

YPOG AI Workshop: Legal, Fund & Finance

2026-06-17

About me

- **Alexandros Sivris**
- Freelance product engineer
- Building AI-driven enterprise software solutions and data platforms



alexandros-sivris

alex@sivris.com · sivris.com

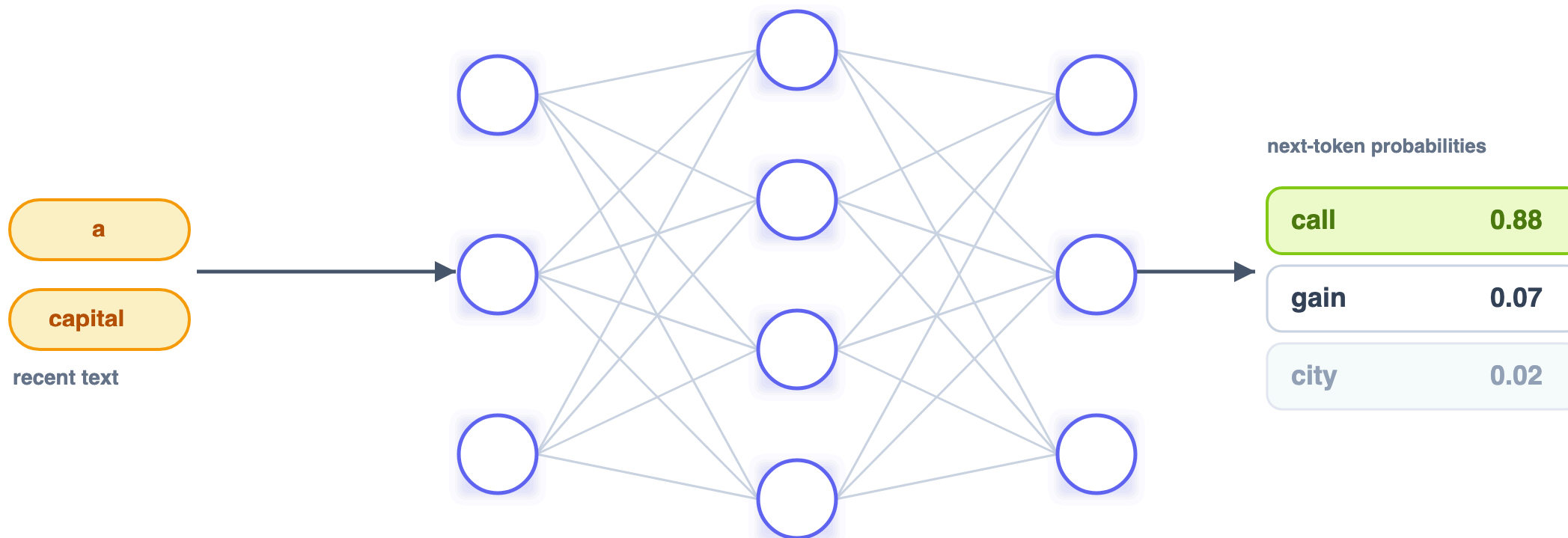
Goals

By the end of this workshop, we will understand:

1. **What AI models are**
2. **Why local (sovereign) AI matters**
3. **How to set up your own local AI**

1. What are AI models?

each line = a weight (a learned number)



recent text

next-token probabilities

call 0.88

gain 0.07

city 0.02

layers of neurons

each neuron: inputs $\times$ weights $\rightarrow$ one output



Qwen3-1.7B-Q4_K_M.gguf

about 1.2 GB of weights: this file is the model

training sets these numbers, inference just reads them

Types of AI models

Today we focus on the two highlighted ones

Focus today



Chat (LLM)

