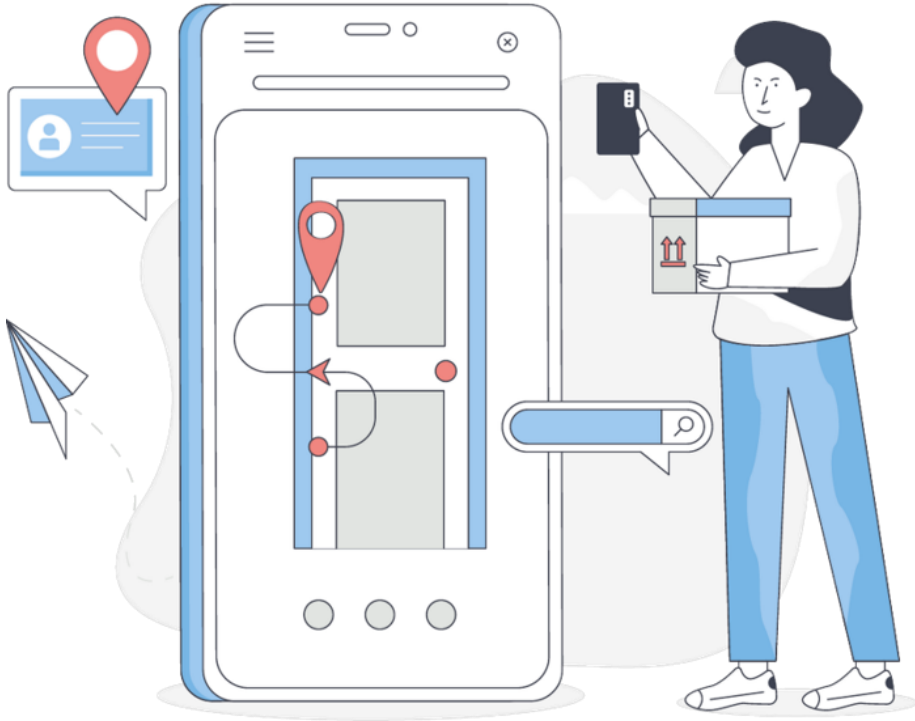




Exploring local LLMs and RAG

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



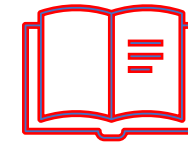
Notes

meeting & research notes



Policies

HR, security, compliance



Contracts

Customers, financial, health

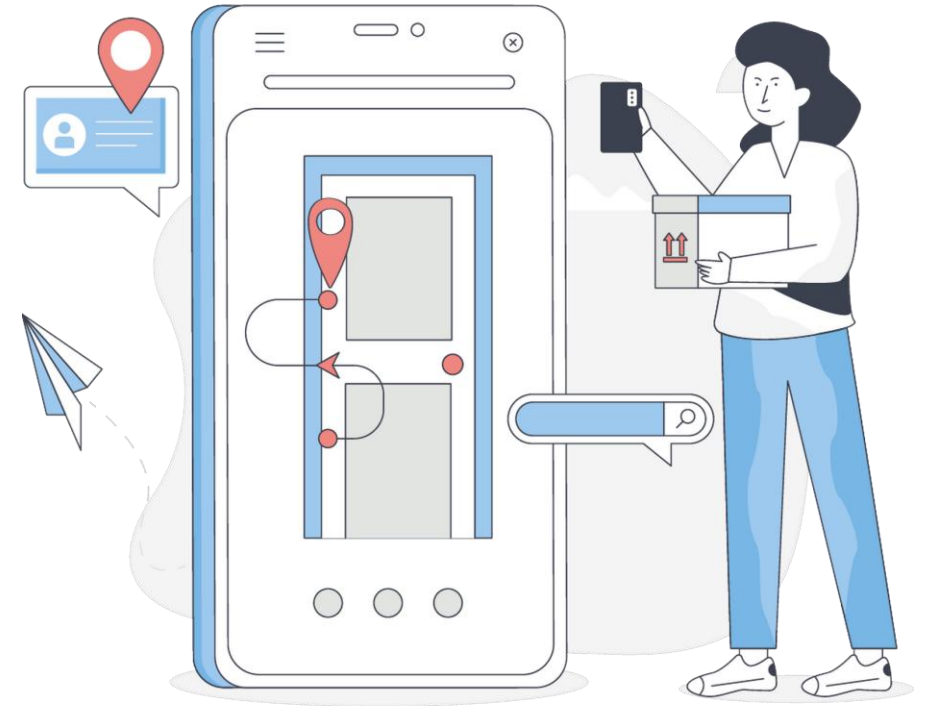


Instrument your workflow

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



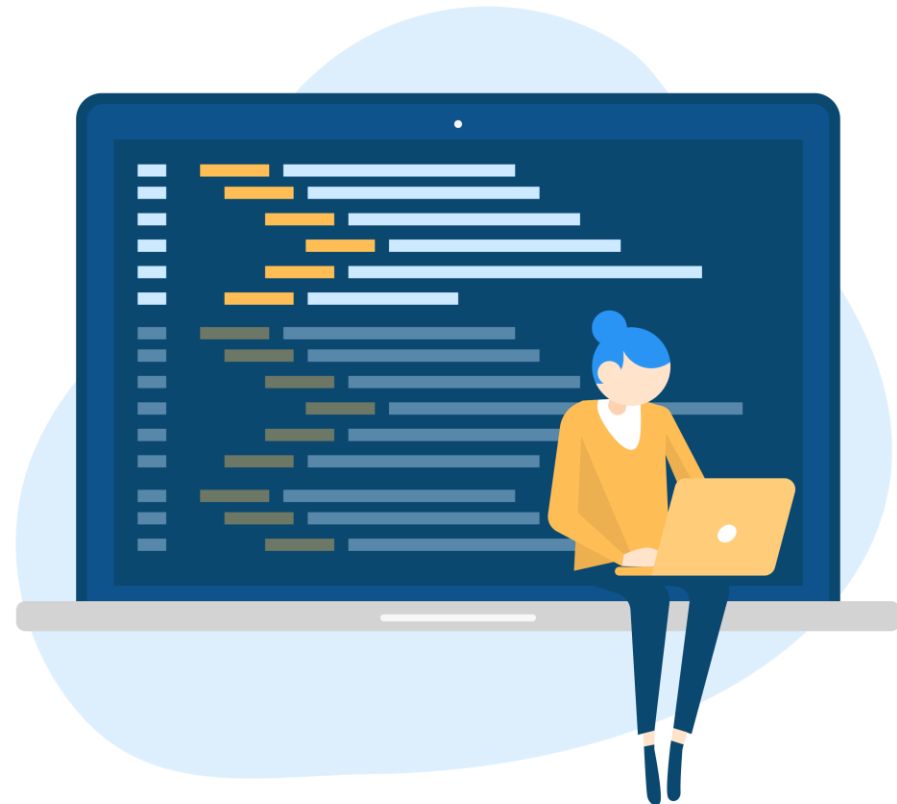
By Naveed Hussain on IconScout

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

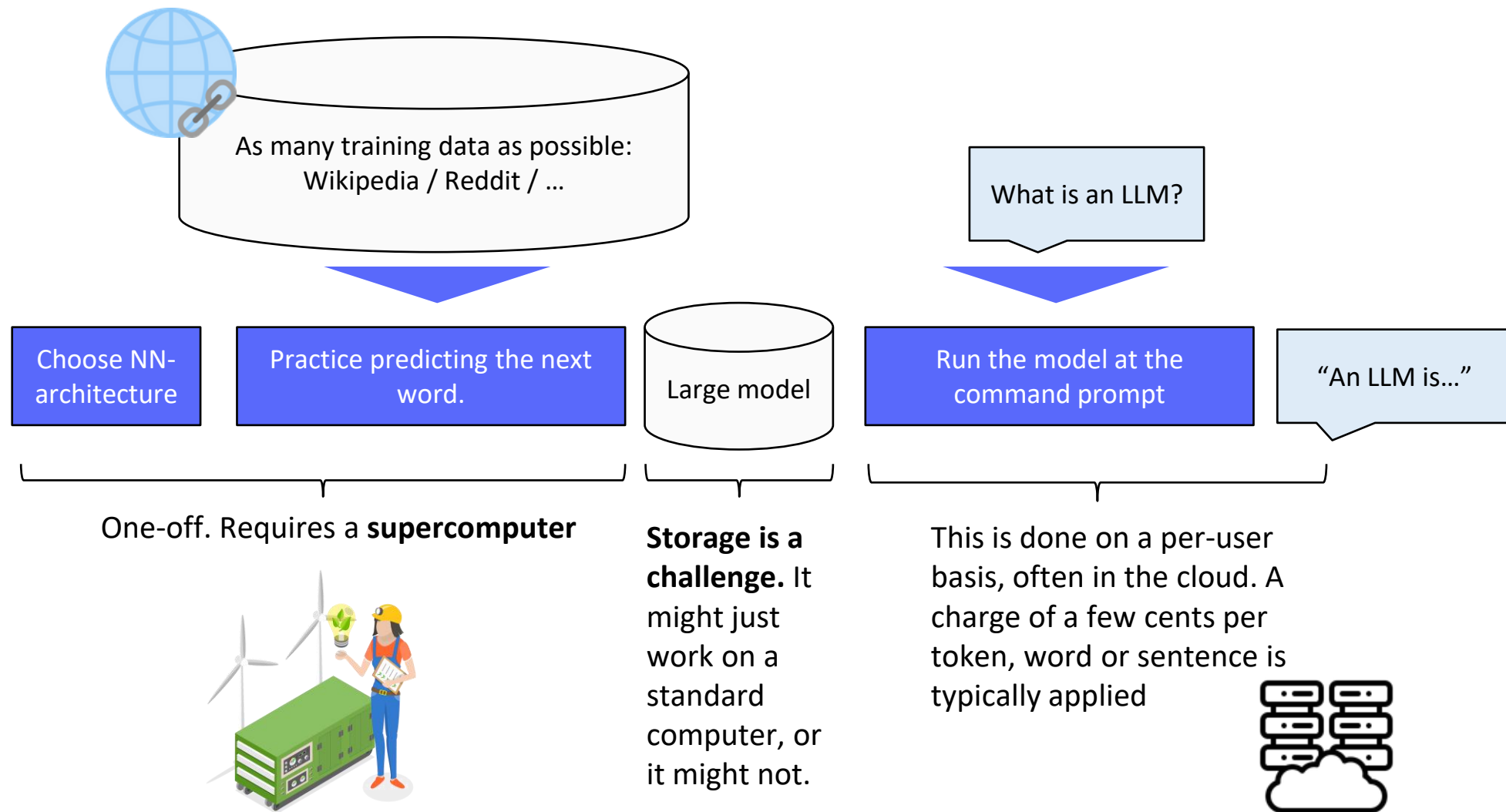


By Deesign Graphics on IconScout

Context engineering

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

What is a Large Language Model?



By Anastasiia Torianyk on IconScout

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

The screenshot shows the Hugging Face website interface. At the top, there's a navigation bar with the Hugging Face logo, a search bar, and links for Models, Datasets, Spaces, Buckets, Docs, Enterprise, Pricing, Log In, and Sign Up. Below the navigation bar, the left sidebar contains tabs for Main, Tasks, Libraries, Languages, Licenses, and Other. The Tasks section is expanded, showing various task categories like Text Generation, Any-to-Any, Image-Text-to-Text, Image-to-Text, Image-to-Image, Text-to-Image, Text-to-Video, and Text-to-Speech. Below the tasks, there's a Parameters section with a slider ranging from <1B to >500B. The Libraries section shows various frameworks like PyTorch, TensorFlow, JAX, Transformers, Diffusers, GGUF, MLX, Transformers.js, Safetensors, sentence-transformers, and ONNX. The main content area displays a list of models with filters for Base only, Inference Available, and Sort: Most likes. The models listed include deepseek-ai/DeepSeek-R1, black-forest-labs/FLUX.1-dev, stabilityai/stable-diffusion-xl-base-1.0, CompVis/stable-diffusion-v1-4, meta-llama/Meta-Llama-3-8B, hexgrad/Kokoro-82M, meta-llama/Llama-3.1-8B-Instruct, openai/whisper-large-v3, black-forest-labs/FLUX.1-schnell, and bigscience/bloom.

