

FROM COPILOT TO A TEAM OF AI SOFTWARE ENGINEERS

Loïc Gouarin



- Research engineer in scientific computing at CNRS
- Co-leader of the HPC@Maths team
- Member of the groupe Calcul board
- Developer of open-source software



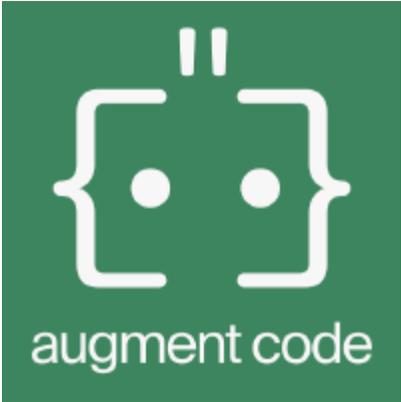
SOME COMMENTS BEFORE WE START

- I'm not an expert of LLM and their construction.
- The world of LLM is vast and the field of possibilities almost infinite.
- Some points may be wrong and lack precision.
- The learning curve is steep.
- I'm a baby of a one month and a half in this world.

WHAT CAN YOU DO WITH GITHUB COPILOT ?

- Code completion as you type (2,000 intelligent code completions a month)
- Chat discussion to explain code, to write tests or documentation, ... (50 Copilot Chat messages a month)
- Use the open files and your GitHub repositories as a knowledge base
- Web search and other agents on the marketplace
- Two models available: Anthropic's Claude 3.5 Sonnet or OpenAI's GPT-4o

THE COPILOT ALTERNATIVES

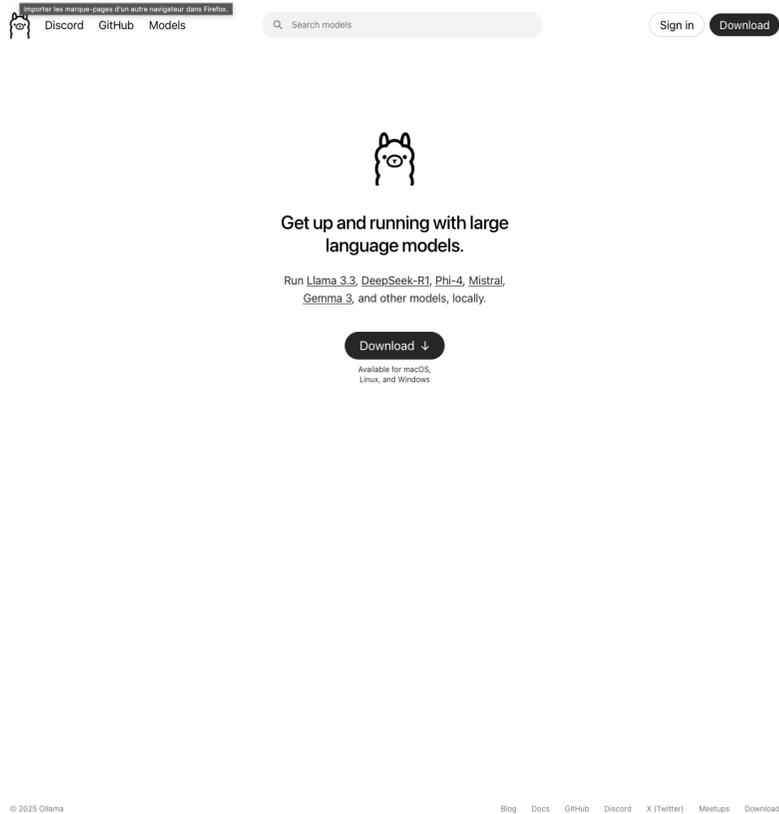


WHAT MODELS DO I NEED TO CHOOSE TO HELP ME DEVELOP MY SOFTWARE ?

DO I HAVE ENOUGH RESOURCES TO USE THEM ?

WHAT ARE THE LIMITATIONS OF THESE MODELS ?

WHERE TO FIND LLM MODELS ?



Importez les marques-pages d'un autre navigateur dans Firefox.

Discord GitHub Models

Search models

Sign in Download



Get up and running with large language models.

Run Llama 3.3, DeepSeek-R1, Phi-4, Mistral, Gemma 3, and other models, locally.

