In de demo werkte het ophalen perfect. Toen kwam productie: de gebruiker stelde dezelfde vraag op drie manieren, de embeddings haalden hun schouders op, en je agent citeerde een passage die met volle overtuiging het tegendeel van de waarheid beweerde. Nu debug je een black box, één fout antwoord tegelijk, en het deel van de stack dat je het minst vertrouwt is precies het deel waarop het hele product steunt.

De reflex is: wissel het embeddingmodel, verhoog top-k, hak in kleinere stukken. Dat sleutelt aan het symptoom en levert de ziekte op: één gelijkenisscore ging je nooit vertellen of een passage relevant, recent of juist is.

The RAG Engineer behandelt retrieval als een engineeringdiscipline, niet als gevoel. Het reikt je GRAIN aan (Grounding, Retrieval, Augmentation, Indexing en Negative-feedback response), vijf fasen die een broze vectorzoekopdracht veranderen in een pijplijn die je kunt meten, verdedigen en die je een agent durft te laten aanroepen zonder dat een mens elk resultaat naleest.

Je leert:

- Hybride retrieval bouwen die dichte vectoren samenvoegt met trefwoordzoeken, zodat exacte termen en zeldzame entiteiten niet langer tussen wal en schip vallen.
- Een reranker toevoegen die ordent op relevantie, niet op rauwe cosinusafstand, en de latentie verdient die hij kost.
- Chunken, embedden en indexeren voor de vraag die je bedient, niet die waar de tutorial van uitging.
- Elk antwoord verankeren in een citeerbare passage, en netjes weigeren wanneer de context de claim niet draagt.
- Retrieval meten als systeem (recall, precisie en getrouwheid op een echte evaluatieset), en op drift reageren voordat je productie verslechtert.

Dit is geen rondleiding door een vectordatabase of een cursus in de binnenkant van transformers. Het is retrieval van operatorklasse: protocollen die je draait wanneer het corpus rommelig is, het latentiebudget echt knijpt, en een agent handelt op wat je pijplijn ook teruggeeft.

Het resultaat is concreet. Je stopt met het uitleveren van zelfverzekerde foute antwoorden en begint met het uitleveren van een retrievallaag die je kunt testen, monitoren en aan een agent kunt overdragen zonder je vingers te kruisen.

Bouw je de retrieval waar een agent op leunt, en weiger je "goed genoeg" te noemen wat je niet kunt meten, dan begin je hier.

drs. Len P. van der Hof

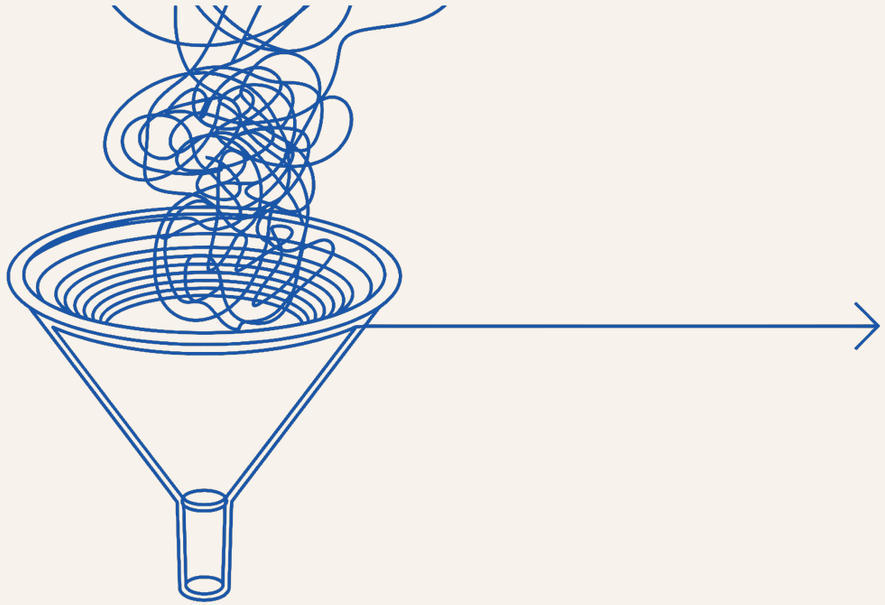
Hij bouwt besturingssystemen voor het strategische zelf. Behandelt ondernemerschap, AI en machine learning, marketing, filosofie, psychologie en gezondheidsoptimalisatie als één engineeringprobleem. MSc in Strategic Entrepreneurship, Rotterdam School of Management, Erasmus Universiteit.

Len P. van der Hof **De RAG- Engineer**

SftSS

De RAG-*Engineer*

Retrieval-systemen waar agents in productie echt op kunnen vertrouwen



Len P. van der Hof

Systems for the Strategic Self