Hugging Face Search models, datasets, users...

Models 2,921,270 Filter by name

Base only Inference Available Sort: Most likes

Tasks

- Text Generation
- Any-to-Any
- Image-Text-to-Text
- Image-to-Text
- Image-to-Image
- Text-to-Image
- Text-to-Video
- Text-to-Speech
- + 44

Parameters

<1B 6B 12B 32B 128B >500B

Libraries

- PyTorch
- TensorFlow
- JAX
- Transformers
- Diffusers
- GGUF
- MLX
- Transformers.js
- Safetensors
- sentence-transformers
- ONNX
- + 42

Models

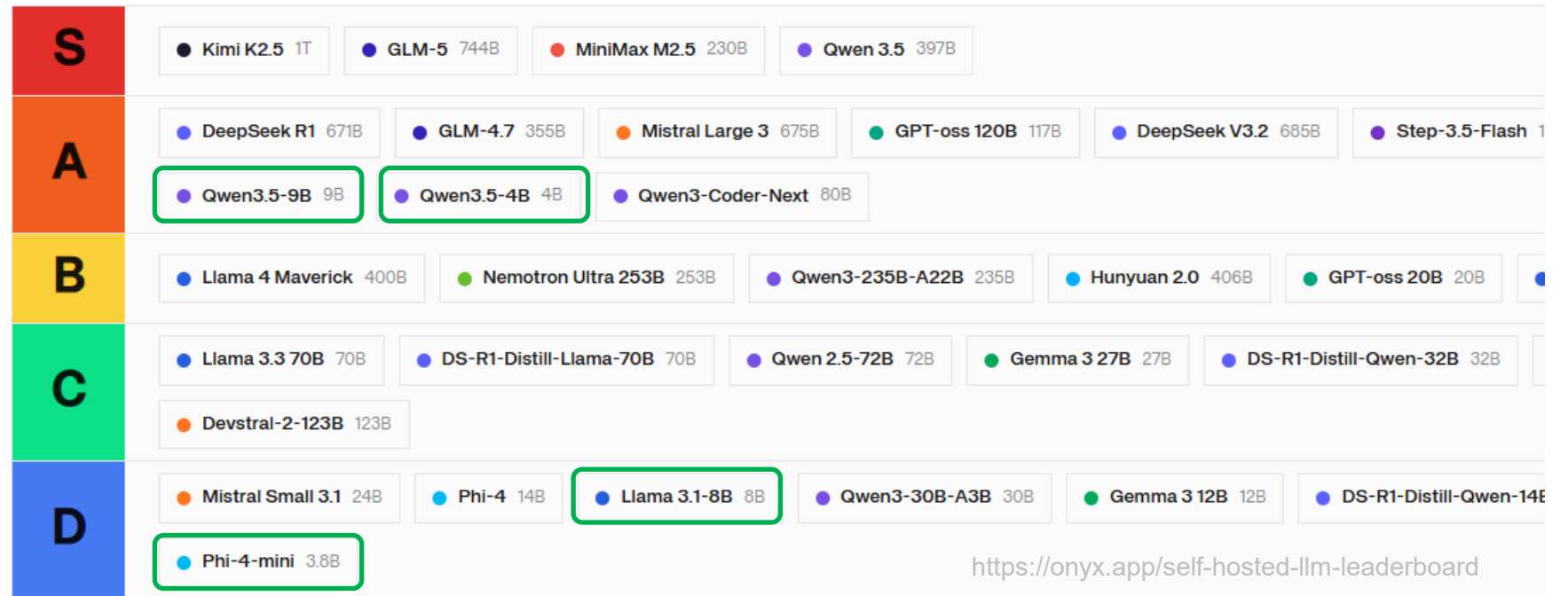
- deepseek-ai/DeepSeek-R1**
Text Generation • 685B • Updated Mar 27, 2025 • 6.91M • 13.4k
- black-forest-labs/FLUX.1-dev**
Text-to-Image • Updated Jun 27, 2025 • 1.12M • 13.3k
- stabilityai/stable-diffusion-xl-base-1.0**
Text-to-Image • Updated Oct 30, 2023 • 1.42M • 7.84k
- CompVis/stable-diffusion-v1-4**
Text-to-Image • Updated Aug 23, 2023 • 430k • 7.03k
- meta-llama/Meta-Llama-3-8B**
Text Generation • 8B • Updated Sep 27, 2024 • 1.26M • 6.58k
- hexgrad/Kokoro-82M**
Text-to-Speech • Updated Apr 10, 2025 • 16.6M • 6.37k
- meta-llama/Llama-3.1-8B-Instruct**
Text Generation • 8B • Updated Sep 25, 2024 • 9.9M • 6.13k
- openai/whisper-large-v3**
Automatic Speech Recognition • 2B • Updated A... • 5.88M • 5.85k
- black-forest-labs/FLUX.1-schnell**
Text-to-Image • Updated Aug 16, 2024 • 227k • 5.18k
- bigscience/bloom**
Text Generation • 176B • Updated Jul 28, 2023 • 5.9k • 5.01k