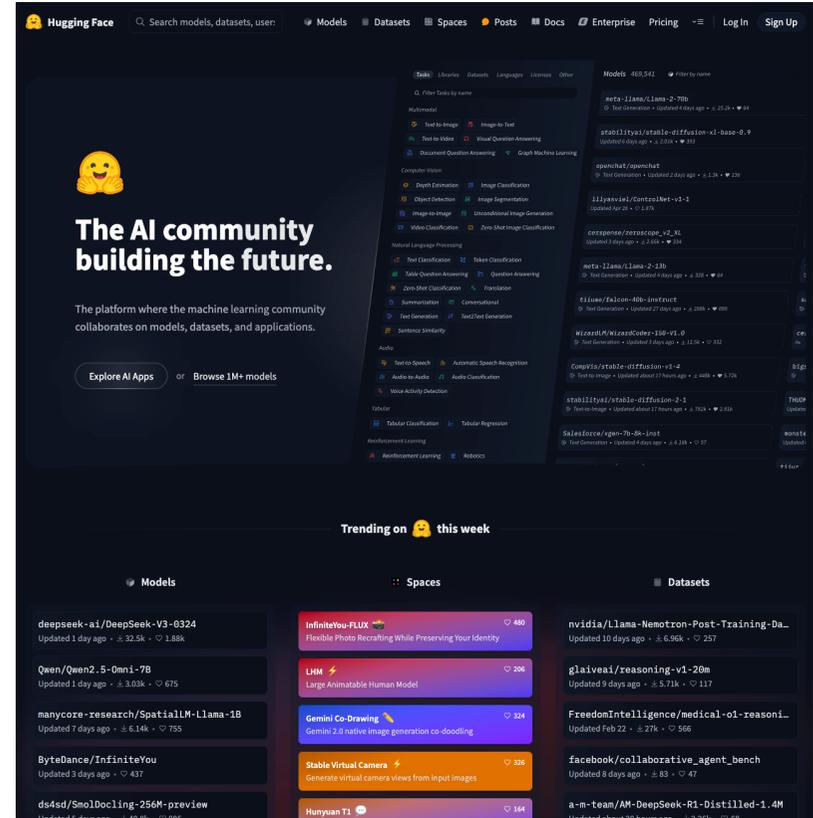
Download ↓

Available for macOS, Linux, and Windows

© 2025 Ollama

Blog Docs GitHub Discord X (Twitter) Meetups Download

Ollama website



Hugging Face

Search models, datasets, users

Models Datasets Spaces Posts Docs Enterprise Pricing Log In Sign Up

469,541 Filter by name

meta-llama/llama-2-70b
Test Generation • Updated 4 days ago • 2.2k • 44

stabilityai/stable-diffusion-v1-base-0.9
Updated 6 days ago • 2.01k • 313

OpenShot/OpenShot
Test Generation • Updated 2 days ago • 1.2k • 210

lllyyrrrr/ControlNet-v1-3
Updated 4 days ago • 1.12k • 144

coqui-ai/corpuscope_v2_xl
Updated 4 days ago • 1.01k • 144

meta-llama/llama-2-13b
Test Generation • Updated 2 days ago • 1.01k • 41

facebook/falcon-40b-instruct
Test Generation • Updated 2 days ago • 1.01k • 400

NVIDIA/nvcrxcode-150-v1.0
Test Generation • Updated 7 days ago • 1.01k • 310

CompVis/stable-diffusion-v1-4
Test to Image • Updated about 17 hours ago • 1.44k • 1,770

stabilityai/stable-diffusion-2-1
Test to Image • Updated about 17 hours ago • 1.71k • 1,111

Salesforce/agent-70-85-Inst
Test Generation • Updated 4 days ago • 1.11k • 177

Trending on 🤖 this week

Models Spaces Datasets

deepseek-ai/DeepSeek-V3-0324
Updated 1 day ago • 32.5k • 1,88k

Qwen/Qwen2.5-Omni-7B
Updated 1 day ago • 3,03k • 675

manycore-research/SpatialM-Llama-1B
Updated 7 days ago • 6.14k • 755

ByteDance/InfiniteYou
Updated 3 days ago • 437

ds4sd/Sm1Docling-256M-preview
Updated 5 days ago • 4.49k • 895

InfiniteYou-FLUX
Flexible Photo Recrafting While Preserving Your Identity
480

LHM
Large Animatable Human Model
206

Gemini Co-Drawing
Gemini 3.0 native image generation co-doodling
334

Stable Virtual Camera
Generate virtual camera views from input images
336

Hunyuan T1
164

nvidia/llama-NeMotron-Post-Training-Da-
Updated 10 days ago • 6.96k • 257

glaiwei/reasoning-v1-20m
Updated 9 days ago • 5.71k • 117

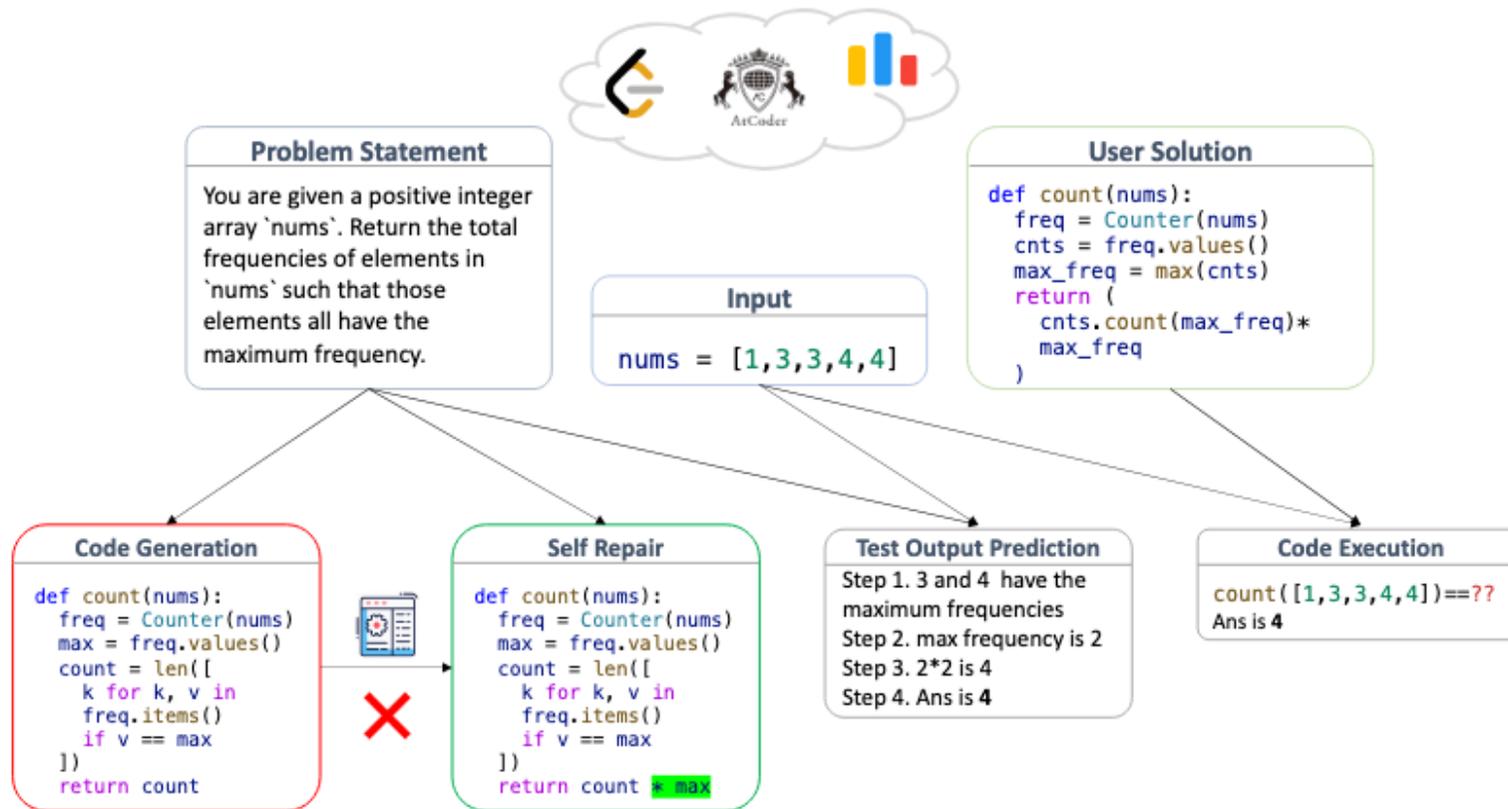
FreedomIntelligence/medical-o1-reasoni-
Updated Feb 22 • 27k • 566

