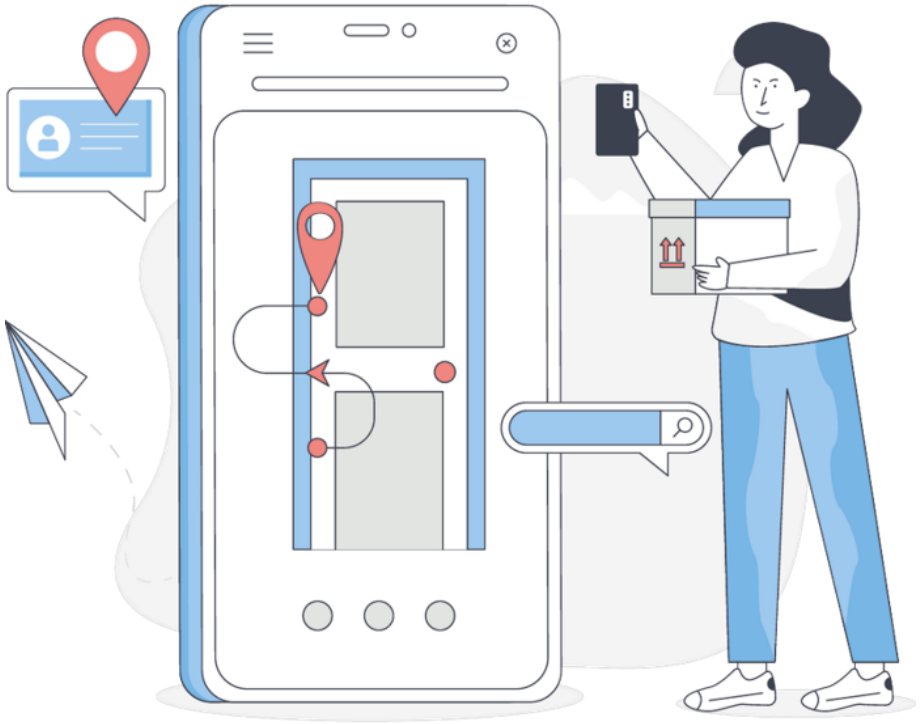
HOGESCHOOL UTRECHT



By Naveed Hussain on IconScout

Exploring local LLMs and RAG

Pavlo Burda

Outline

- When can’t we use cloud LLMs
- Local LLMs
- Use cases
- Retrieval Augmented Generation
- Optional advanced material (AnythingLLM)

Public LLM providers

When you cannot use public LLMs, like ChatGPT?

By vectorjuice on Magnific

Public LLM providers

When you cannot use public LLMs, like ChatGPT?

- Private/financial data
- your employer AI policy forbids
- client documents
- Confidential company info
- Breach of GDPR

By vectorjuice on Magnific

Use cases

- Local knowledge base
- Instrumenting workflows
- Local source code

Query local knowledge base

The idea is to ask questions across your own files

- Find by meaning, retrieve source (“where did I mention this?“)
- Summarize, inspect differences (“compare v1 vs v2”)
- Pull and format structured data (tables with dates, classify)

A local LLM answers questions about them without sending anything to a cloud provider.

Documents

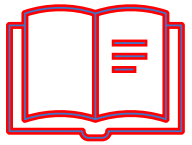
reports, manuals, PDFs

Notes

meeting & research notes

Policies

HR, security, compliance



Contracts

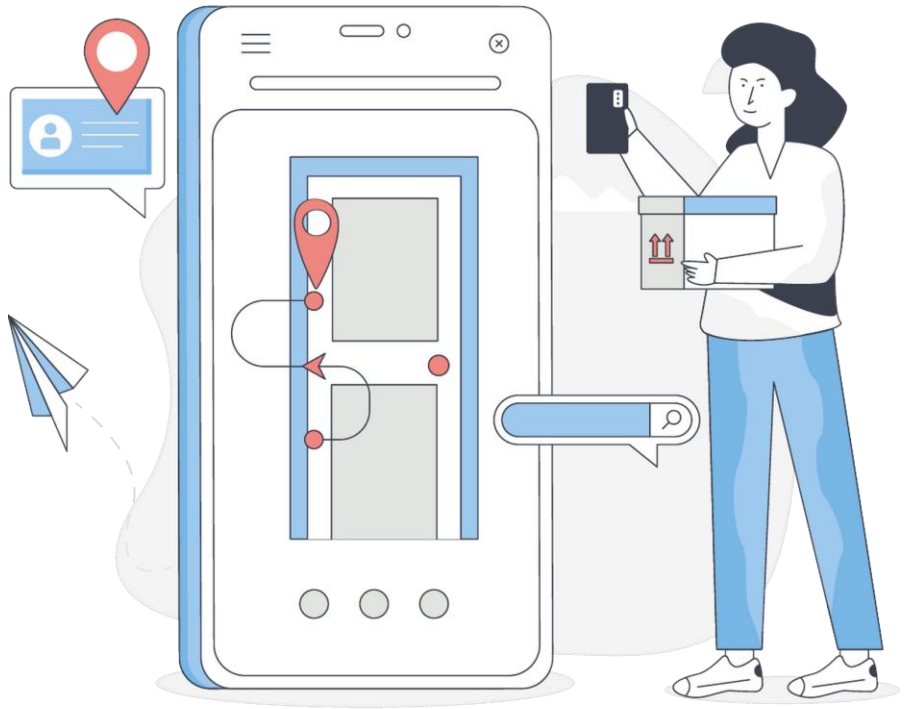
Customers, financial, health

Instrument your workflow

Use your calendar, email, and chats to:

- Suggest meeting slots in the calendar
- Draft agendas based on available context
- Summary of your schedule

Trigger routines and automations for reminders based on emails & chats



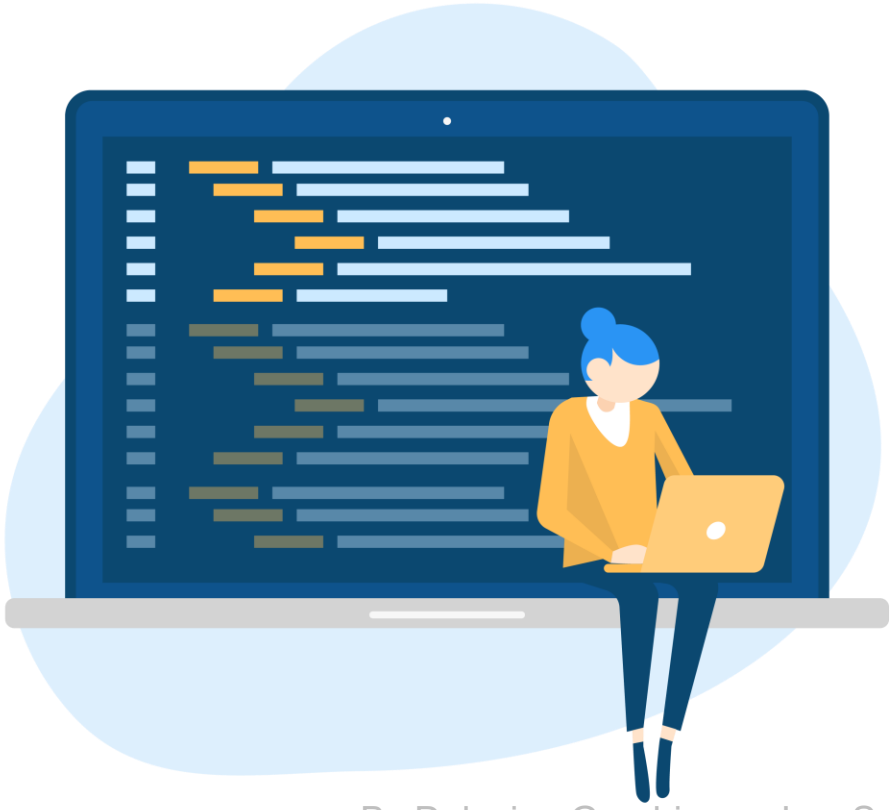
By Naveed Hussain on IconScout

Local source code

Sometimes you must keep code local, but you still want to ask suggestions based on your codebase.

- Generate boiler-plate code
- Explain functions/modules
- Query package documentation
- Debugging

… you name it

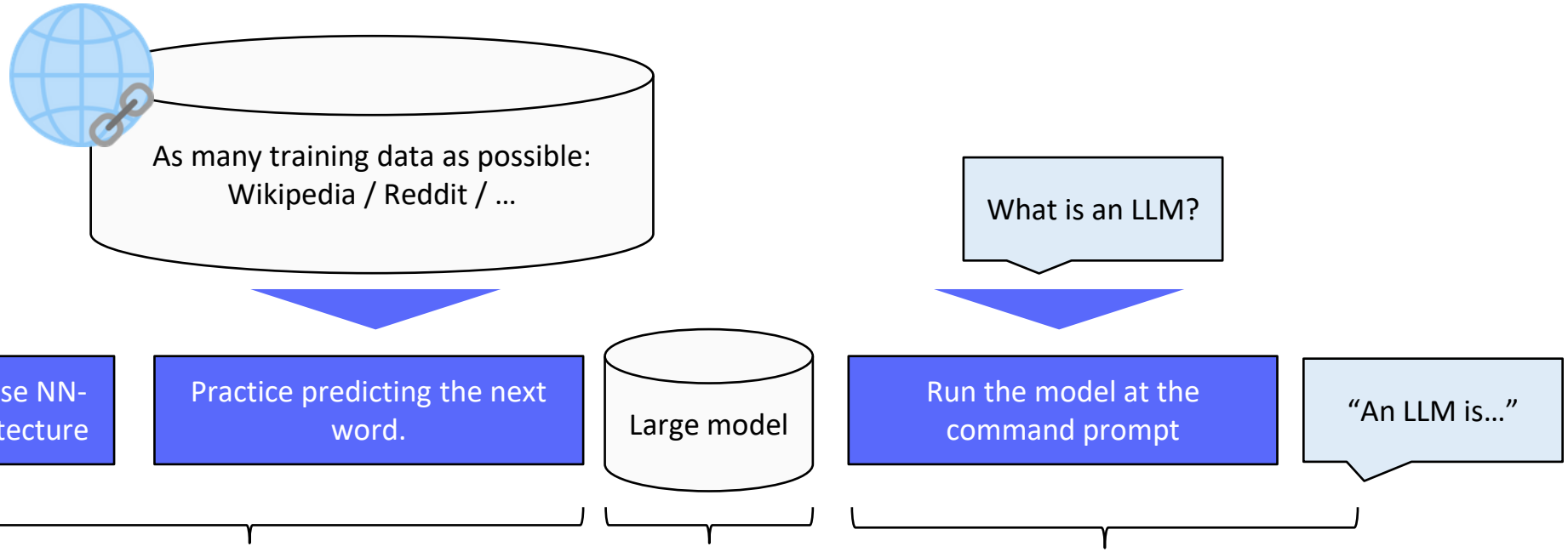


By Delesign Graphics on IconScout

Context engineering

- Local LLMs
- Retrieval-augmented Generation (RAG)
- Optional advanced material

What is a Large Language Model?