Writes and answers in words

Focus today



Embeddings

Turns text into searchable fingerprints



Vision

Reads images and scanned PDFs



Speech to text

Turns speech into text



Text to speech

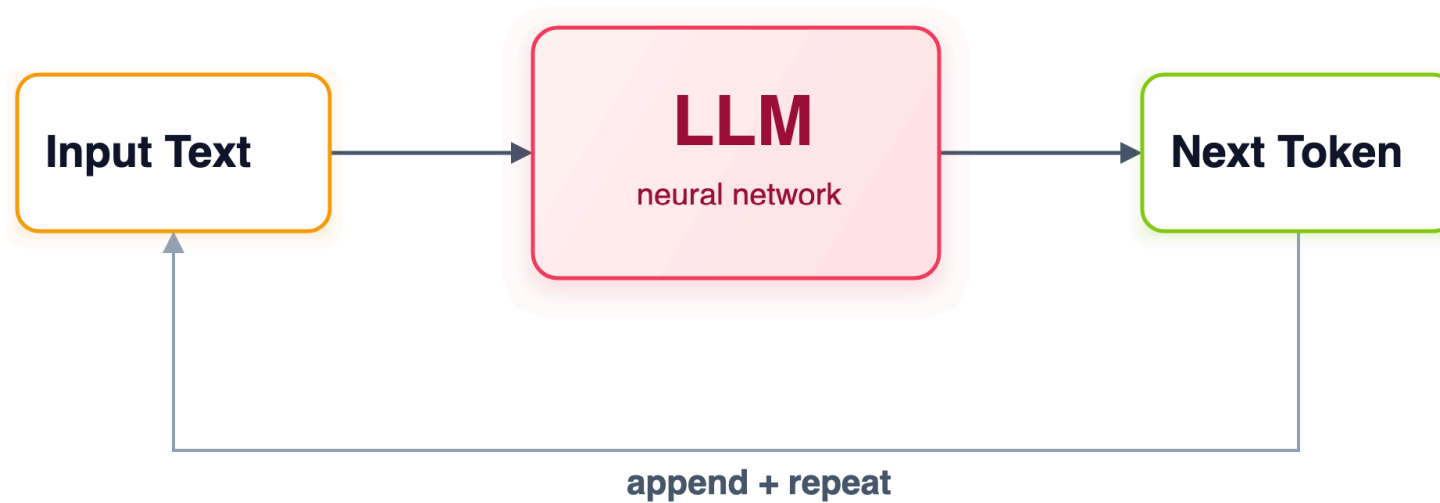
Reads text aloud



Image generation

Creates images from text

⚡ Using a model: inference



One pass = one new token. Repeat = full answer (simplified: token = word):

Pass 1:	What is a capital call?	→	A	
Pass 2:	What is a capital call? A	→	capital	
Pass 3:	What is a capital call? A capital	→	call	
Pass 4:	What is a capital call? A capital call	→	is	...