facebook/collaborative_agent_bench
Updated 8 days ago • 83 • 47

a-m-team/AM-DeepSeek-R1-Distilled-1.4M
Updated about 20 hours ago • 2.26k • 88

Hugging Face website

LEADERBOARDS



LiveCodeBench: <https://arxiv.org/pdf/2403.07974>

LEADERBOARDS

LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code

Paper Code Data Home

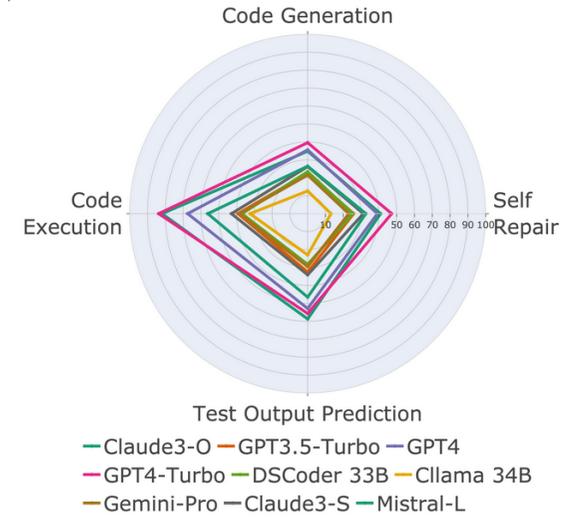
Code Generation Self Repair Test Output Prediction Code Execution

349 problems selected in the current time window. You can change start or end date to change the time window.

We estimate cutoff dates based on release date and performance variation. Feel free to adjust the slider to see the leaderboard at different time windows. Please offer feedback if you find any issues!



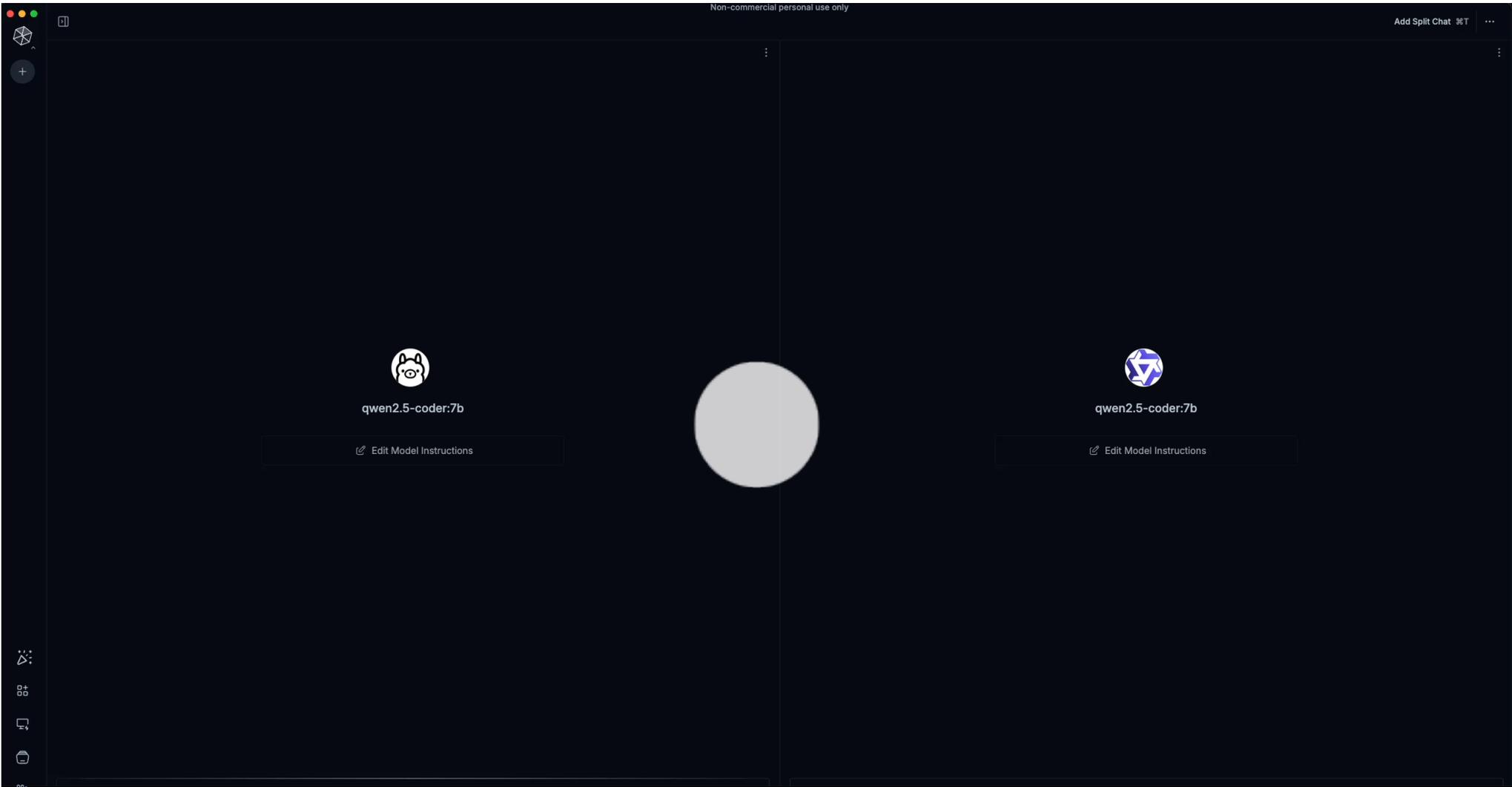
Rank	Model	Pass@1 ↓	Easy-Pass@1	Medium-Pas
	GPT-40-2024-05-13	45.6	88.3	33.2
1	GPT-4-Turbo-2024-04-09	44.7	85.3	33
2	GPT-4-Turbo-1106	39.7	84.4	24
3	GPT-4-0613	36.9	78.4	21.2
4	Gemini-Pro-1.5-May	35.7	76	19.4
5	Claude-3-Opus	35.4	78.8	16.3
	Codestral-Latest	32.2	69	18.7
6	Gemini-Flash-1.5-May	30	68.1	12.6
7	LLama3-70b-Ins	28.3	60.7	15.8
8	Claude-3-Sonnet	26.9	67.6	6.3
9	Gemini-Pro-1.5-April (n=1)	26.9	56.5	14.3
10	Mixtral-8x22B-Ins	26.4	59.8	12.7
11	Mistral-Large	26	60.2	10.9
12	GPT-3.5-Turbo-0125	24.4	54.0	10.0