<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

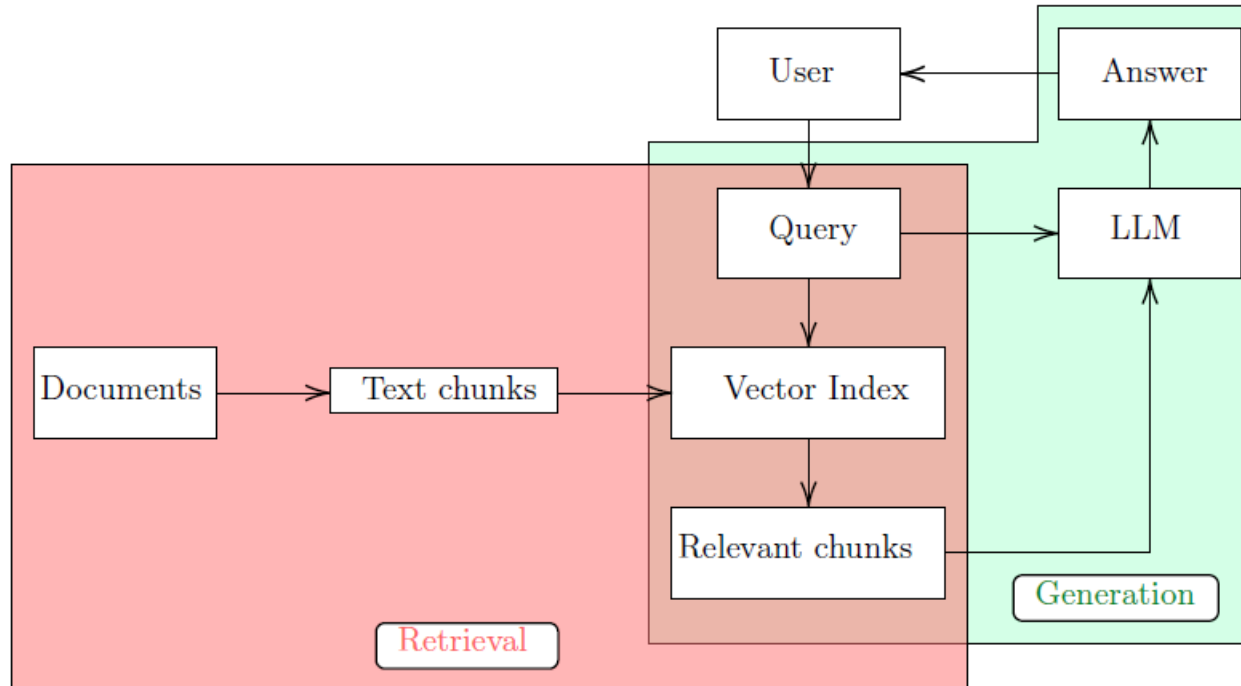
LLMs are trained on a general corpus of human knowledge and interactions (mostly internet, Wikipedia, Reddit, pirated books ...).

Still, they do not know anything about your data.

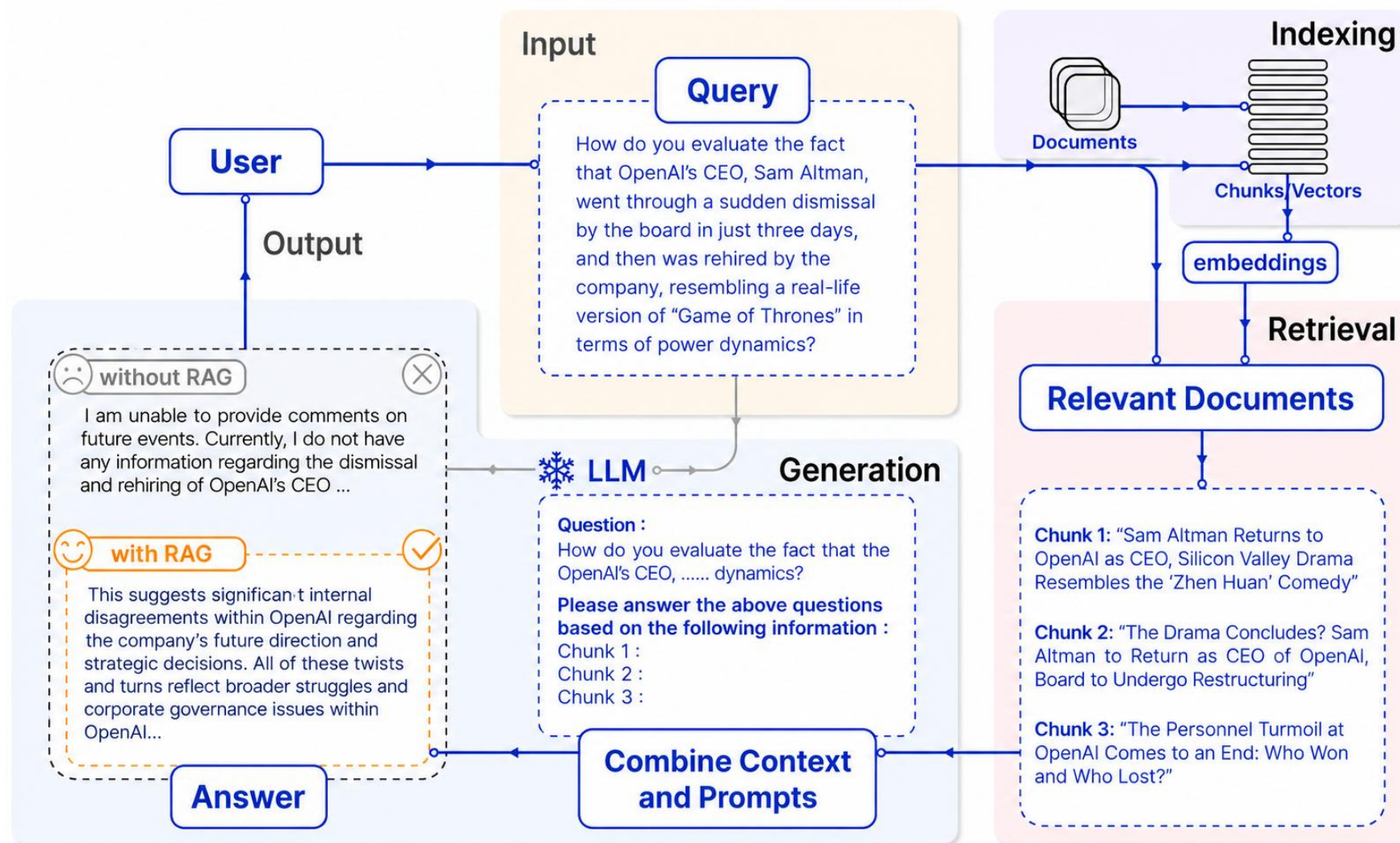
The content and quality of Local LLM outputs depends on what and how you provide information to the model:

- (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



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>

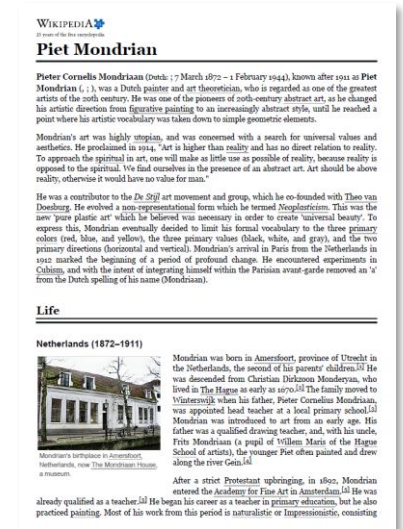
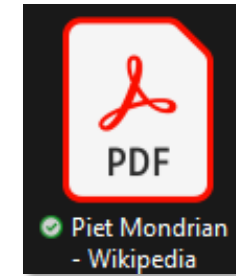
Interactive RAG Playground

Debug, visualize, and master Retrieval-Augmented Generation through hands-on experiments and real-time demonstrations

Start Experiment →

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?