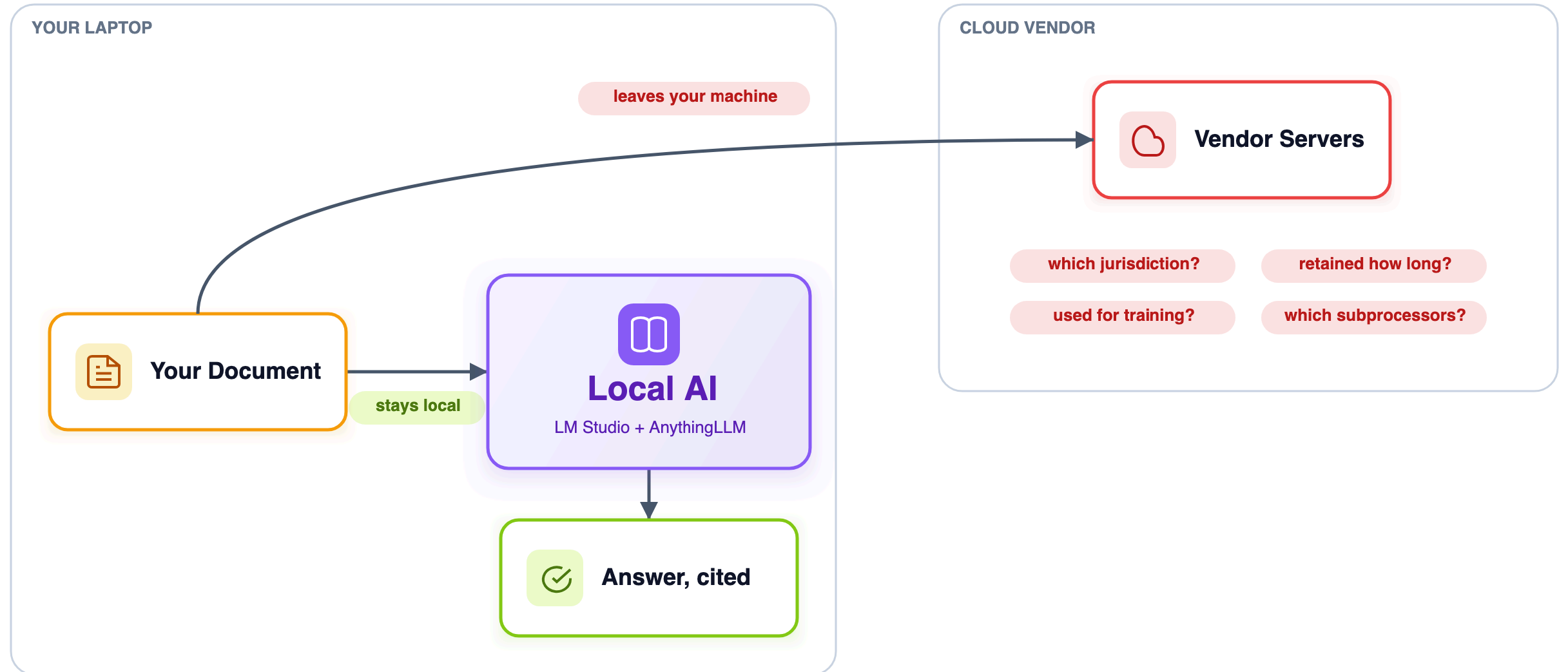
Building a model: training

- Reads billions of examples, guesses the next token, corrects itself, repeats
- Example: What is a capital call? →
 -  A capital call is a demand for committed capital... → reinforce these weights
 -  A capital call is a banana... → adjust weights away
- Millions of corrections → model gets better at continuing any text

2. Why local AI?

Where does your prompt go?