Other leaderboards

- Code evaluation
 - BigCode's Models Leaderboard
 - BigCode's BigCodeBench
 - Meta's CyberSecEval
- Mathematics abilities
 - NPHardEval

CAN I USE LLM MODELS LOCALLY ?



0:00 / 0:30

1 GPU A100 - Juliet

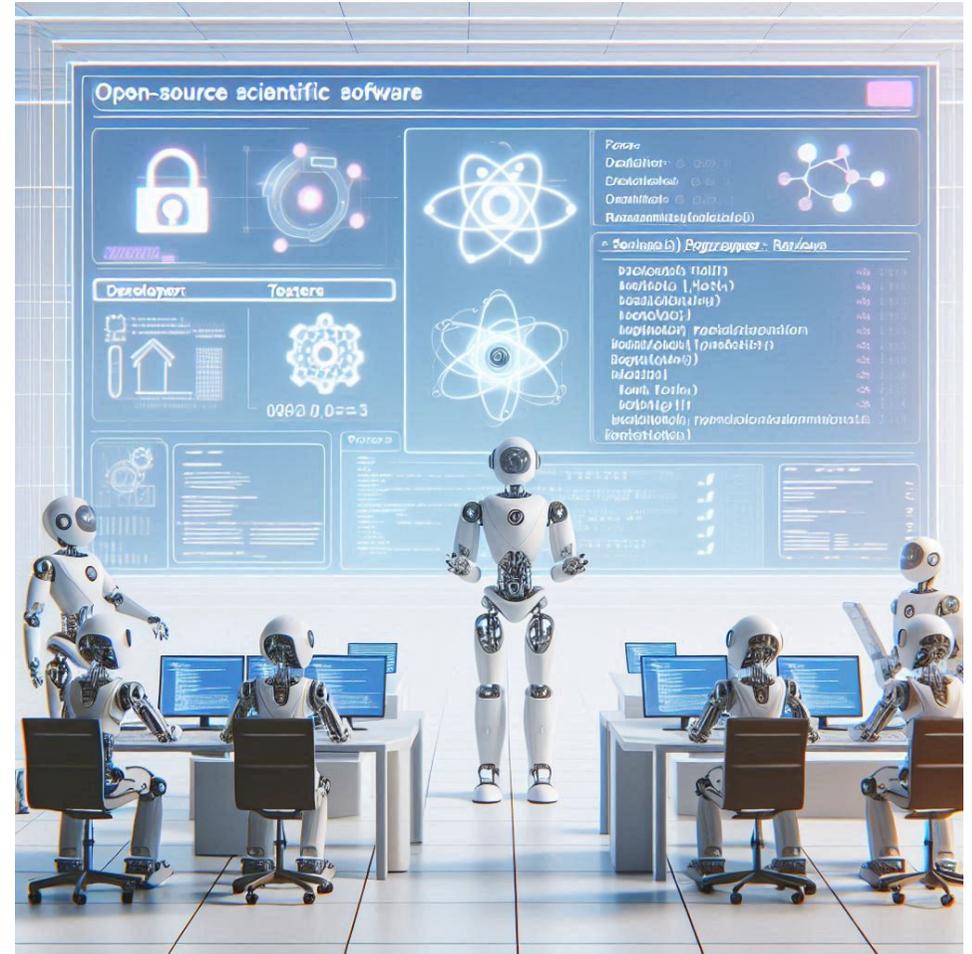
MacBook Pro M1



- **Offline-first, online-ready**
Works seamlessly offline while supporting online models.
- **Parallel multiverse chats**
Compare responses from different AI models in real-time.
- **Unified access to models**
Supports models from Hugging Face, Ollama, and Open Router.
- **Prompt management**
Offers a library of prompts and allows custom additions.
- **Ultimate privacy**
No personal data leaves the user's machine