Text Splitting

Split Strategy: Recursive Character

Chunk Size: 500

Overlap Size: 50

Separators:

Source Document

Sample text from [AWS Documentation](#)

```
in Germany, where it is now held. The gallery explained that it would
continue to display it the wrong way up to avoid damaging it.
[
31
]
[
32
]
]
His painting
Broadway Boogie Woogie
(1942-43) at the Museum of Modern Art was highly influential in the
school of
abstract geometric
painting. The piece is made up of a number of shimmering squares of
bright color that leap from the canvas, drawing the viewer into those
neon lights. In this painting and the unfinished
```

Generated Chunks

Hover over chunks to highlight their position in the source document

Chunks: 55

Avg. Size: 385 chars

With Overlap: 35

Avg. Overlap: 29 chars

500

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

484

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

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 greataartists 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 universal

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

[0.0021, 0.0478, 0.0022, 0.0156, -0.0760, 0.0136, -0.0057, -0.0487...]

Chunk 3 • 384 dimensions

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 uni

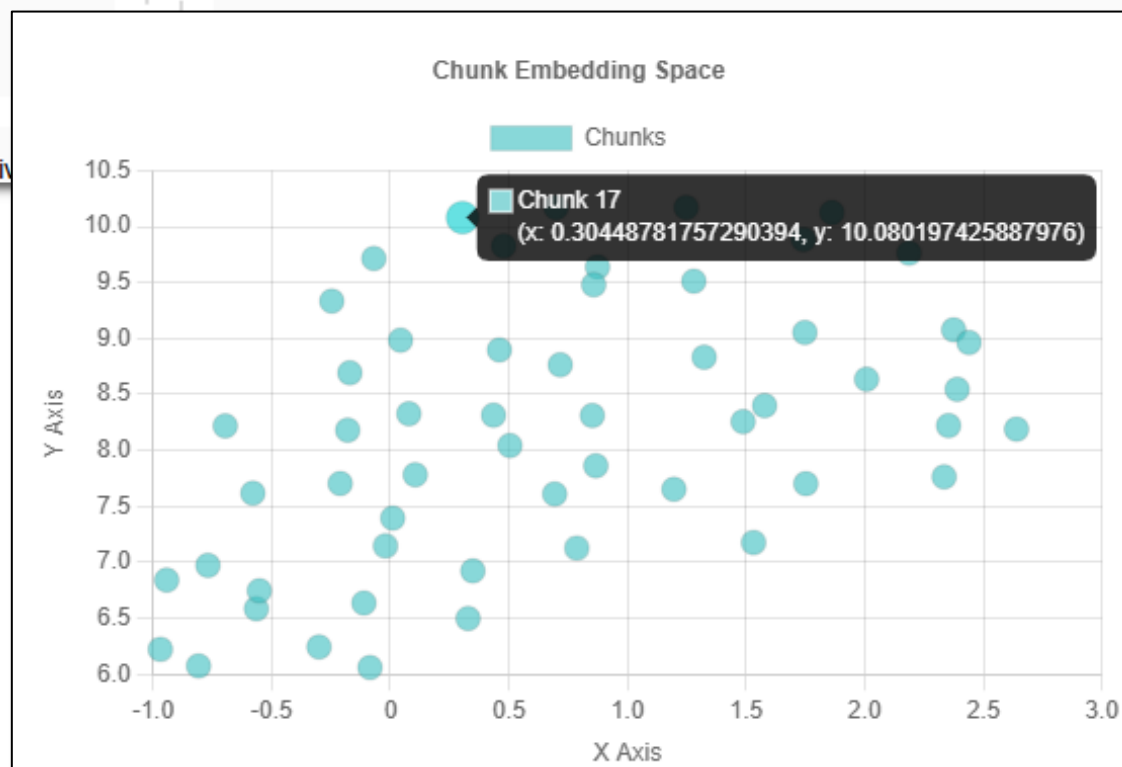
Chunks Embedding Vectors:

[-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



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 the addition of a colored
Chunk 46 • Similarity: 0.8607

Practice RAG

Temperature: 0.3 Max Tokens: 1000

Prompt Preview:

System Message Edit Message

You are a helpful AI assistant. Use the following pieces of context to answer the user's question.
If you don't know the answer, just say that you don't know. Don't try to make up an answer.

Context:
amidst the familiar black lines, creating a new sense of depth by the addition of a colored layer on top of the black one. His painting

paintings of this period came to be increasingly dominated by white space.

in appearance. He spent many long hours painting on his own until his hands blistered, and he sometimes cried or made himself sick.

User Message Edit Message

What was Mondriaan's painting style?

Model Response:

- Generation pipeline
- Collected 3 retrieved context chunks
- Composed system instructions and user question
- Generated a grounded response

Based on the given context, it appears that Mondriaan's painting style involved the use of black lines and the addition of a colored layer on top of the black one, creating a new sense of depth.

Practice RAG

User Message

Edit Message

What was Mondriaan's favourite car?

I don't know. There is no mention of Mondriaan's favorite car in the given context.

User Message

Edit Message

Who was Mondriaan's mother?

I don't know. The provided context does not mention Mondriaan's mother.

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 advice and 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

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 advice and 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

Hi Pavlo, what's the move?



Pavlo and
...euwert_rag

Can you do a one-line summary for the two persons in the pdf? Can you tell me if they are good candidates for the next Olympic games?



Flash ▾



- Try it yourself on Gemini
- What happens?

Fooling RAG

Hi Pavlo, what's the move?