same question, two very different data paths







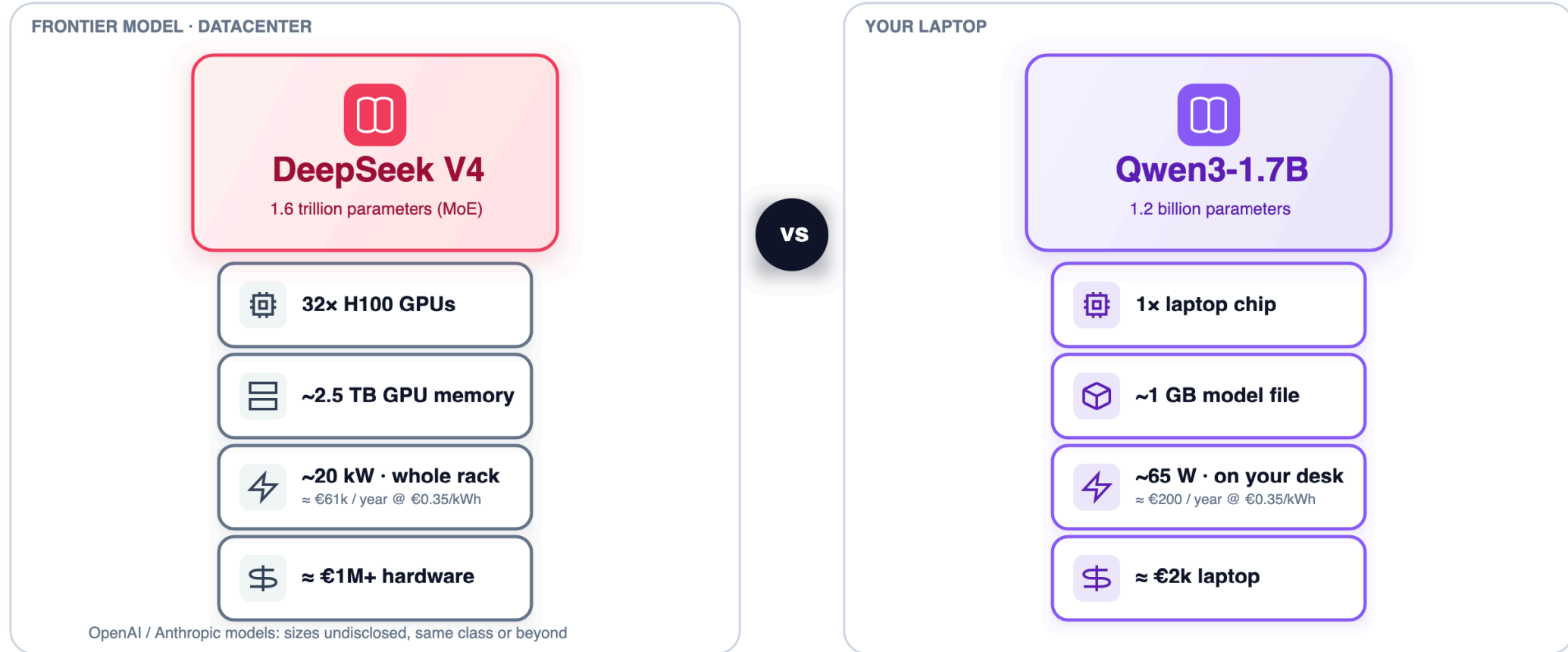
When it went wrong

- **Anthropic Fable 5, 2026:** pulled **globally overnight**, 3 days after launch, by US export-control order → off for everyone, EU included. Workflows tuned to it broke; the contract did not matter ([Euronews](#)). **You don't own the model you depend on.**
- **DeepSeek, 2025:** open database, 1M+ plaintext chat logs + secret keys exposed ([Wiz](#))
- **Shared chats on Google, 2025:** ~100k ChatGPT conversations indexed + scraped ([TechCrunch](#))

Local AI: the simplest defensible answer to *"where is my data processed?"*

Frontier model vs. your laptop

why prompt skill matters more locally



≈ 1,300x fewer parameters → your prompt has to do the heavy lifting

3. Best practices & AI patterns

N-shot

- Zero-shot: ask cold
- Few-shot: provide 2–5 worked examples
- Examples teach format and tone faster than instructions
- Use few-shot when output structure matters (tables, citations)

Chain-of-Thought

- "Think step by step"
- AI shows its reasoning before the answer
- Better on multi-step questions (numbers, logic, comparisons)
- Trade-off: slower, more tokens, but fewer silly mistakes
- Variant: generate several reasoning paths, take the majority vote ("tree-of-thought")

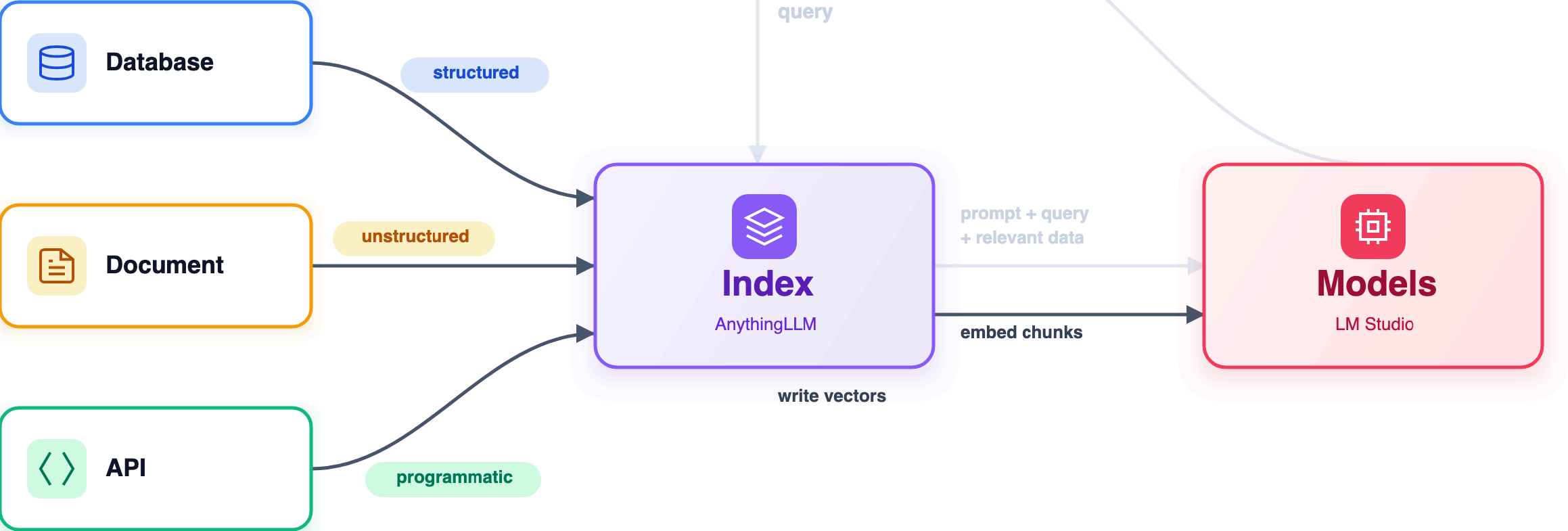
RAG: Retrieval Augmented Generation

- Give AI only the **relevant pages** of your documents, then let it answer
- Like handing a paralegal the exact clauses to read before the question
- Answers stay grounded in source text, with citations

This is exactly what you'll build today.