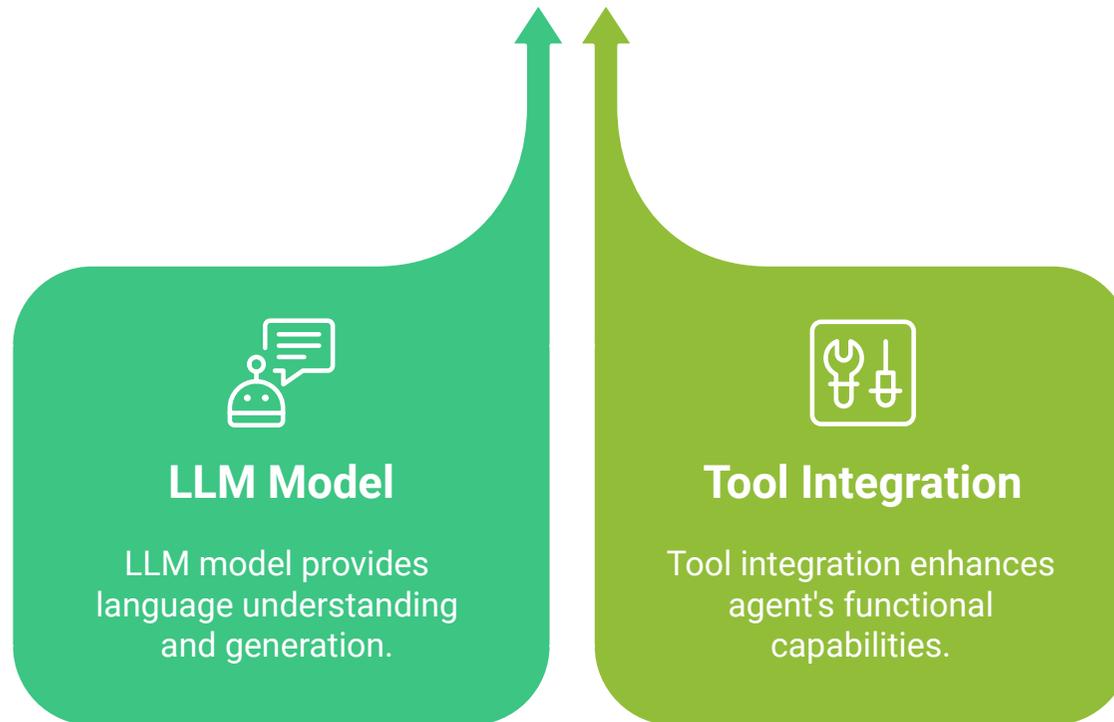
BUT FINALLY WHAT DO I NEED ?

- To have an AI assistant that knows
 - the programming language I use
 - the documentation of my third-party libraries
 - my software and the mathematics behind it
- Specialized models for each step of development
- Not to iterate over LLM models again and again
- To have a memory of what was done

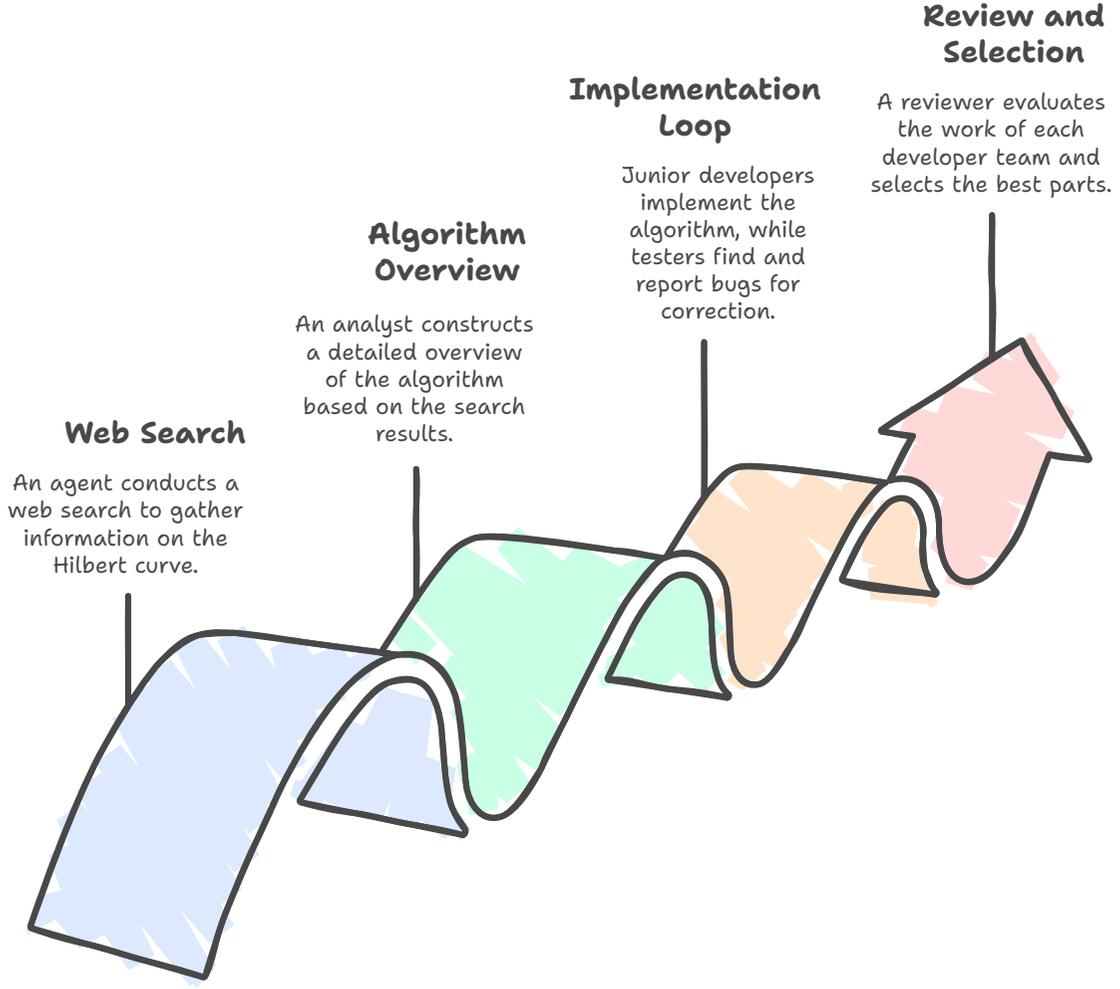


WHAT WE CALL AN AGENT ?