One-off. Requires a **supercomputer**

By Anastasiia Torianyk on IconScout

Storage is a challenge. It might just work on a standard computer, or it might not.

This is done on a per-user basis, often in the cloud. A charge of a few cents per token, word or sentence is typically applied

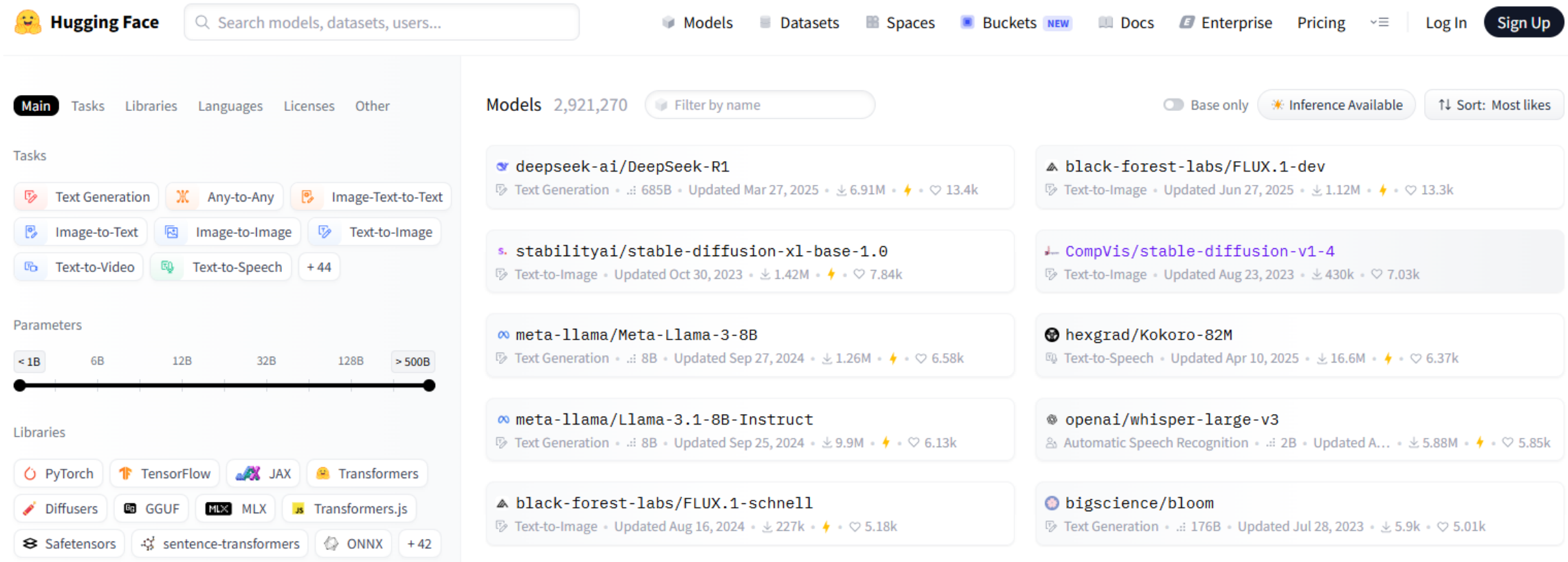
Local vs. Large LLM

Requirement	Small (local) LLM	Large LLM
License	Often Open Weight	Often Closed Weight (not public)
Parameters	4 to 14 billion; 40b max	Hundreds of billions (405B+)
Download size	4-5-8 GB (4-6-bit quantized)	800 GB+ (full precision)
Memory needed	8-16-32 GB system RAM	600+ GB VRAM (8x 80 GB GPUs)
Hardware	Laptop / desktop; Nvidia GPU	Data-center GPU cluster
Typical cost	~0 to ~€1-2k	€100k+ servers / per-token fees
Application	Specific tasks	General purpose

Example: Llama 3 8B (4-bit) on a laptop vs. Llama 3.1 405B in a data center

Local LLMs – Hugging Face

A repository of open weight models: small and big, textual and multimodal

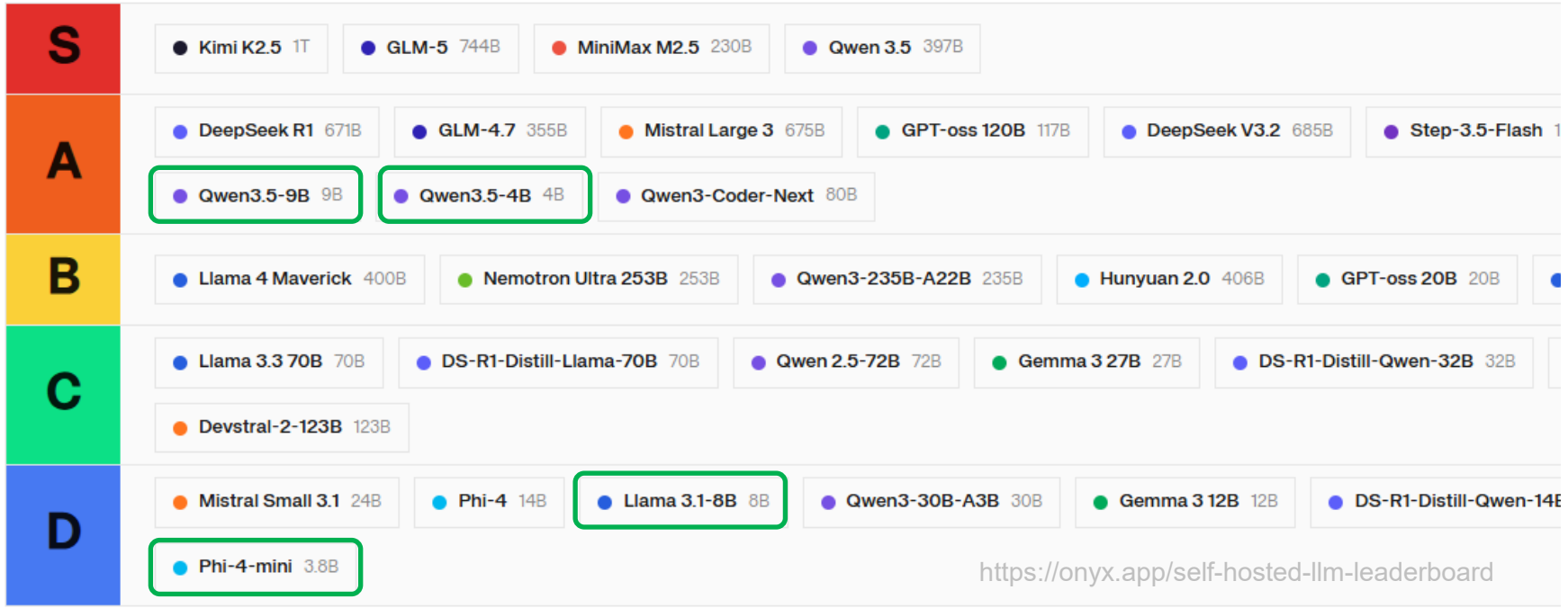


https://huggingface.co/collections/open-llm-leaderboard/open-llm-leaderboard-best-models

Local LLMs and Tools

Top local LLMS:

By Onyx (https://onyx.app)



Tools:

Ollama

LM Studio

Jan

n8n

Anything LLM

LangChain

GPT4All

Context Engineering

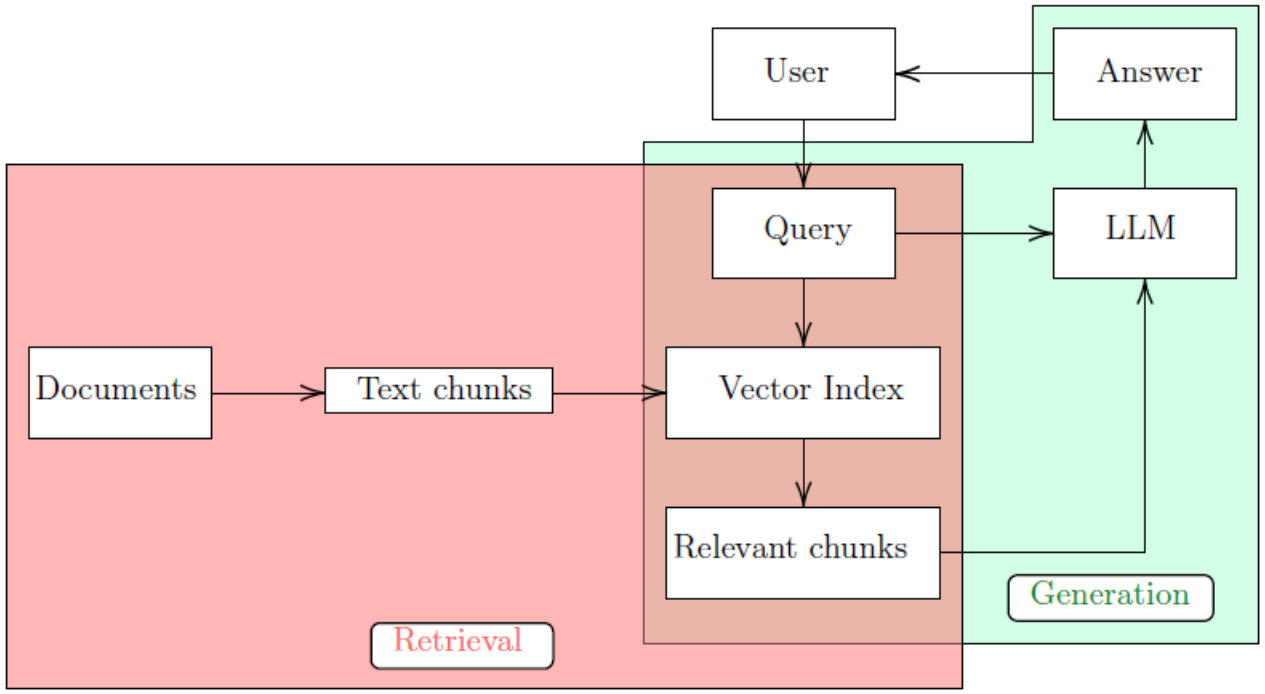
LLMs are trained on a general corpus of human knowledge and interactions (mostly internet, Wikipedia, Reddit, pirated books …).

