe. It hands you GRAIN (Grounding, Retrieval, Augmentation, Indexing, and Negative-feedback response), five stages that turn a brittle vector lookup into a pipeline you can measure, defend, and trust an agent to call without a human reading every result.

You will learn to:

- Build hybrid retrieval that fuses dense vectors with keyword search, so exact terms and rare entities stop falling through the gap.
- Add a reranker that reorders on relevance, not raw cosine distance, and earns the latency it costs.
- Chunk, embed, and index for the query you serve, not the one the tutorial assumed.
- Ground every answer in a citable span, and refuse cleanly when the context cannot support the claim.
- Measure retrieval as a system (recall, precision, and faithfulness on a real eval set), and answer drift before production degrades.

This is not a vector-database walkthrough or a course in transformer internals. It is operator-grade retrieval: protocols you run when the corpus is messy, the latency budget bites, and an agent acts on whatever your pipeline returns.

The result is concrete. You stop shipping confident wrong answers and start shipping a retrieval layer you can test, monitor, and hand to an agent without crossing your fingers.

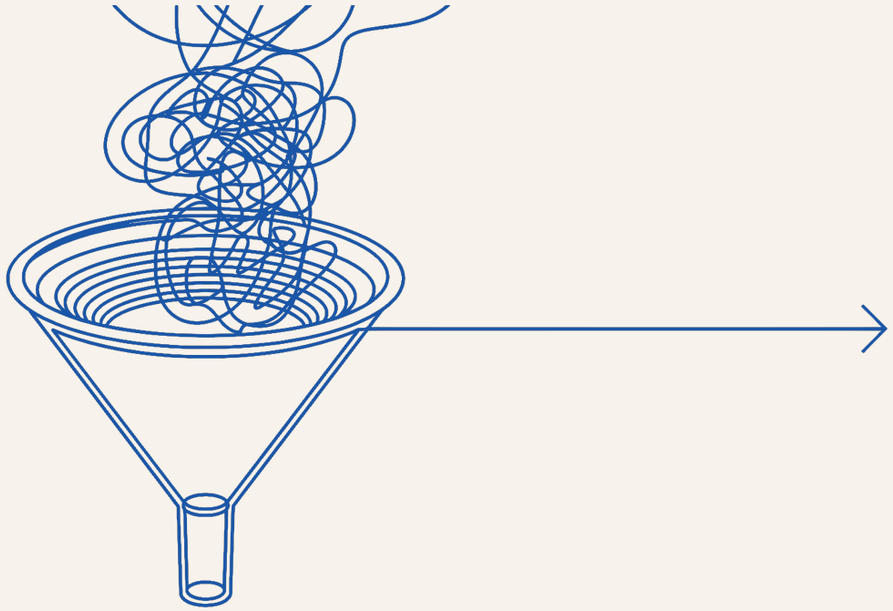
If you build the retrieval an agent depends on and refuse to call "good enough" a system you cannot measure, start here.

Len P. van der Hof, MSc
He builds operating systems for the strategic self. Treats entrepreneurship, AI and machine learning, marketing, philosophy, psychology, and health optimization as one engineering problem. MSc in Strategic Entrepreneurship, Rotterdam School of Management, Erasmus University.

Len P. van der Hof The RAG Engineer
SftSS

The *RAG* Engineer

Retrieval systems that agents can actually trust in production



Len P. van der Hof

Systems for the Strategic Self