Simple Agent



Multi-agents for software development



USE CASE: DEPIXELIZING

SIGGRAPH 2011



[Johannes Kopf](#)

Microsoft Research

[Dani Lischinski](#)

The Hebrew University



Nearest-neighbor result (Original: 40 x 16 pixels)



Our result

Naïve upsampling of pixel art images leads to unsatisfactory results. Our algorithm extracts a smooth, resolution-independent vector representation from the image which is suitable for high-resolution display devices (Image © Nintendo Co., Ltd.).

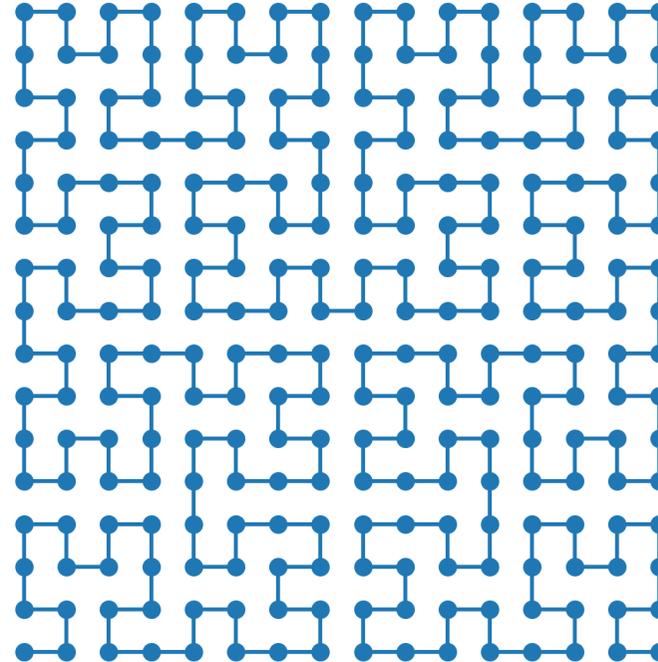
USE CASE: HILBERT CURVE

```
def rot(n, x, y, rx, ry):
    if ry == 0:
        if rx == 1:
            x = n - 1 - x
            y = n - 1 - y
        return y, x
    return x, y

def d2xy(n: int, d: int):
    t = d
    x = y = 0
    s = 1

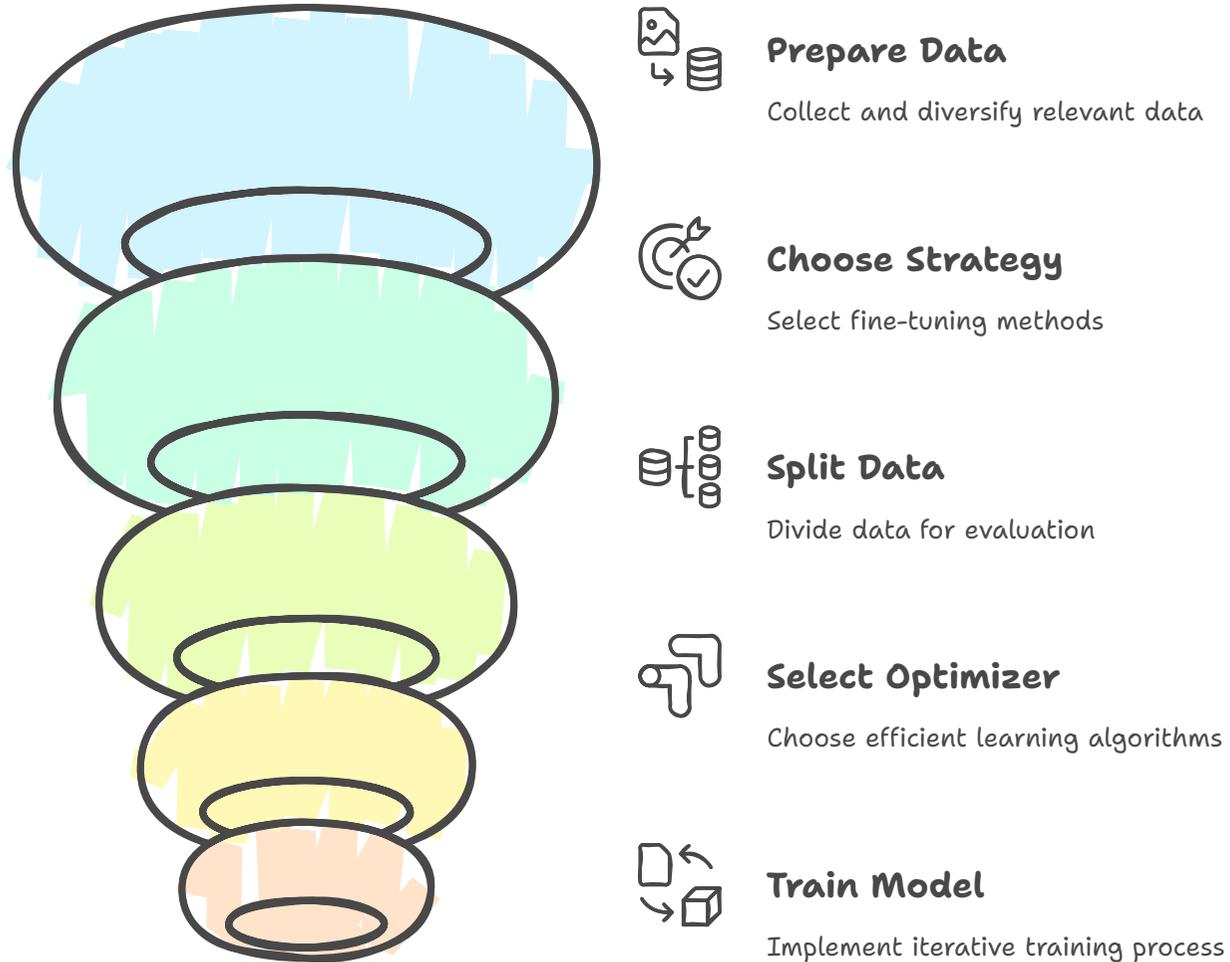
    while (s < n):
        rx = 1 & (t//2)
        ry = 1 & (t ^ rx)
        x, y = rot(s, x, y, rx, ry)
        x += s * rx
        y += s * ry
        t = t//4
        s *= 2
    return x, y

if __name__ == "__main__":
    x = y = 0
    n = 8
    coords = []
    for i in range(1<<n):
        coords.append(d2xy(1<<n, i))
```