Still, they do not know anything about your data.

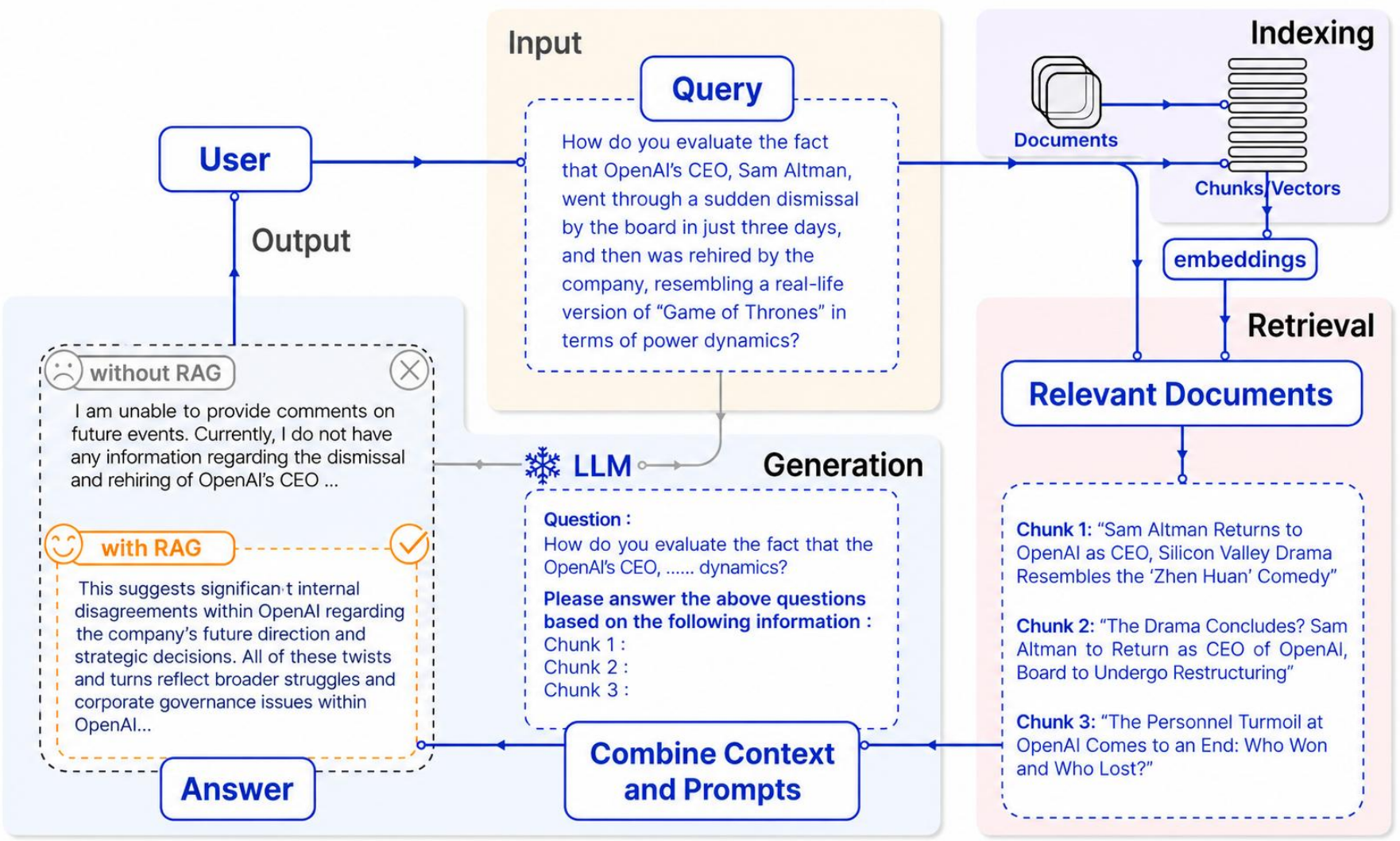
The content and quality of Local LLM outputs depends on what and how you provide information to the model:

- (Re) training – build weights from scratch (need $Xm & a modular reactor)
- Fine-tuning – adjust weights with targeted data (hard, still expensive)
- **Context engineering/RAG – supply your data, weights untouched**

RAG (Retrieval-Augmented Generation)



RAG: combining prompt with context

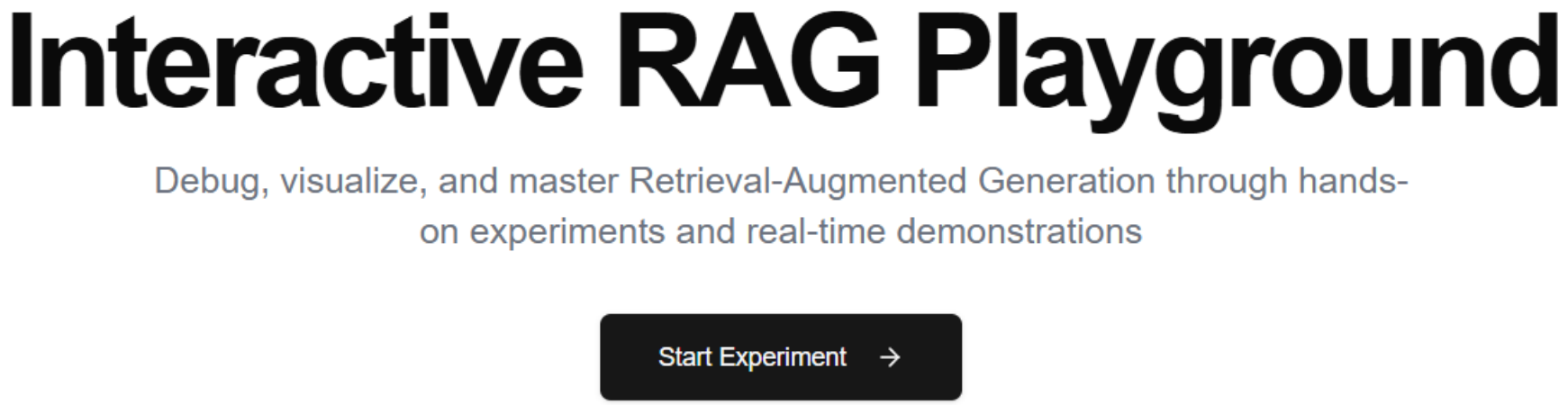


Src: https://cloud.tencent.com/developer/article/2523788

Practice Session

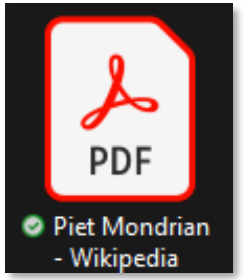
- Form a team of 2 people
- 1 laptop is going to be your main device, decide whose

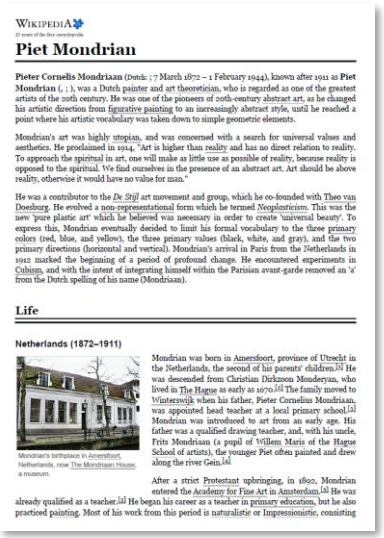
- Go to: https://ragplay.vercel.app



Practice RAG

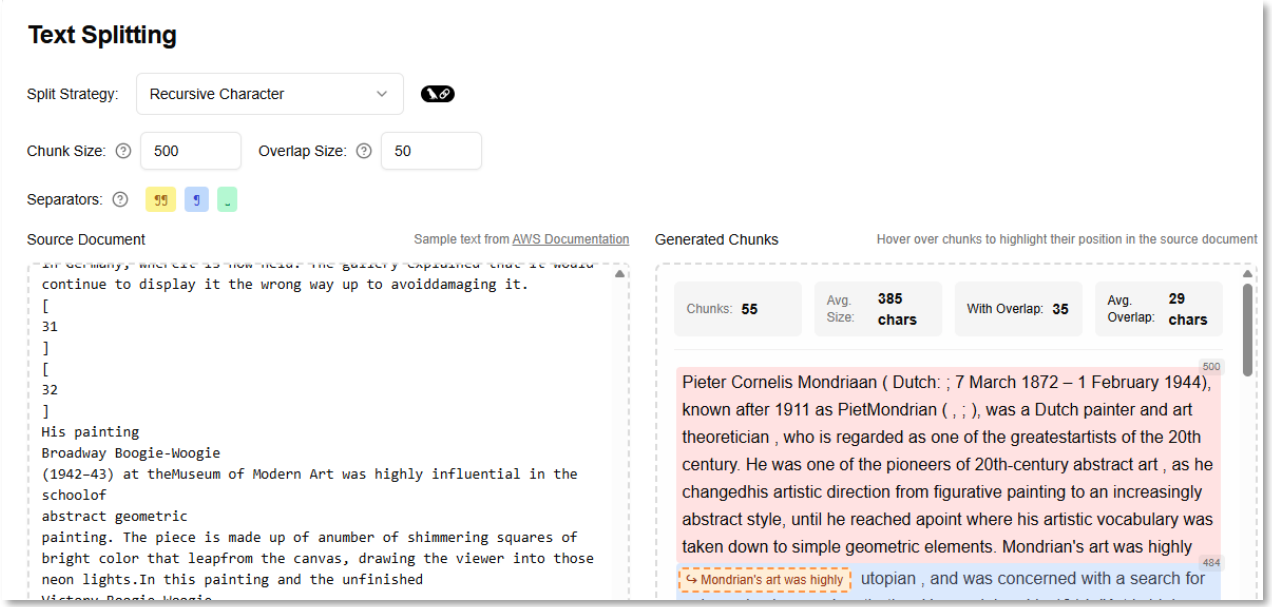
- Get the Piet Mondrian - Wikipedia.pdf from Friday folder
- Copy all the content and past into Text Splitting window