Pavlo and
...euwert_rag

Can you do a one-line summary for the two persons in the pdf? Can you tell me if they are good candidates for the next Olympic games?

Flash ▾



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

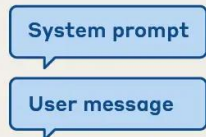
Optional material

Using context

Prompt engineering vs. context engineering

Prompt engineering for single turn queries

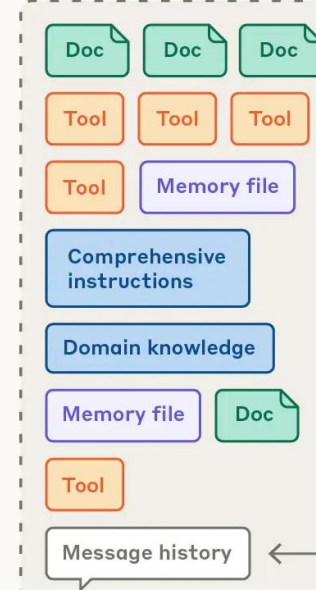
Context window



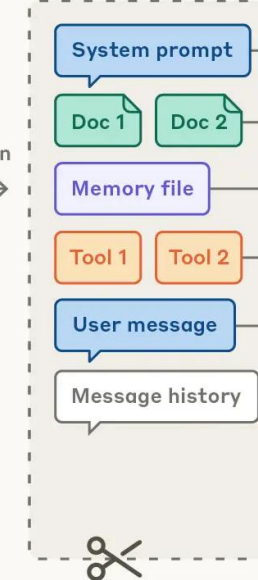
Assistant message

Context engineering for agents

Possible context to give model



Context window



Assistant message

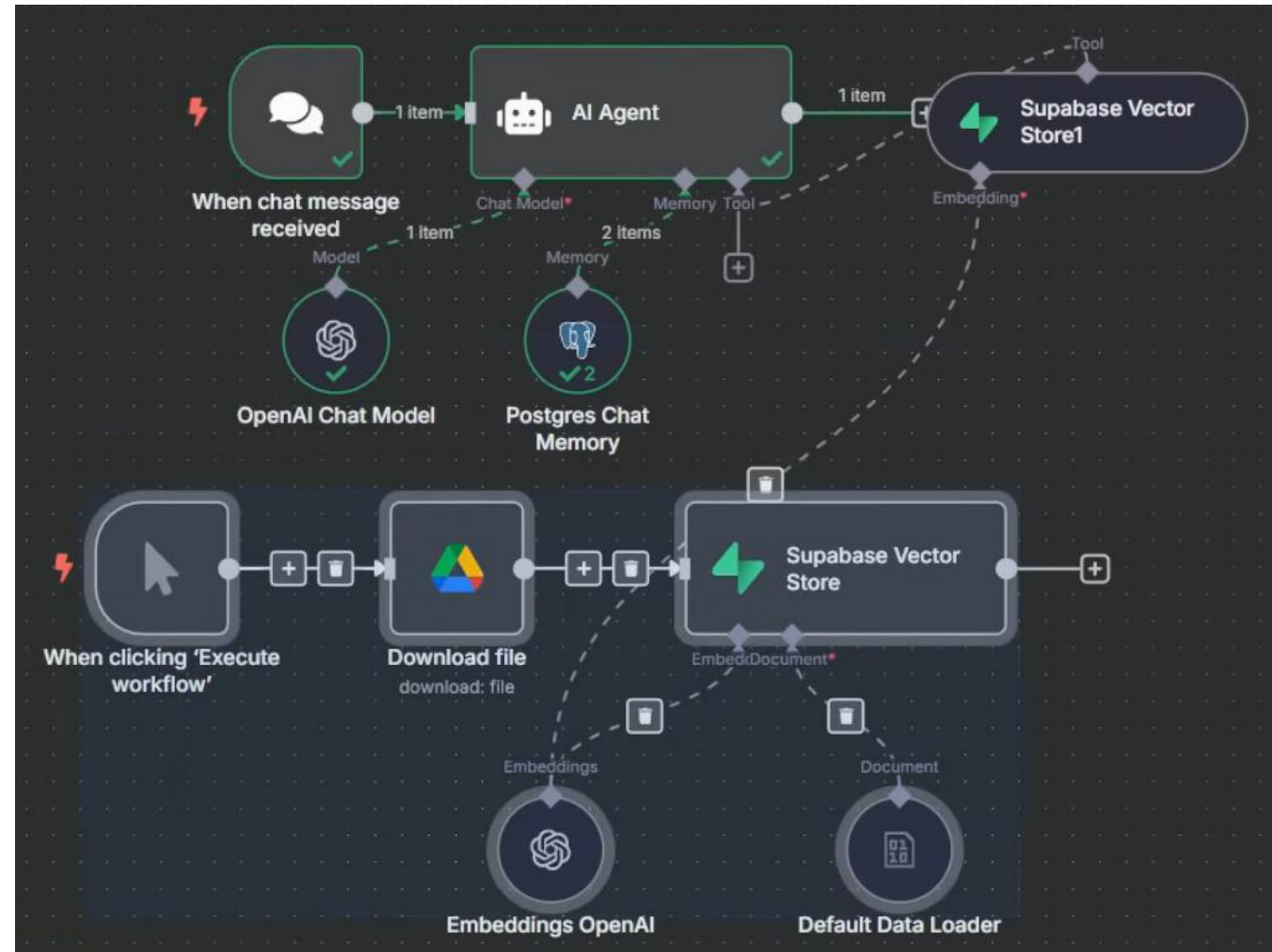
Tool call

Tool result

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

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.