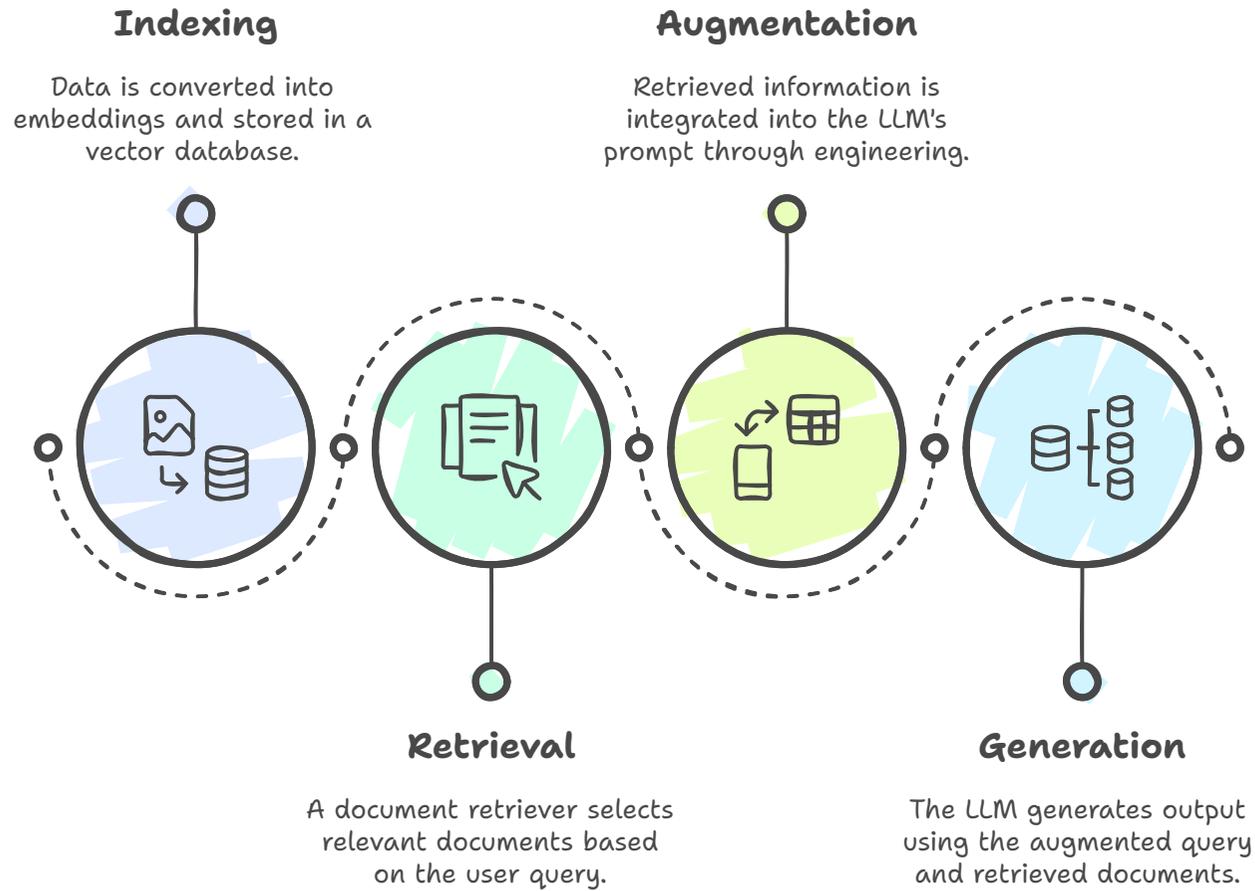


HOW CAN I ENRICH A GENERIC LLM MODEL ?

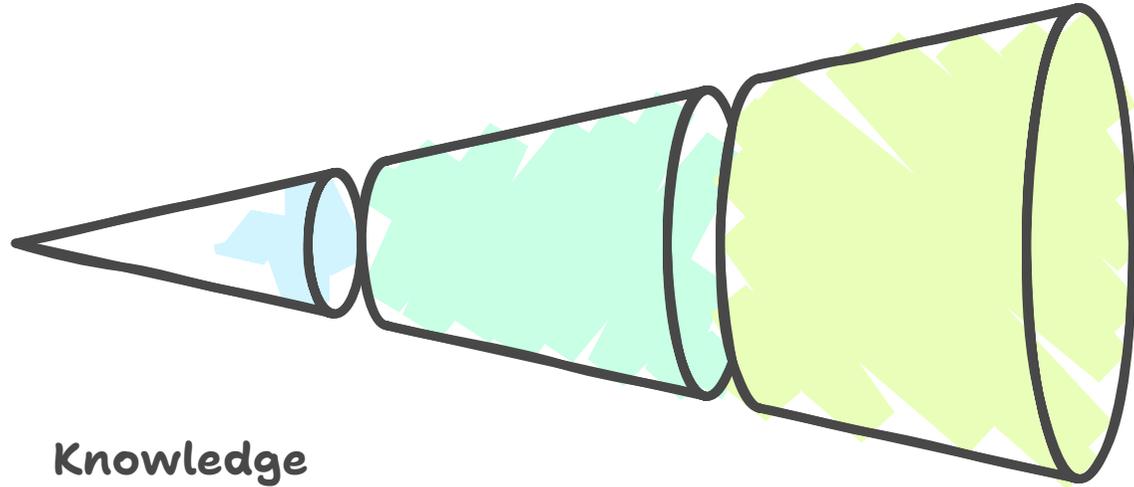
Fine-Tuning a Large Language Model



Retrieval-Augmented Generation Process



Cache augmented generation



Knowledge Preloading

Documents are encoded into a KV cache

Inference Execution

User queries are processed using the KV cache

Cache Resetting

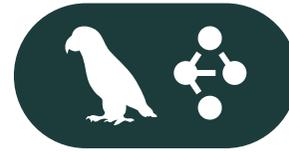
The cache is reset for new sessions



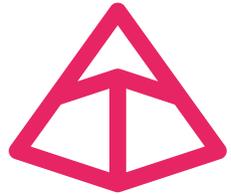
- Web search (Tavily, DuckDuckGo, Brave, ...)
- Python script execution
- GitHub interactions
- ...