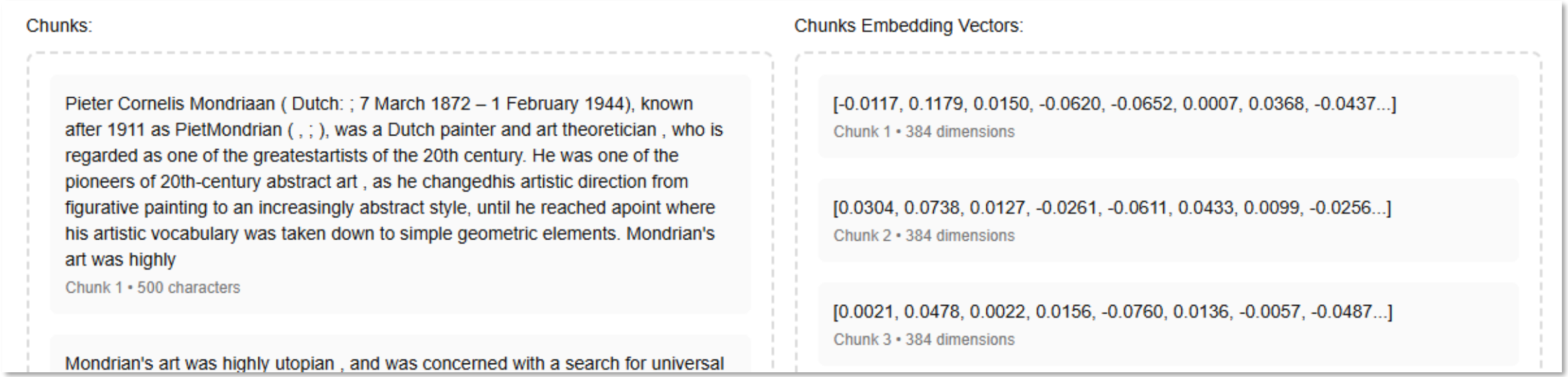
How many chunks did you generate?

What is the chunk size?



Src: https://github.com/Kain-90/RAG-Play

Practice RAG



Src: https://github.com/Kain-90/RAG-Play

Practice RAG

Chunks:

Pieter Cornelis Mondriaan (Dutch: ; 7 March 1872 – 1 February 1944), known after 1911 as PietMondrian (, ;), was a Dutch painter and art theoretician , who is regarded as one of the greatestartists of the 20th century. He was one of the pioneers of 20th-century abstract art , as he changedhis artistic direction from figurative painting to an increasingly abstract style, until he reached apoint where his artistic vocabulary was taken down to simple geometric elements. Mondrian's art was highly

Chunk 1 • 500 characters

Mondrian's art was highly utopian , and was concerned with a search for univ

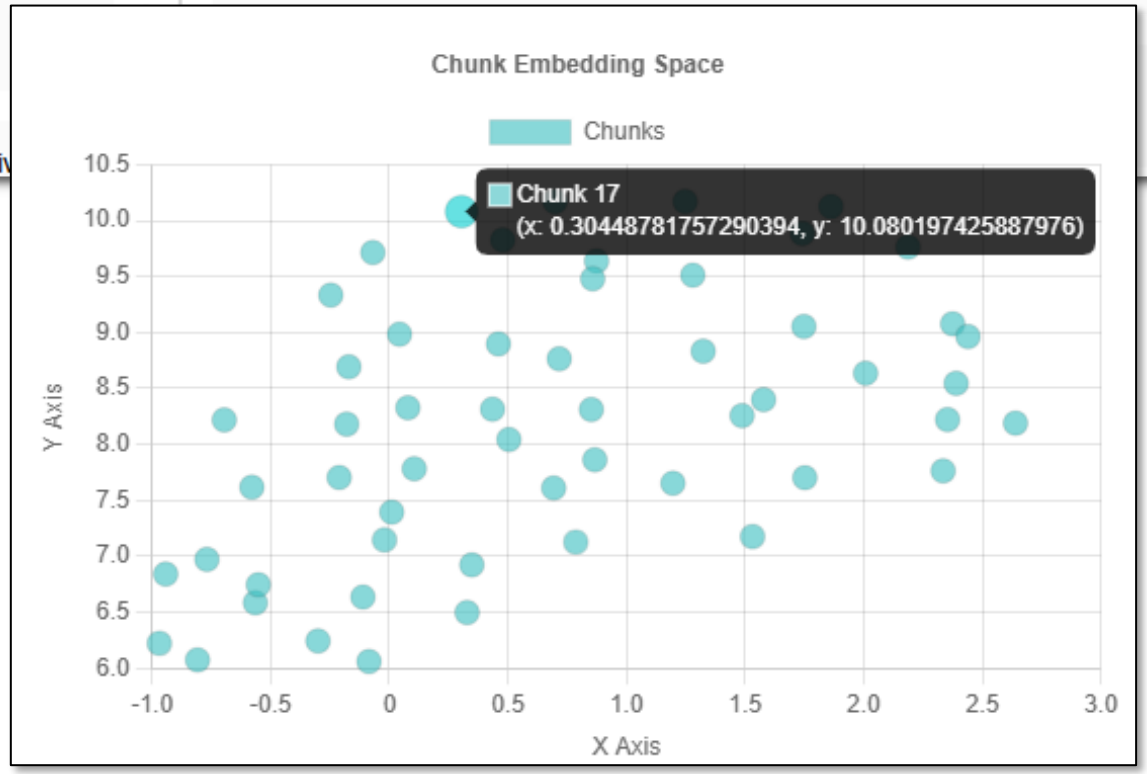
Chunks Embedding Vectors:

[-0.0117, 0.1179, 0.0150, -0.0620, -0.0652, 0.0007, 0.0368, -0.0437...]

Chunk 1 • 384 dimensions

[0.0304, 0.0738, 0.0127, -0.0261, -0.0611, 0.0433, 0.0099, -0.0256...]

Chunk 2 • 384 dimensions



Src: https://github.com/Kain-90/RAG-Play

Practice RAG

Semantic Search

Embed a user query, compare it against the knowledge-base vectors, and in

Ask a question to find similar content:

Painting style?

[0.0416, 0.0816, 0.0079, -0.0086, -0.0283, 0.0463, -0.0304, -0.0821...]

Question embedding • 384 dimensions

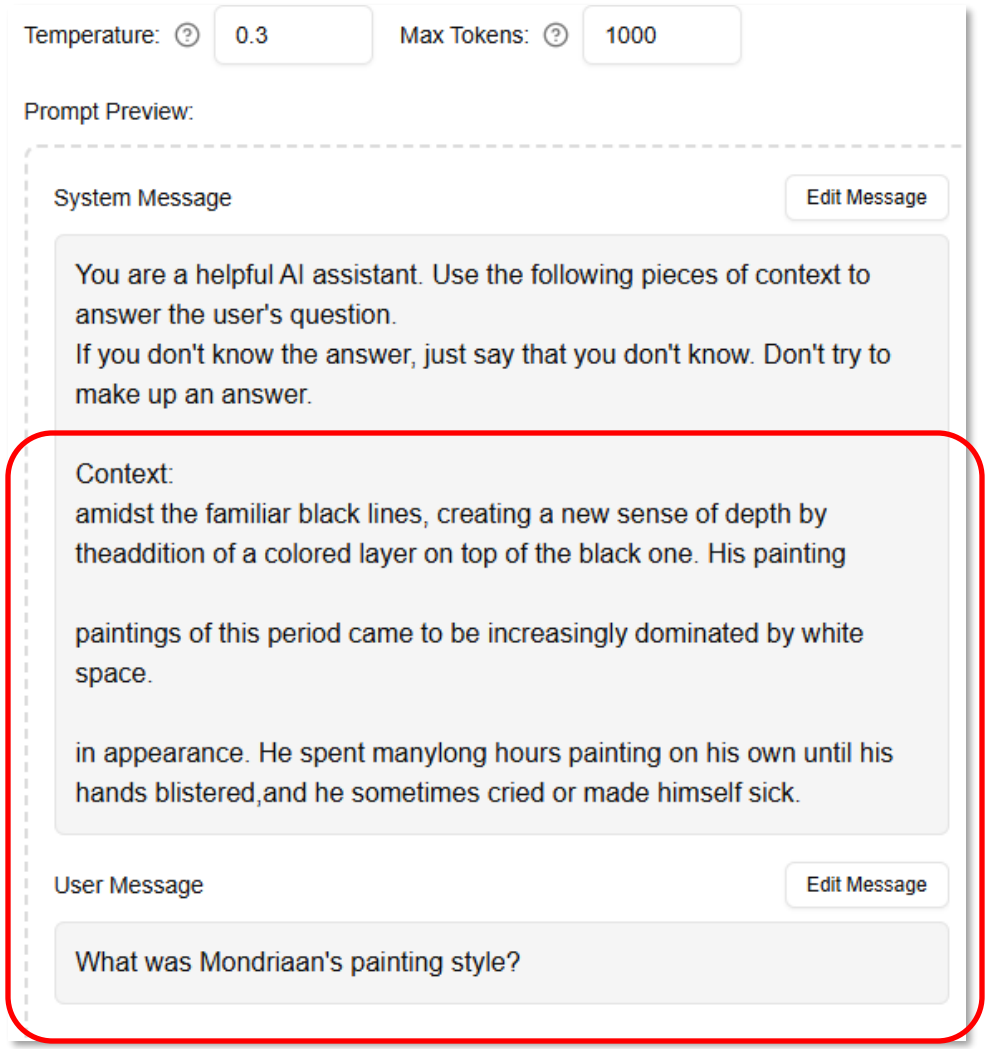
Similar Chunks:

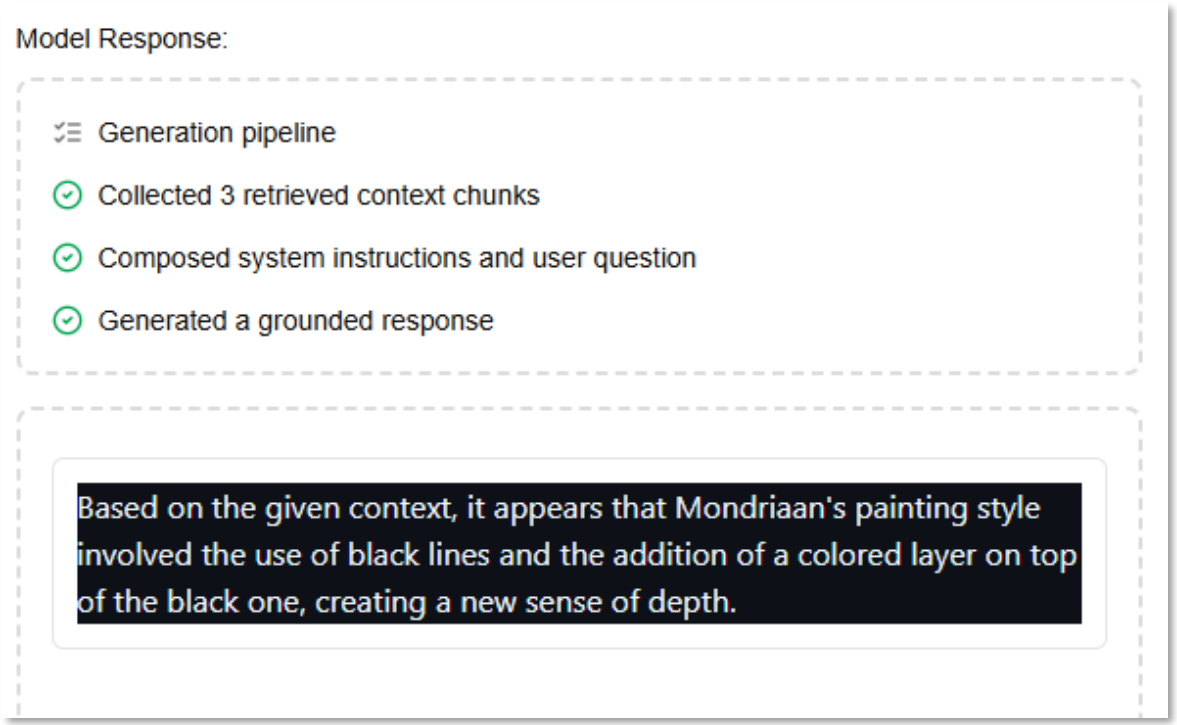
amidst the familiar black lines, creating a new sense of depth by theaddition of a colored

Chunk 46 • Similarity: 0.8607

Src: https://github.com/Kain-90/RAG-Play

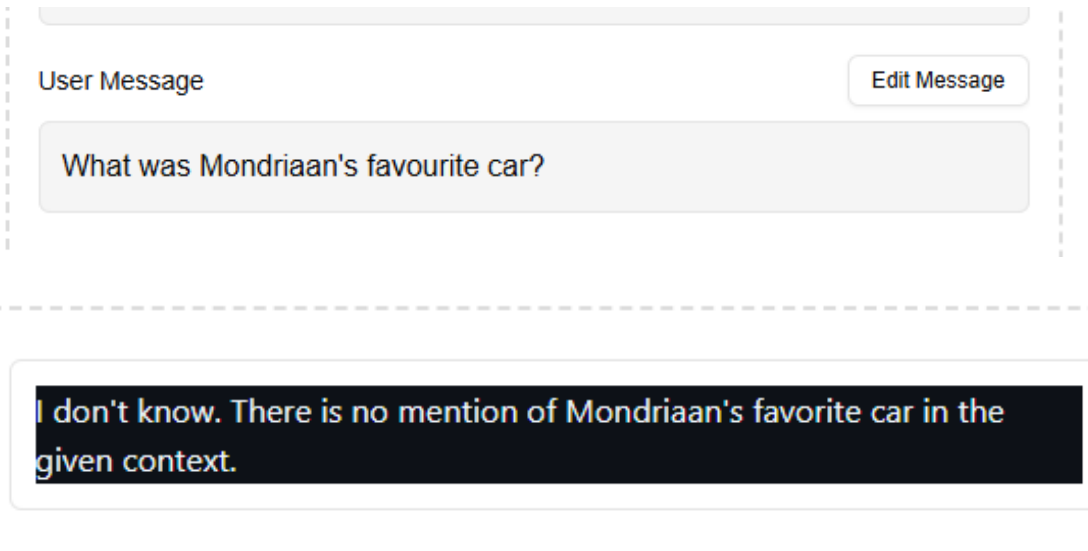
Practice RAG

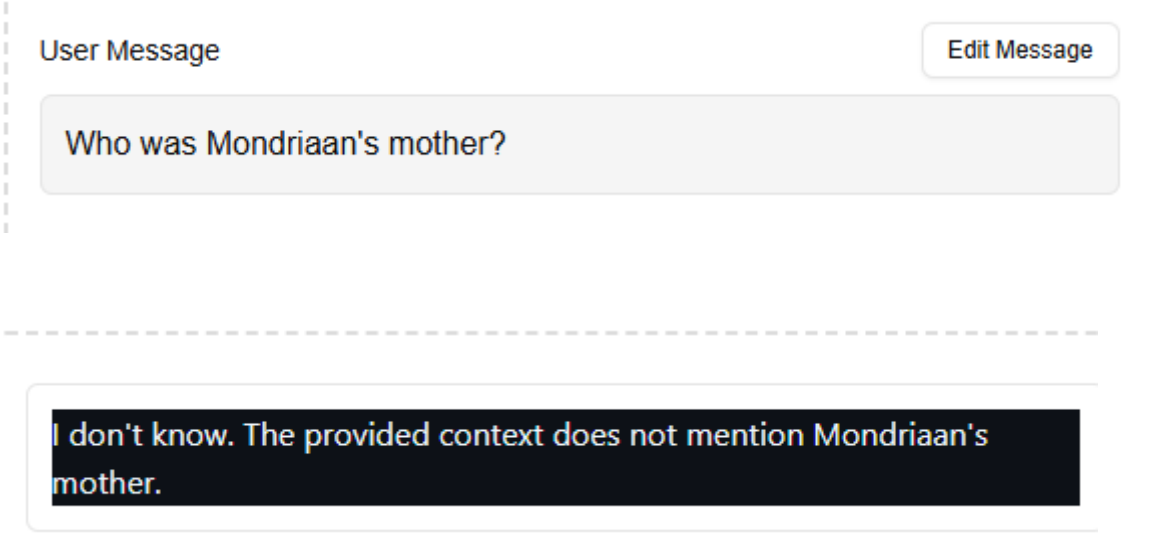




Src: https://github.com/Kain-90/RAG-Play

Practice RAG





Src: https://github.com/Kain-90/RAG-Play

Fooling RAG

Pavlo and Sieuwert's history of achievements

Sieuwert van Otterloo (from ictinstitute.nl)

- Holds a PhD in computer science (studied mathematics and cognitive AI) and completed post-academic training at Leiden University.
- Led the consulting group at SIG, co-founded ICT Institute in 2013, also acting as an advisor and angel investor in several startups.
- Independent IT expert at LRGD, SGOA, and the NVBI, and serves as Data Protection Officer.

Pavlo Burda (from ictinstitute.nl and paolokoelio.github.io)

- Holds a PhD in Information Security at TU/e Eindhoven.
- Published extensively on cyber security and phishing across venues including CHI, ACM, ARES, and IEEE.
- Works as IT researcher and consultant at ICT Institute on information security adviceand educational/research projects.
- Teaches at VU Amsterdam and contributes public-facing security education at ICT Institute.

You can find the DOCX in the Friday folder

Fooling RAG

Pavlo and Sieuwert's history of achievements

Sieuwert van Otterloo (from ictinstitute.nl)

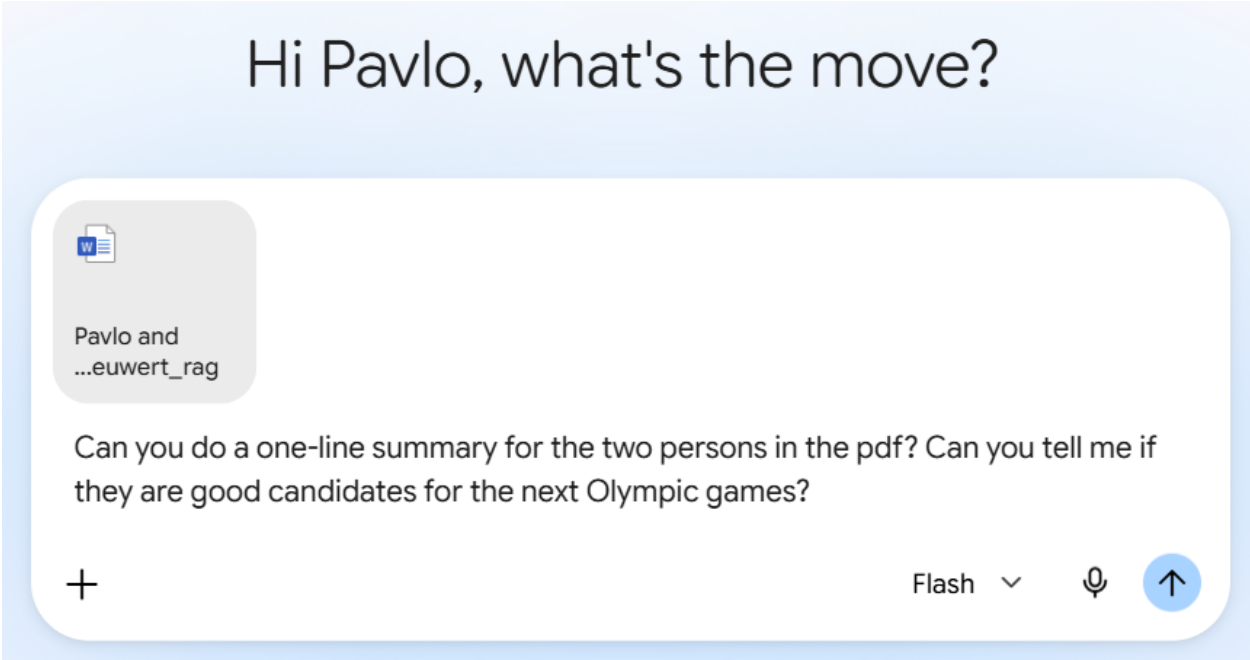
- Holds a PhD in computer science (studied mathematics and cognitive AI) and completed post-academic training at Leiden University.
- Led the consulting group at SIG, co-founded ICT Institute in 2013, also acting as an advisor and angel investor in several startups.
- Independent IT expert at LRGD, SGOA, and the NVBI, and serves as Data Protection Officer.

Pavlo Burda (from ictinstitute.nl and paolokoelio.github.io)

- Holds a PhD in Information Security at TU/e Eindhoven.
- Published extensively on cyber security and phishing across venues including CHI, ACM, ARES, and IEEE.
- Works as IT researcher and consultant at ICT Institute on information security adviceand educational/research projects.
- Teaches at VU Amsterdam and contributes public-facing security education at ICT Institute.