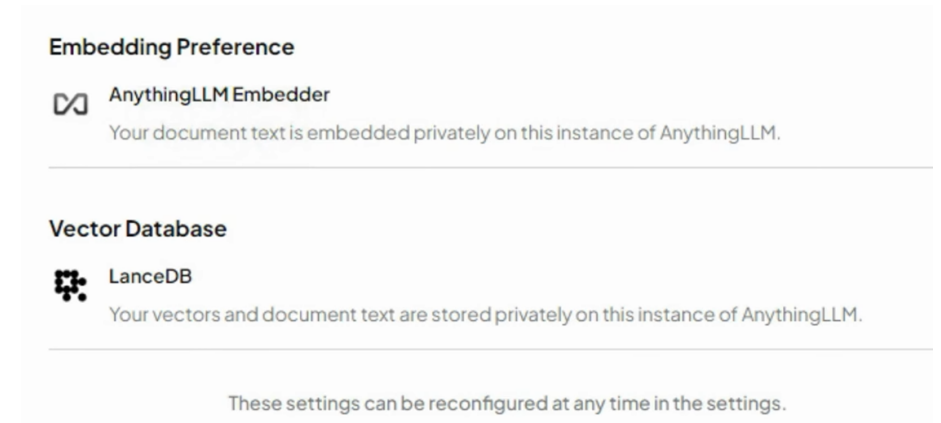
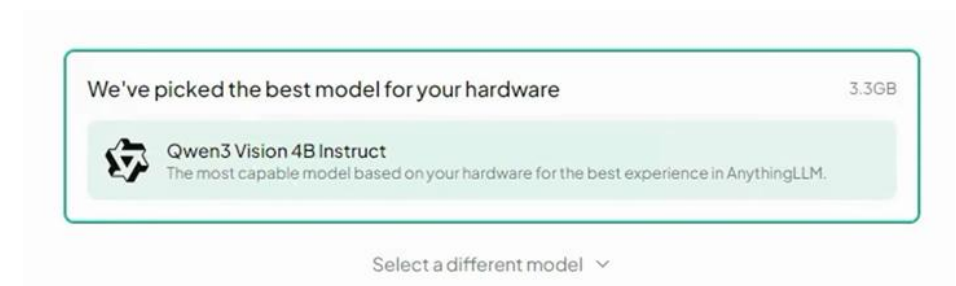
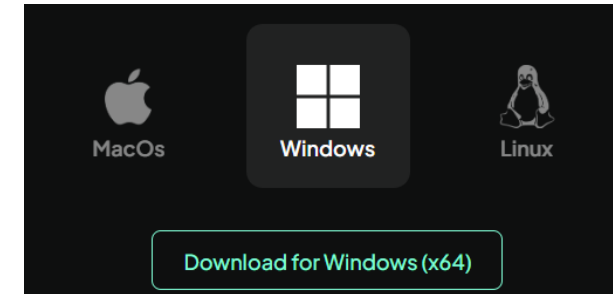
Src: From Zero to RAG Agent: Full Beginner's Course (no code) - <https://www.youtube.com/watch?v=cCD303XsUjI>

OPTIONAL – Advanced Practice Local LLMs

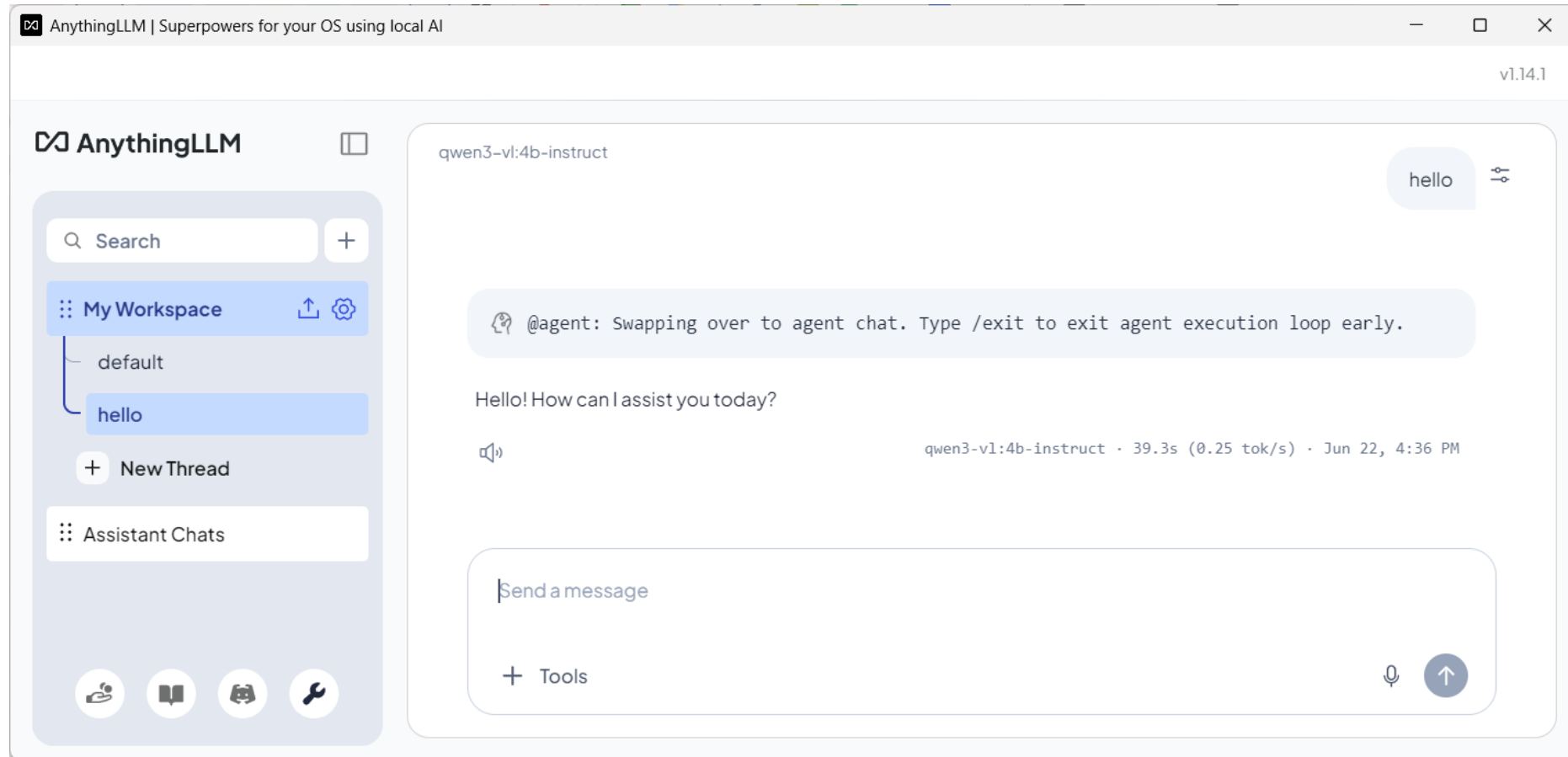
- Requirements:
 - 1 laptop is going to be your main device, decide whose
 - Python, Jupyter notebook engine (and IDE)
 - 1 program AnywhereLLM

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

Practice – Exploring the infrastructure

LLM Preference

These are the credentials and settings for your preferred LLM chat & embedding provider. It is important that these keys are current and correct, or else AnythingLLM will not work.