LangChain



LangGraph



Pydantic



PydanticAI

THE MULTI AGENT TOOLS

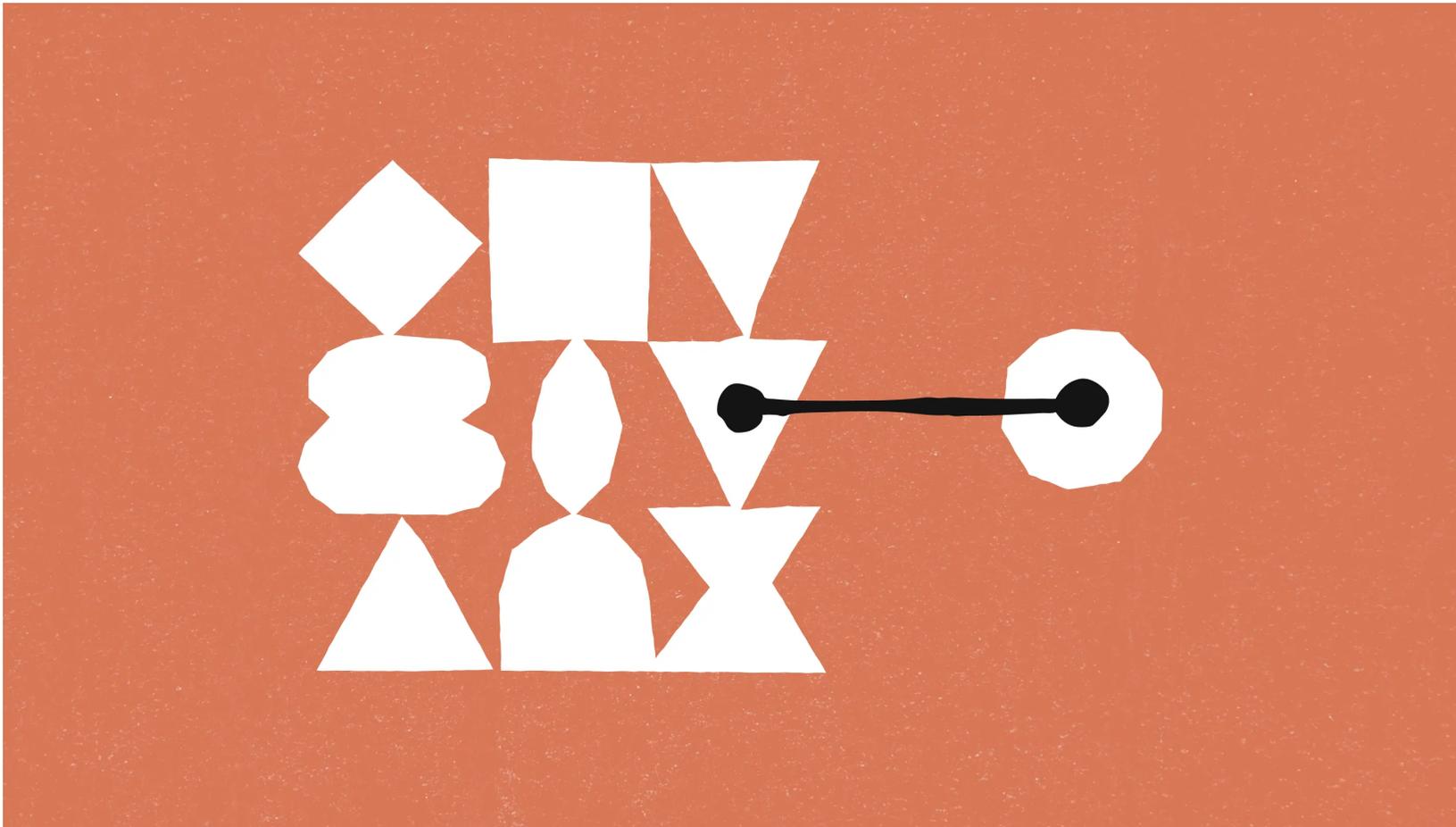
- **CrewAI** is a lean, lightning-fast Python framework built entirely from scratch—completely independent of LangChain or other agent frameworks.
- **Langflow** is a new, visual framework for building multi-agent and RAG applications. It is open-source, Python-powered, fully customizable, and LLM and vector store agnostic.
- **AutoAgent** is a Fully-Automated and highly Self-Developing framework that enables users to create and deploy LLM agents through Natural Language Alone.
- **SmolAgents** is the simplest framework out there to build powerful agents!
- ...

SOME DEMOS

SO, WHAT DID I LEARN?

- The context is probably the most important.
- XML can help you structure your output messages.
- It's not that simple to get the tools and the LLM to talk to each other.
- There are so many software packages out there that it's hard to choose the right one.
- I won't be getting a team of AI assistants working on my software developments any time soon.

But I'm not done yet !!



Model Context Protocol

REFERENCES

- More about RAG
[Advanced RAG Techniques: Elevating Your Retrieval-Augmented Generation Systems](#)
- More about CAG
[Don't Do RAG: When Cache-Augmented Generation is All You Need for Knowledge Tasks](#)
- If there is only one article that you need to read
[A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges](#)
- Prompt examples on various topics
[fabric is an open-source framework for augmenting humans using AI](#)

THANK YOU FOR YOUR ATTENTION
AND THANKS TO MESONET
ESPECIALLY THE ROMEO TEAM TO LET ME USE THEIR COMPUTING RESOURCES !

<https://github.com/gouarin/2025-04-ia4dev>
<https://github.com/gouarin/llm4code>