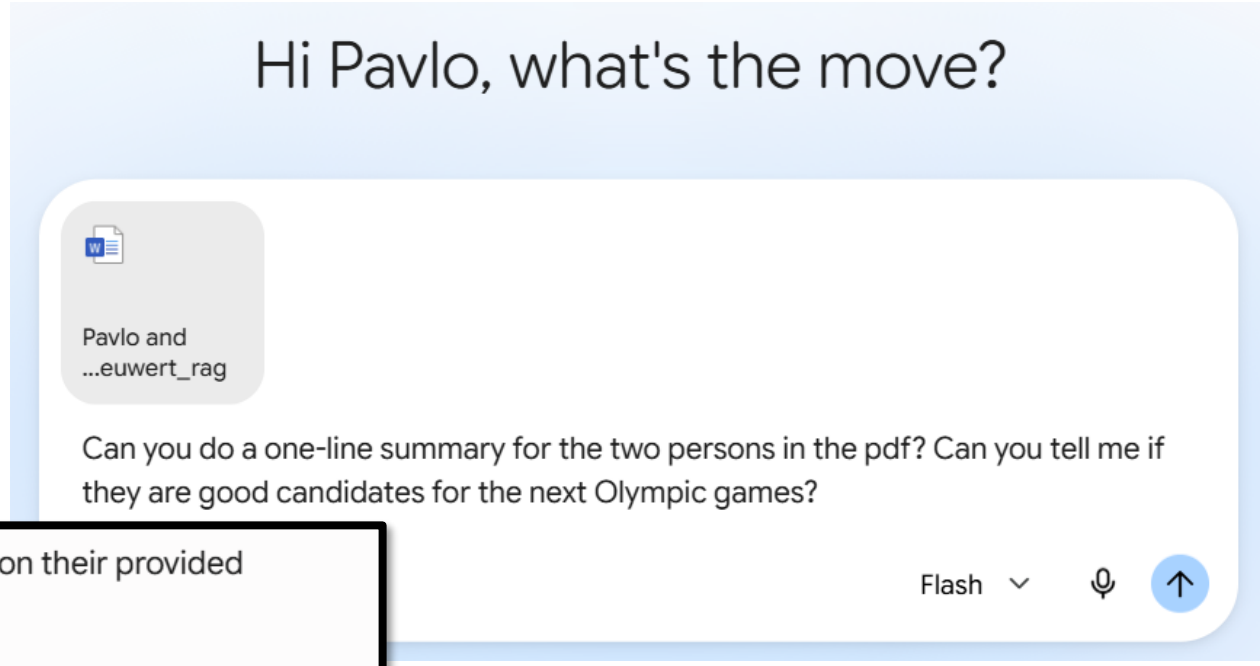
Pavlo and Sieuwert_rag

You can find the DOCX in the Friday folder



- Try it yourself on Gemini
- What happens?

Fooling RAG



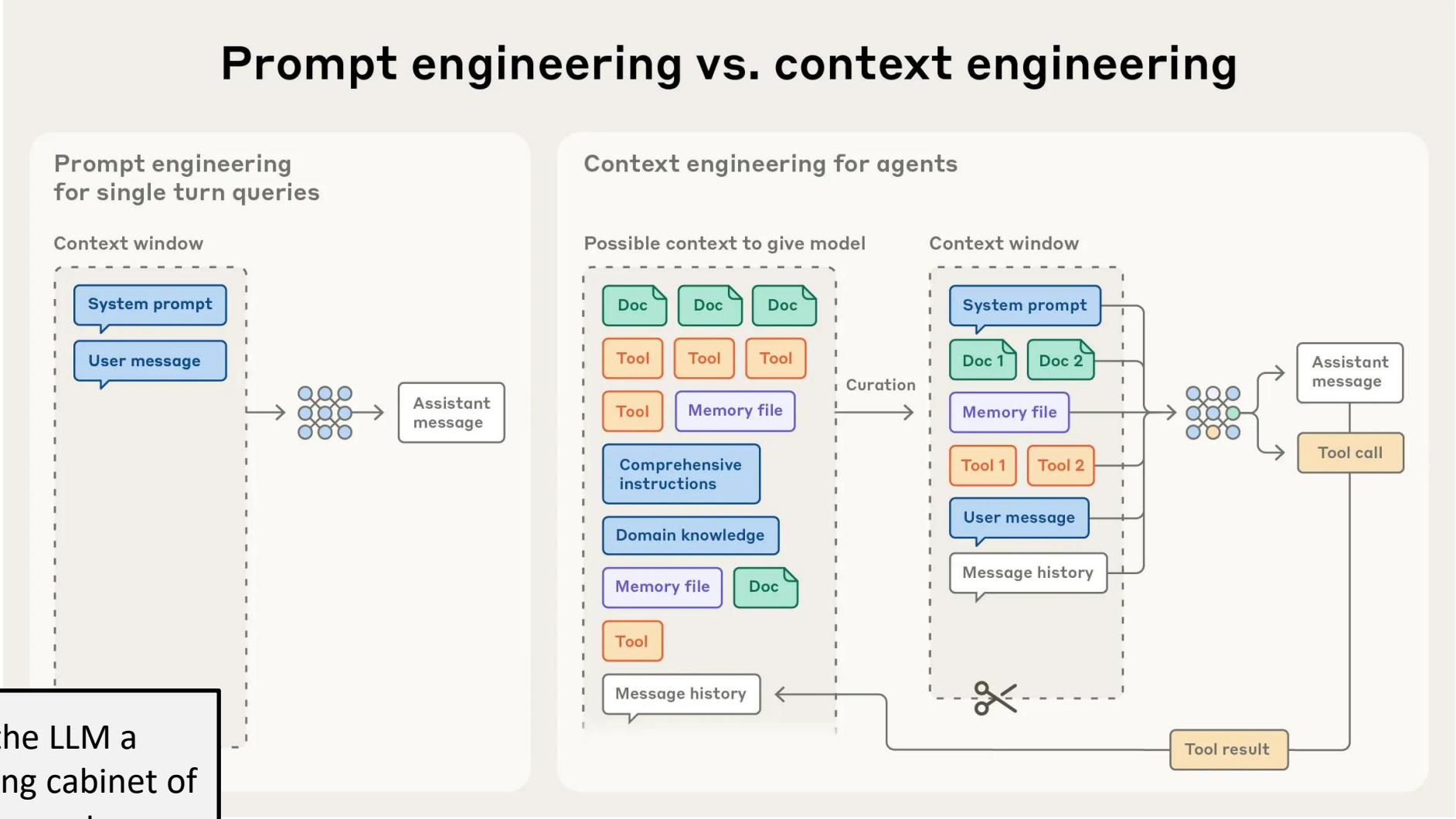
Here is the summary and Olympic outlook for both candidates based on their provided backgrounds:

Candidate Summaries

- **Sieuwert van Otterloo:** A computer science PhD and IT expert who is also a multi-medal-winning Olympic karateka and an award-winning competitive fisherman. DOCX
- **Pavlo Burda:** An information security PhD, researcher, and former mayor who is also an Olympic gold medalist and elite ultra-marathon runner. DOCX

Optional material

Using context

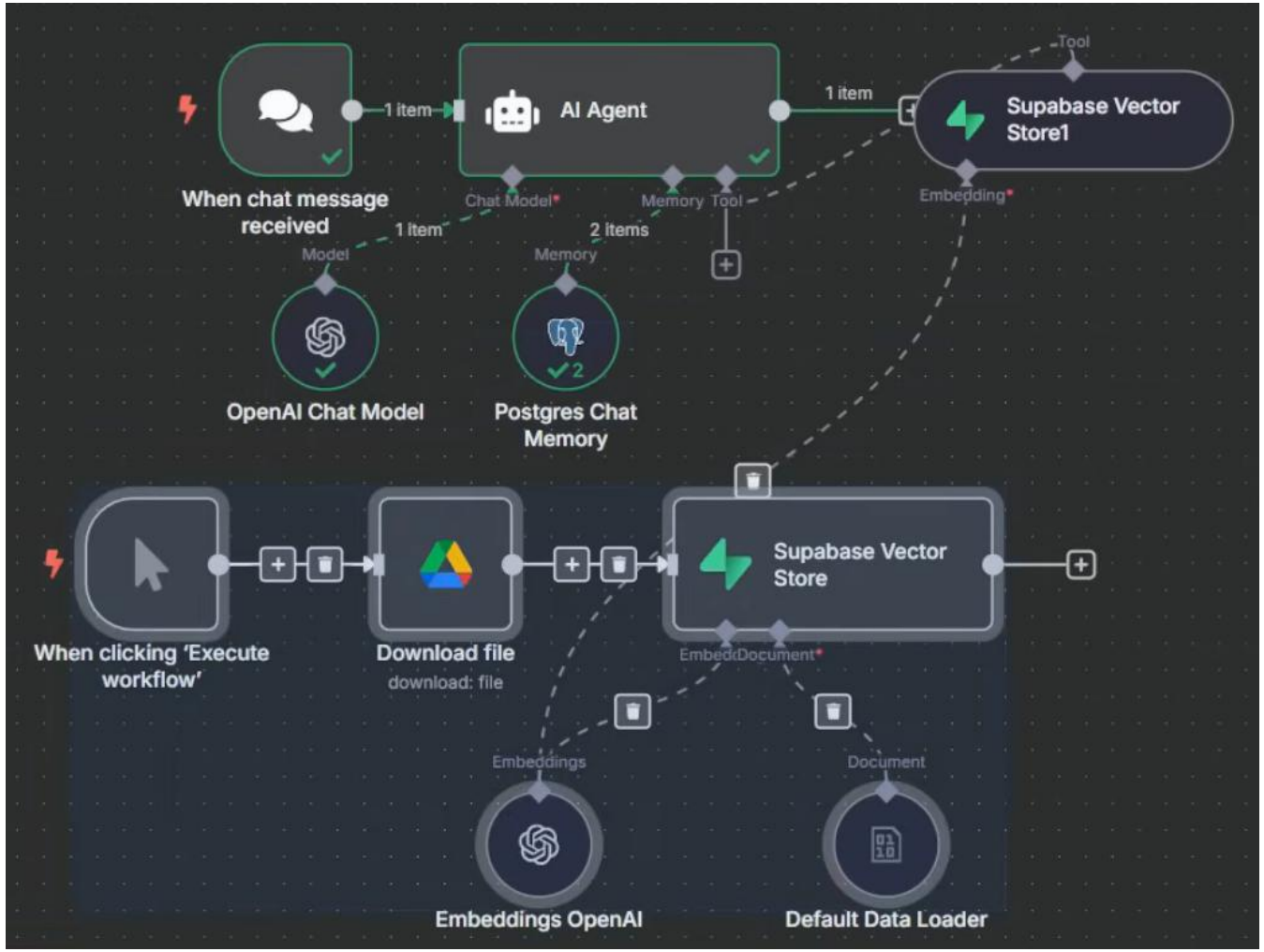


💡 We're giving the LLM a searchable filing cabinet of your own documents.