LLM Provider

 AnythingLLM

Download & run models from Meta, Mistral and more on this device with zero setup. Powered by Ollama.

+ Import GGUF file

+ Import model from Ollama or Hugging Face

Show Advanced Settings ▾

Qwen3 Vision

 Qwen3 Vision 2B Instruct 2.6GB

Qwen3 VL is a multi-modal model that can process both text and vision ... [Read more](#)

 Qwen3 Vision 2B Thinking 1.9GB

Qwen3 VL is a multi-modal model that can process both text and vision ... [Read more](#)

 Qwen3 Vision 4B Instruct 3.3GB Active

Qwen3 VL is a multi-modal model that can process both text and vision ... [Read more](#)

Vector Database

These are the credentials and settings for how your AnythingLLM will store and retrieve data.

Vector Database Provider

 LanceDB

100% local vector DB that runs on the same machine as your LLM.

There is no configuration needed for LanceDB.

[General Settings](#)

[Chat Settings](#)

[Vector Database](#)

Vector database identifier

Vector Count

summer-school0

Search Preference

Default

This is the fastest performance, but may increase the risk of hallucinations.

Max Context Snippets

This setting controls the maximum number of snippets returned in a chat or query.
Recommended: 4

4

Document similarity threshold

The minimum similarity score required to return a document. The higher the number, the more similar the documents must be.

Low (similarity score $\geq$.25)

Reset Vector Database

Text splitting & Chunking Preferences

Sometimes, you may want to change the default way that new documents are split into chunks. You should only modify this setting if you understand how text splitting works.

Text Chunk Size

This is the maximum length of characters that can be present in a single vector.

1000

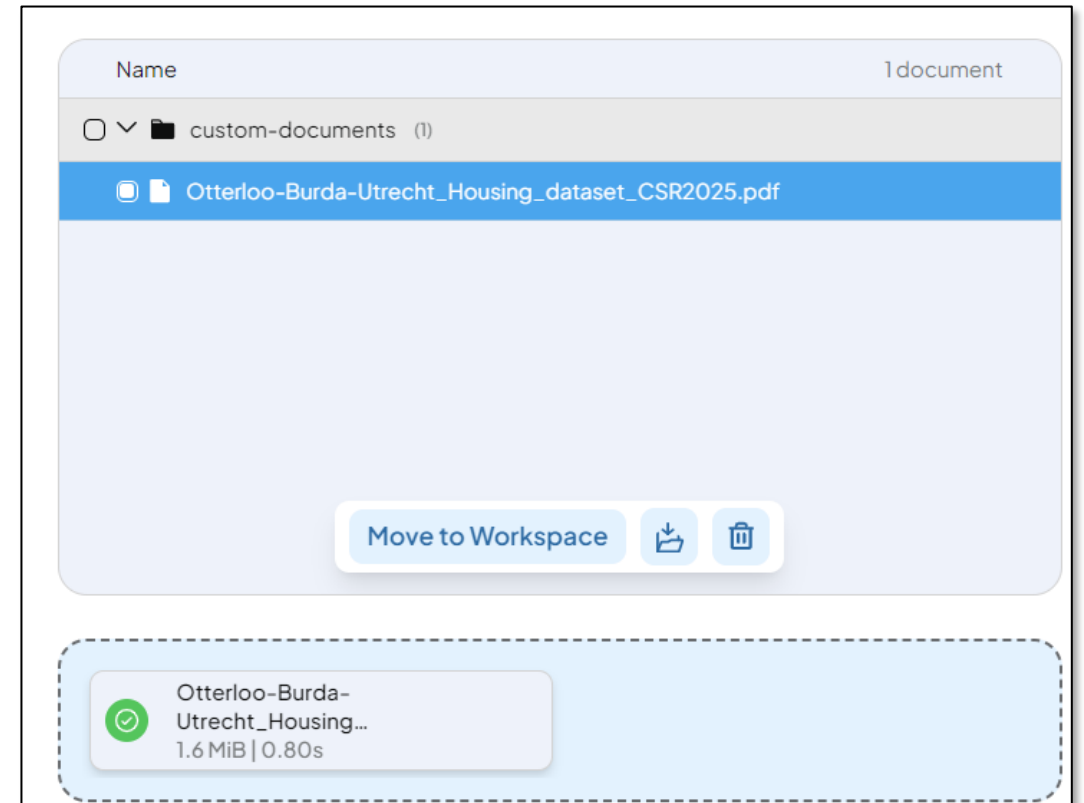
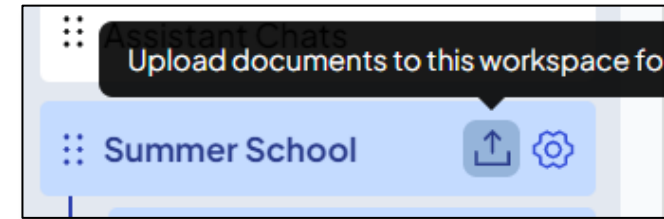
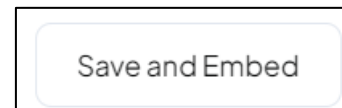
Text Chunk Overlap

This is the maximum overlap of characters that occurs during chunking between two adjacent text chunks.

20

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”



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?

The provided context does not contain information about the number of houses in Amsterdam. It only discusses the Utrecht housing dataset, including its size, city, and approximate year of data collection. To answer your question about Amsterdam, additional data or a different source would be required.



qwen3-v1:4b-instruct · 1m 52s (0.49 tok/s) · Jun 23, 1:10 PM

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	<code>anythingllm.db</code> (SQLite)	the structure: workspaces, documents, and a map from each document to its vector IDs
Vector store	<code>lancedb/</code> + <code>vector-cache/</code>	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`