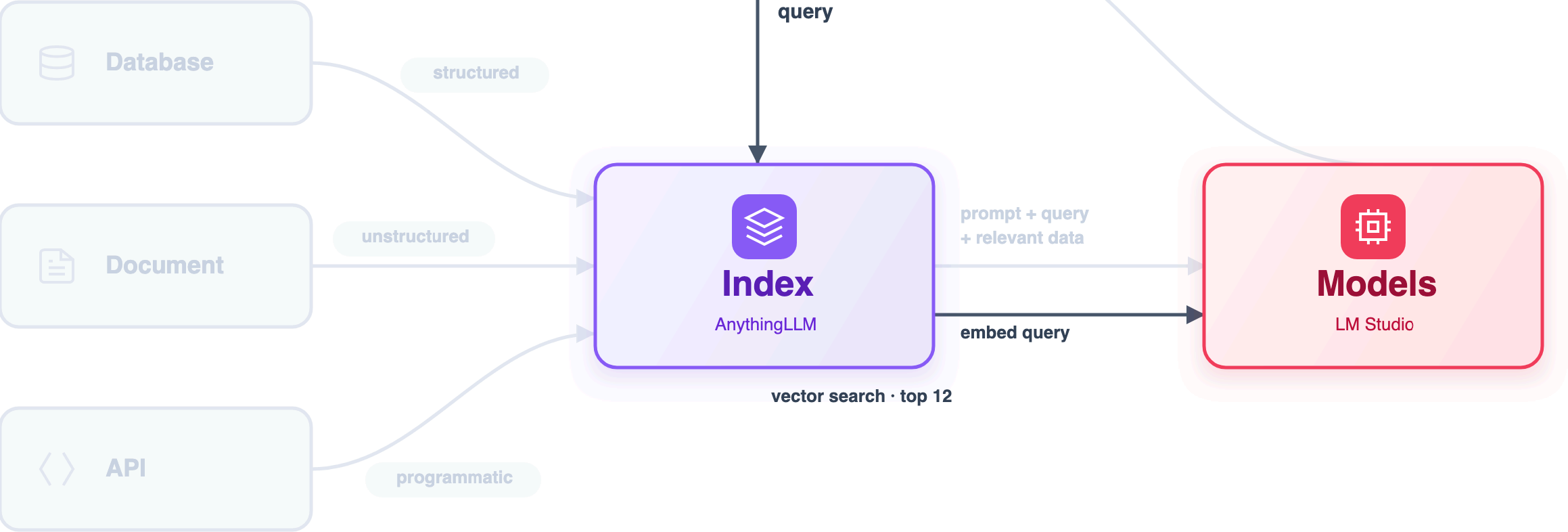
Phase 1 — Ingestion

Your data



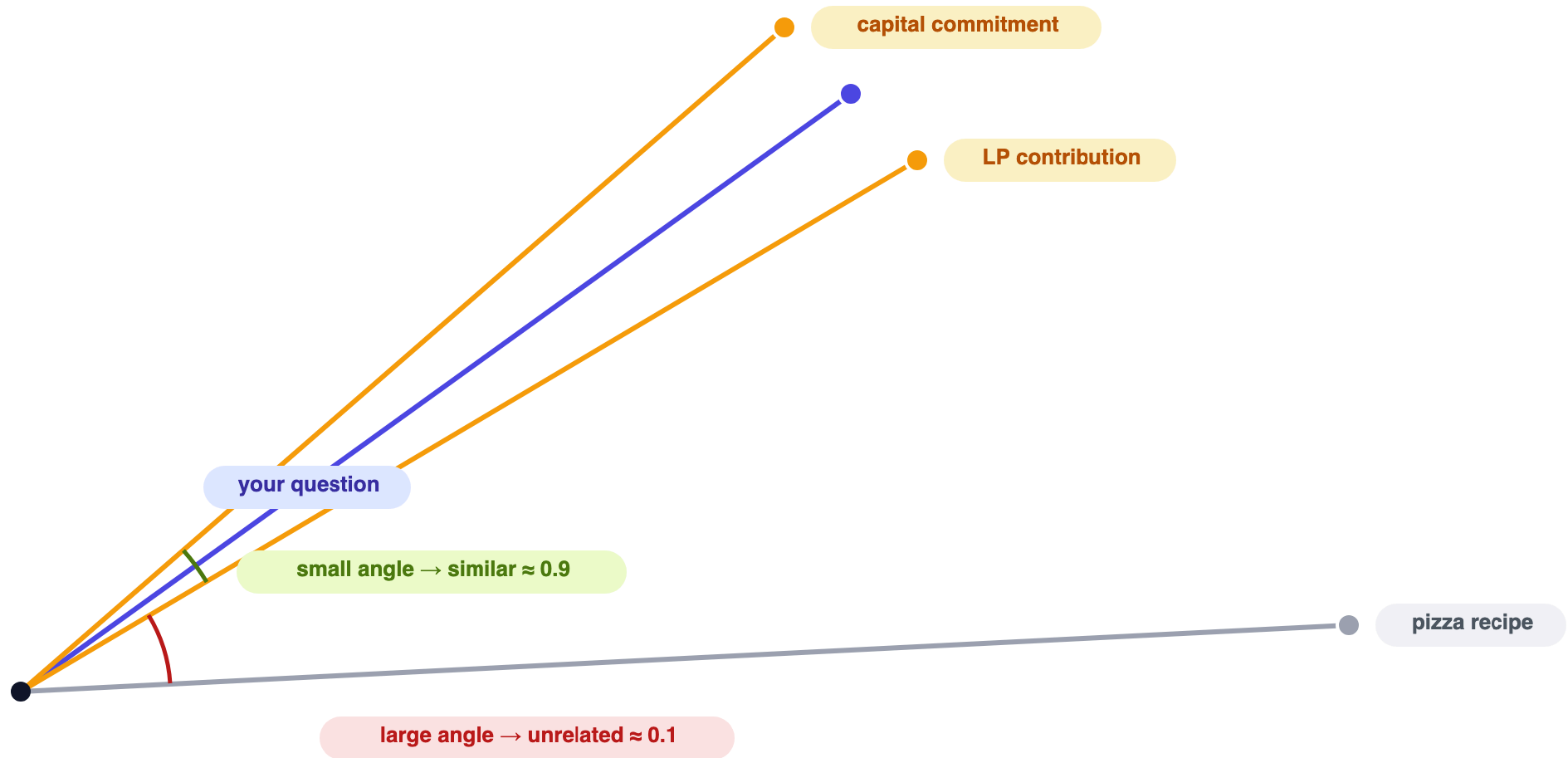
Phase 2 — Retrieval

Your data

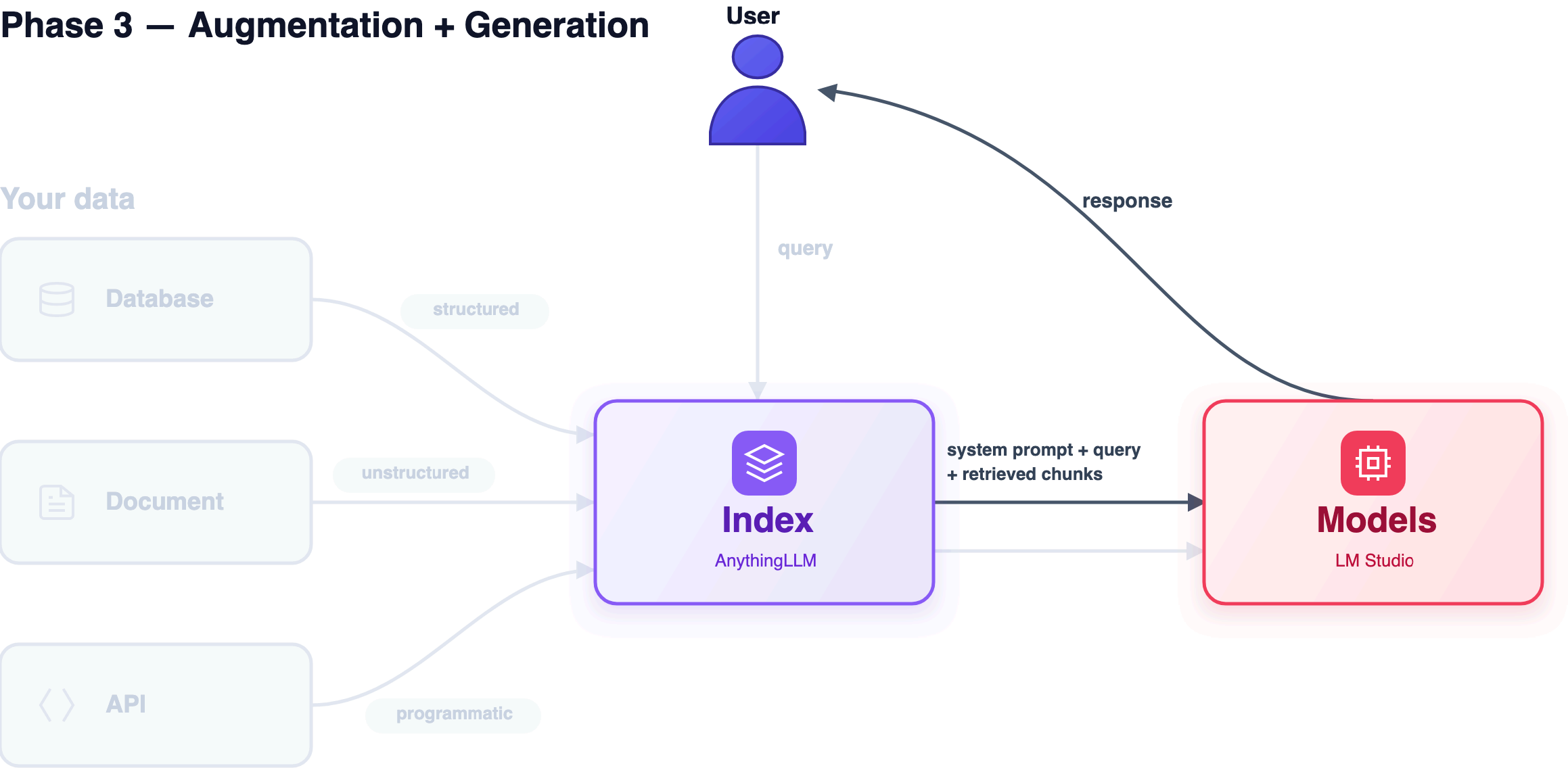


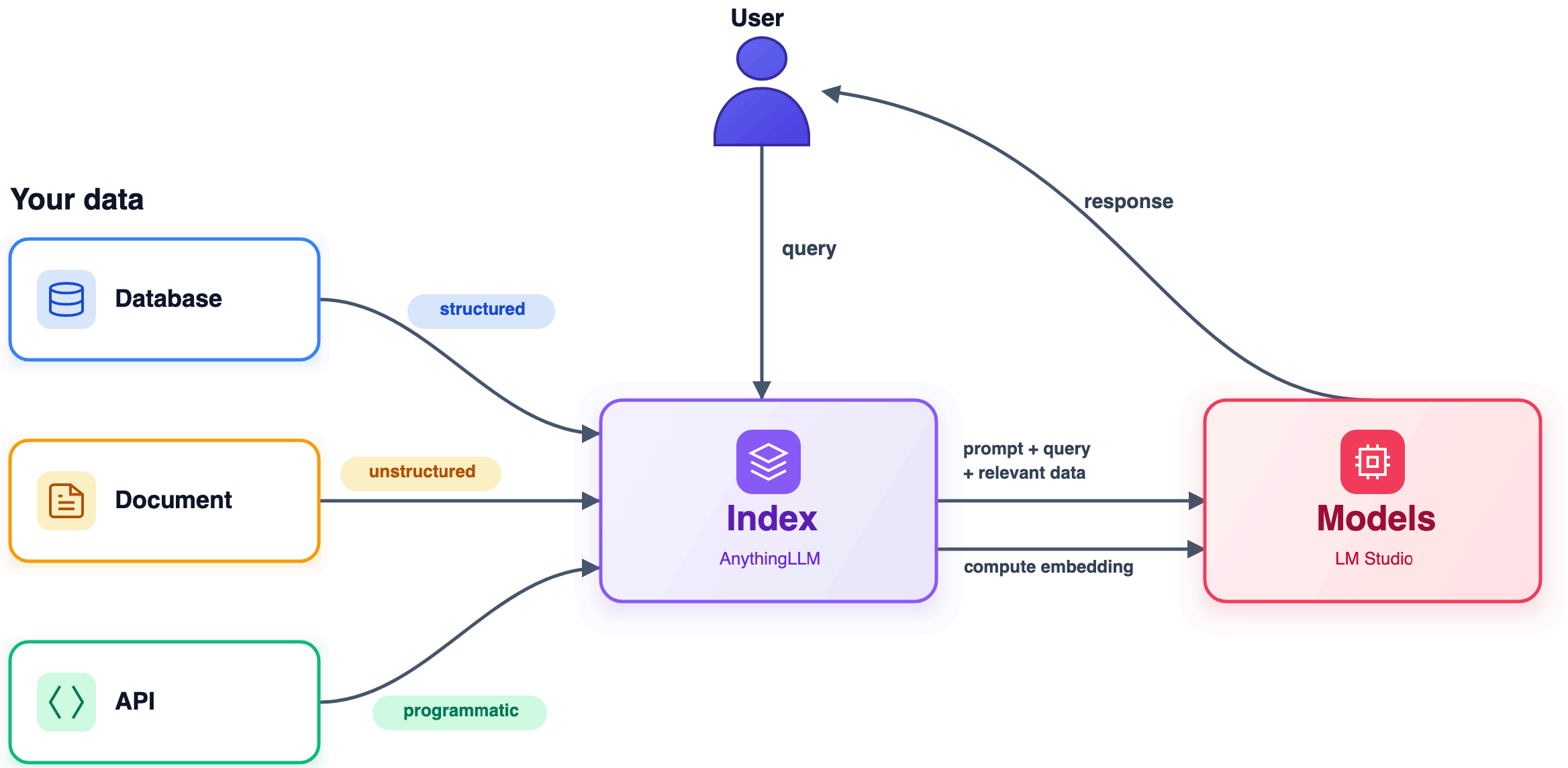
Embeddings: a map of meaning

similar meaning → small angle between vectors → high score

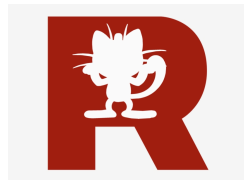


Phase 3 — Augmentation + Generation





4. Hands-on: Set up your local AI stack



Meet the fund: Team Rocket Opportunities Fund I LP

- **Fund:** Cayman Islands Exempted LP · final close **30 Sep 2023**
- **GP:** Team Rocket GP Ltd. · **IM:** Team Rocket Capital Management LP