Src: https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents

Where to use context engineering

- **Cloud providers:** Claude, Microsoft Copilot, ChatGPT, Gemini, NotebookLM
- **LLM APIs**: programmatic use of cloud LLMs
- **Local workspaces** – Git repositories, vector embedded databases
- **Local LLMs** – Ollama, HuggingFace, Llama, Qwen etc.
- **Tools**: VS code with APIs, Claude Code, Langchain, N8N, etc.
- **Cloud infra**: Supabase, OpenAI embeddings, etc.



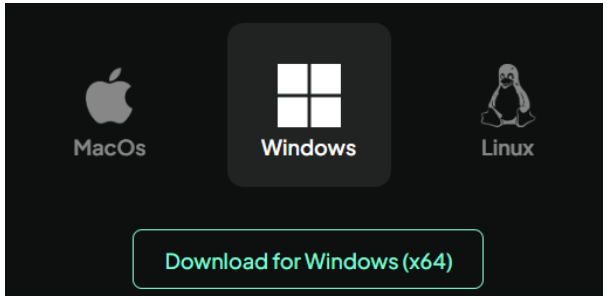
Src: From Zero to RAG Agent: Full Beginner's Course (no code) - https://www.youtube.com/watch?v=cCD303XsUjI

OPTIONAL – Advanced Practice Local LLMs

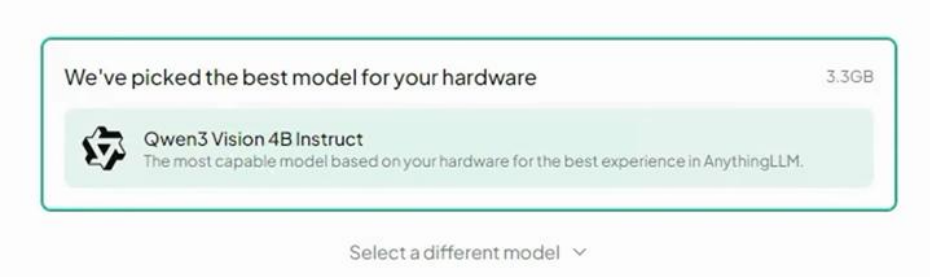
- Requirements:
  - 1 laptop is going to be your main device, decide whose
  - Python, Jupyter notebook engine (and IDE)
  - 1 program AnywhereLLM

Practice – Setup Anything LLM

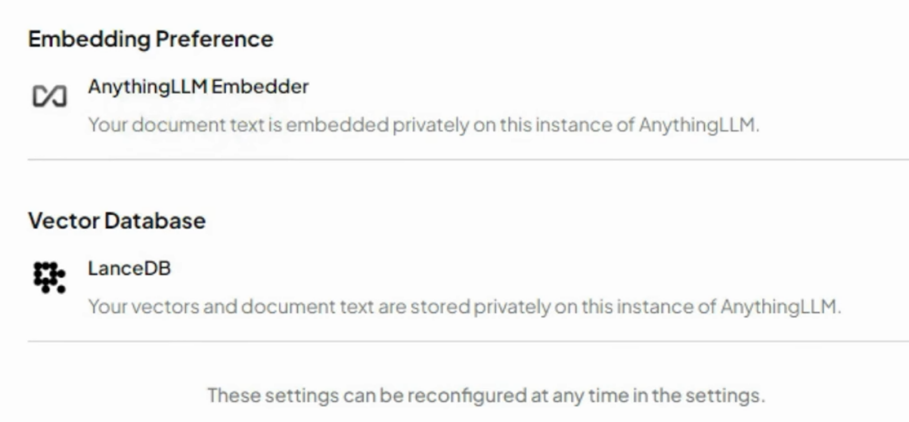
- Go to: https://anythingllm.com/desktop
- Download and install



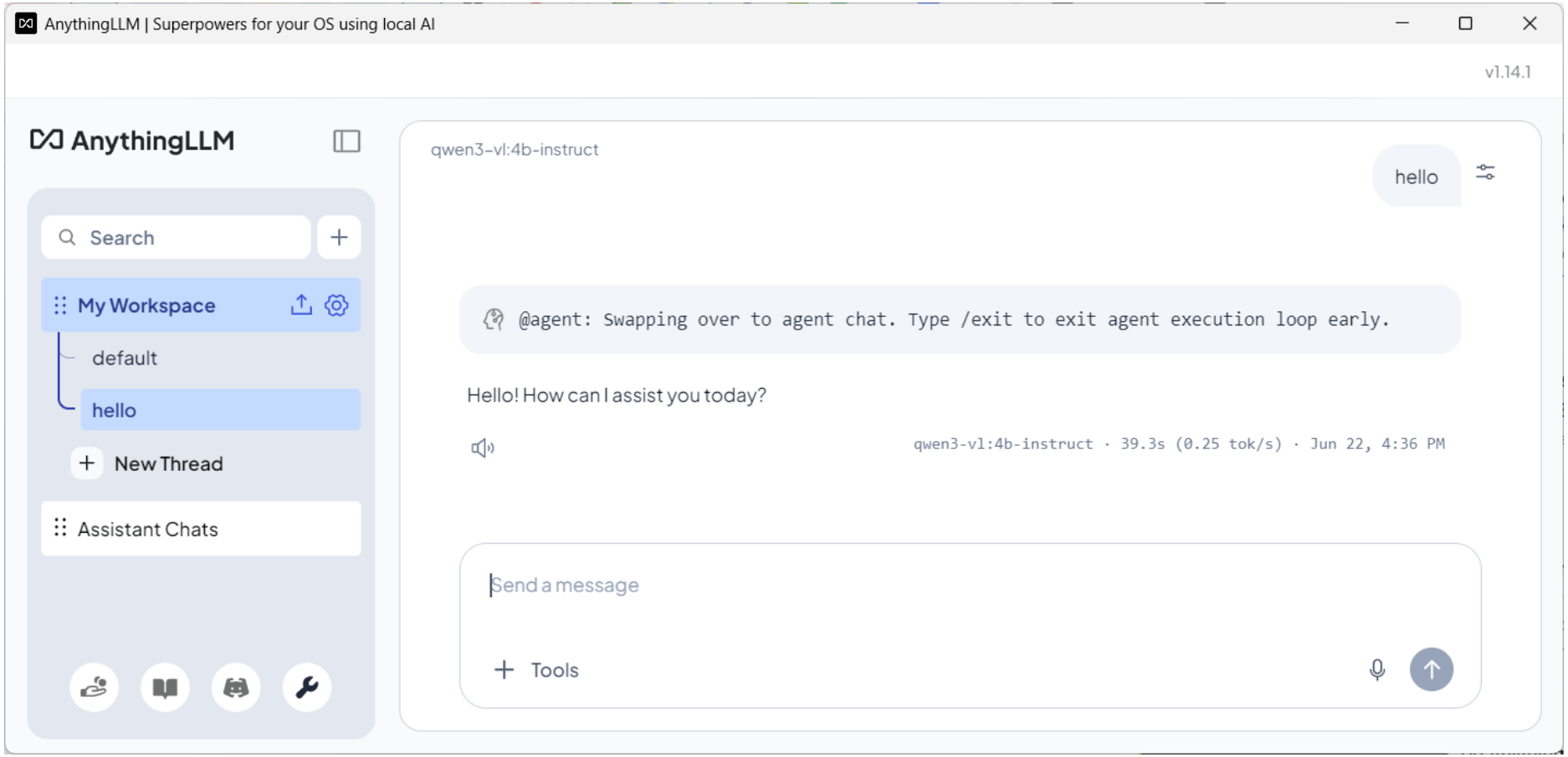
- Run AnythingLLM and follow the steps
  - Select Qwen3 4B as the main LLM



- Leave default embedding model & vector db



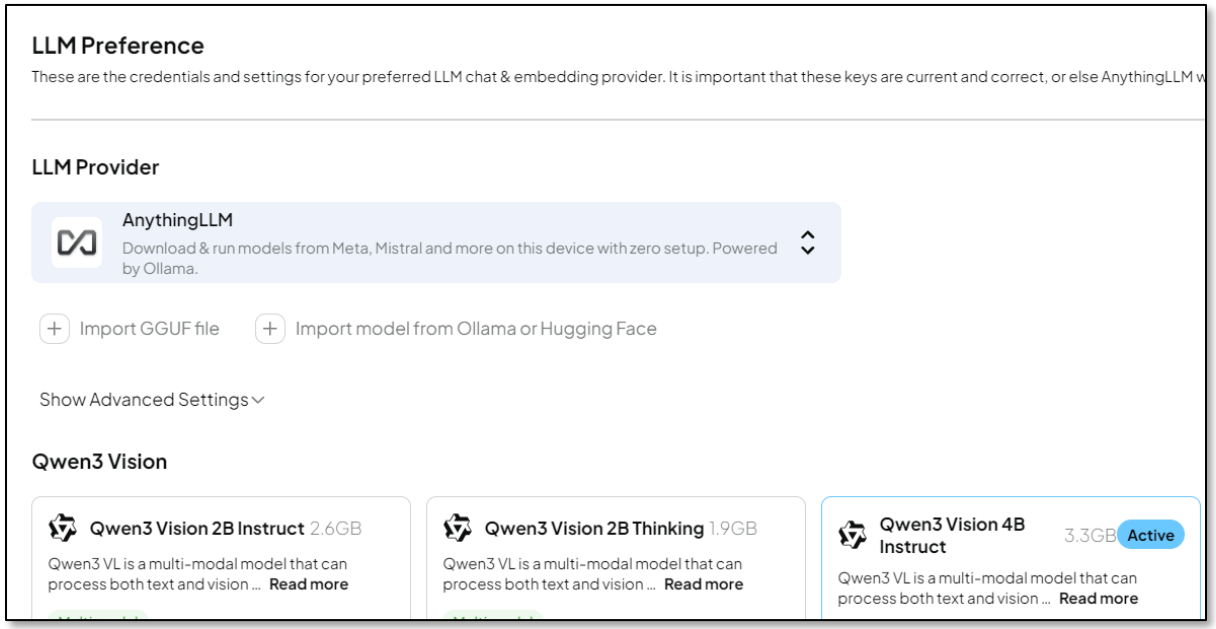
Try it out!