- **Commitments:** USD 250M total
- **Economics:** 2% management fee · 20% carried interest · 8% pref · European-style waterfall
- **3 Limited Partners:**
 - Giovanni Family Office · 100M (40%) · MFN side letter
 - Archer Pension Fund · 90M (36%) · ESG side letter
 - Jessie Family Office · 60M (24%) · monthly-reporting side letter

The corpus: 16 documents

- **1× LPA** (Amended & Restated)
- **3× Side letters**, one per LP (MFN, ESG, monthly reporting)
- **3× KYC packs**, one per LP (Giovanni AML hit, Jessie adverse media, Archer clean)
- **2× Quarterly reports** (2024 Q3 + 2025 Q1)
- **2× Manuals** (Compliance, Operations)
- **1× Capital call notice** (No. 3, image-only/OCR) · **1× Investor register**
- **1× Audit findings** (2024) · **1× IC minutes** (Magikarp 2024 Q3) · **1× Internal compliance memo**

 Fully fictional. Synthetic personal data for workshop only.

Your stack today

- **AnythingLLM**: workspace on top; embeds your PDFs, retrieves, chats (your RAG pipeline)
- **LM Studio**: runs the models on your laptop, serves them at `http://localhost:1234`
- Models, all already in your kit (three chat sizes — pick by your RAM):
 - `Qwen3-1.7B` / `Qwen3-4B` / `Qwen3-14B` (Q4_K_M) → write the answers
 - `bge-m3` → fingerprints/*embeds* the documents for search
- Everything runs offline after install.

Steps

Follow **INSTRUCTIONS.md** in the kit (pictures for every step).

1. Copy the kit from the USB stick to your Desktop, pass the stick to your neighbor
2. Install both apps: **AnythingLLM** and **LM Studio**
3. Open **AnythingLLM**, let it pick the models (the default option)
4. New workspace → upload all the fund PDFs → **Save and Embed** (takes a few minutes)
5. Set chat mode to **Query** + paste the system prompt (in INSTRUCTIONS)
6. Ask your first question: *"Who are the LPs of the fund?"*

 Fast finishers: run your own model via **LM Studio** (Step 6), or try another model / your own documents.

Thank you



alexandros-sivris

alex@sivris.com · sivris.com