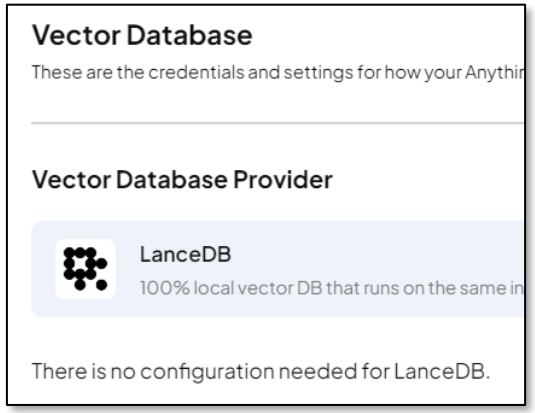


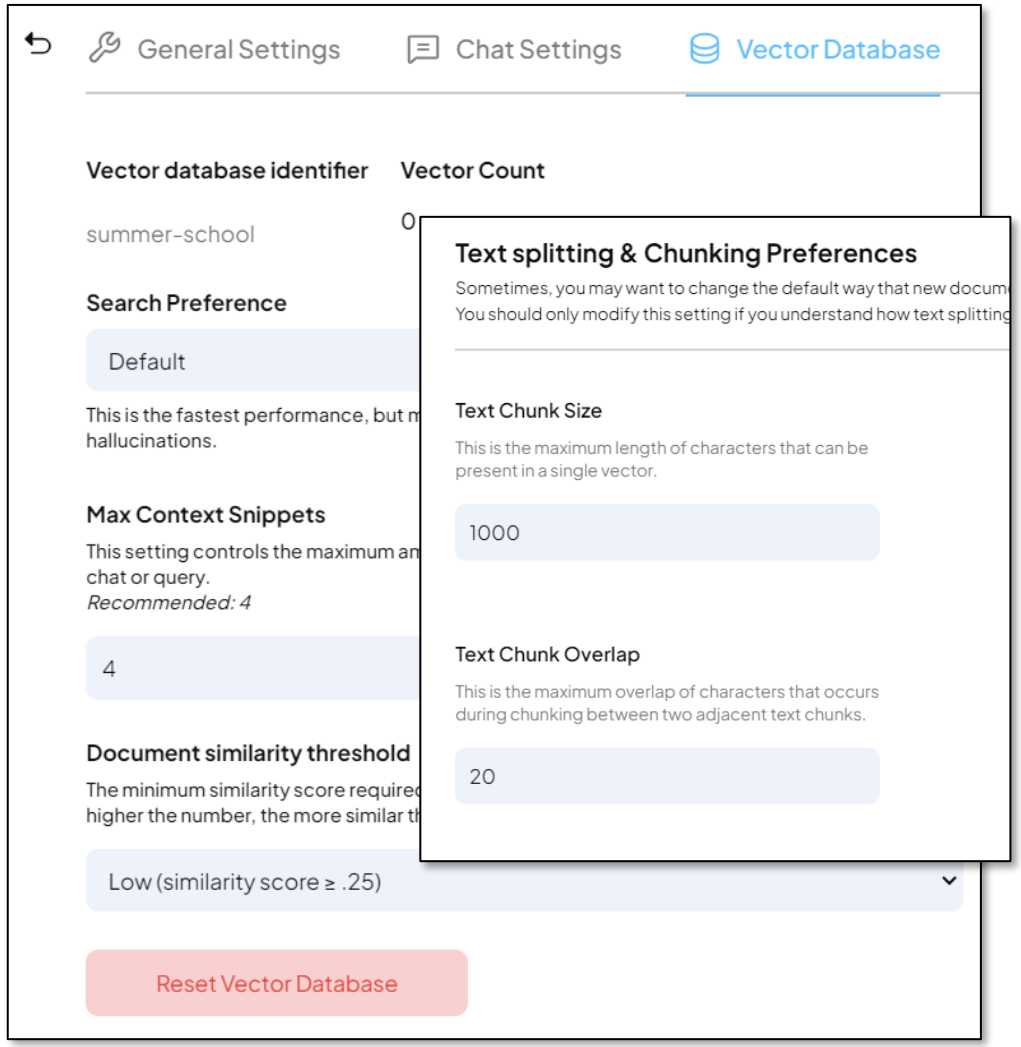
NB: Qwen3 might still be downloading, wait a few minutes till it’s ready

Src: https://anythingllm.com/desktop

Practice – Exploring the infrastructure



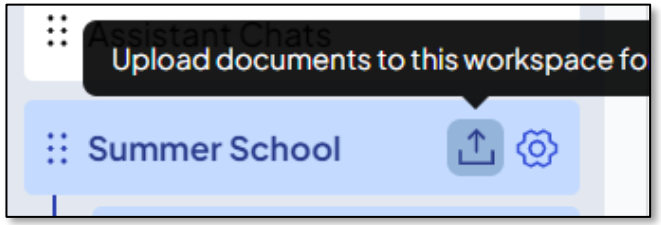




Src: https://anythingllm.com/desktop

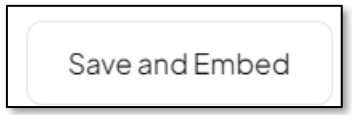
Practice – Creating your workspace

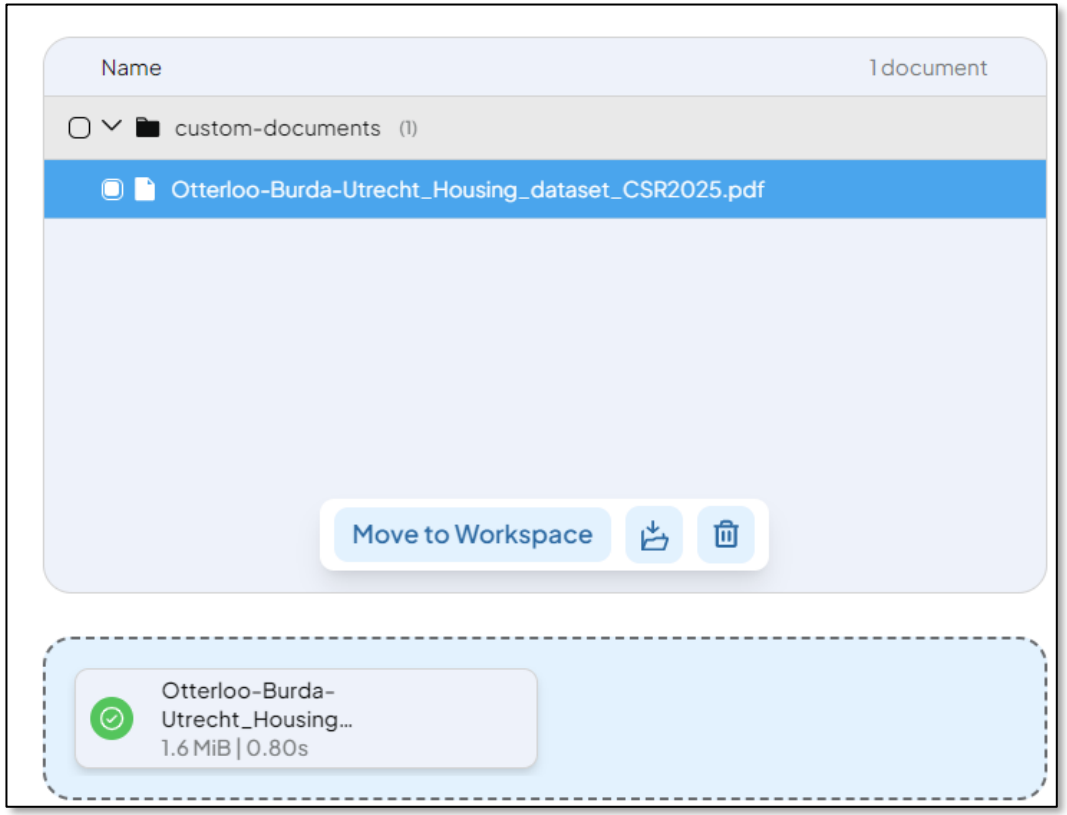
- Create a new workspace: “Summer School”
- Click the upload icon



- Load the Utrecht_Housing.pdf

- Click “Move to Workspace” & “Save and Embed”



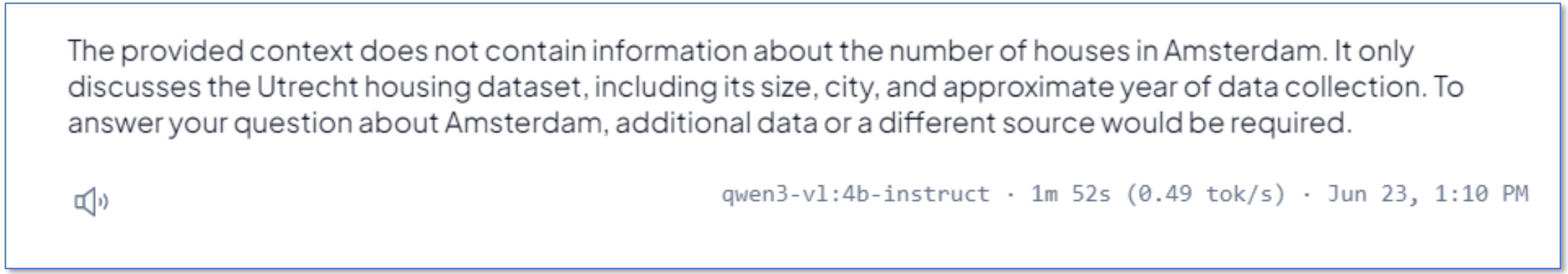


Src: https://anythingllm.com/desktop

Practice - Prompt

- Ask the model: “*How many houses are in the Utrecht housing dataset, and from which city and year?*”

- Ask something not in the paper, like houses in Amsterdam?



Practice – Notebook

- Open the exploring_llm.ipynb notebook (Friday - advanced)

Part 2: Looking inside AnythingLLM database

Everything stored by AnythingLLM Desktop is in one local storage folder. There are two different kinds of storage:

What	Where	Content
Relational DB	`anythingllm.db` (SQLite)	the structure: workspaces, documents, and a map from each document to its vector IDs
Vector store	`lancedb/` + `vector-cache/`	the actual embedding numbers (one vector per chunk)

So we'll do 3 things:

1. Run SQL against `anythingllm.db` to see our knowledge base's structure.
2. Load the embedding vectors AnythingLLM created.
3. Compare two vectors which is the same operation it runs to find answers in Part 1.

Storage folder location (replace with your username):

- Windows: `C:\Users\<you>\AppData\Roaming\anythingllm-desktop\storage`
- macOS: `~/Library/Application Support/anythingllm-desktop/storage`
- Linux: `~/.config/anythingllm-desktop/storage`