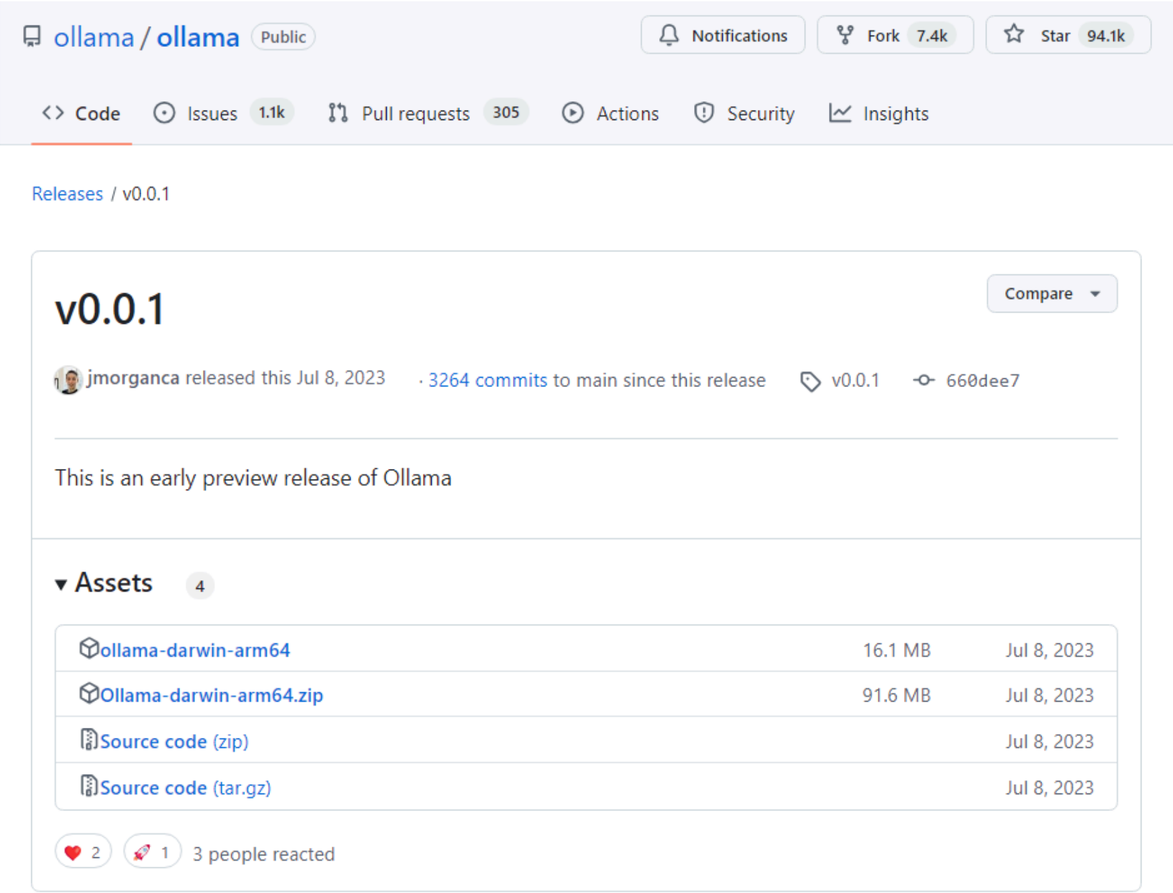
AI TOOL: OLLAMA

There are a lot of interpretations of what Ollama is, like: It's a library, a framework, an open-source lightweight platform, it's quite docker-like, an LLM server, an LLM puller app, etc.

But to simplify, Ollama is: **"The runner and container of Open Large Language Models, locally and without the need of internet through the CLI".**



Stands for Omni-Layer Learning Language Acquisition Model.

BACKGROUND

Ollama was released for the first in 2023 by **Michael Chiang** and **Jeffrey Morgan** in Palo Alto, CA.

It is **based on llama.cpp** (an implementation of the Llama architecture in plan C/C++).

They published the first revision (**v0.0.1**) on **Github**, and began by **supporting** the 1st LLM: **Llama2**.

PRICING?

PRICING?

REQUIREMENTS, COMPATIBILITY INTEGRATIONS

SUPPORTED OPERATING SYSTEMS

HARDWARE REQUIREMENTS

- **RAM**: At least 8GB available
- **GPU:** Nvidia GPUs with compute capability 5.0+ and some AMD Radeon cards.

INTEGRATIONS & LIBRARIES

INTERFACES

Open-webUI

Chatbot-ollama

CHARACTERISTICS

- PRIVACY
- MULTI-MODAL INPUTS
- PASSING AN ARGUMENT WITHIN A PROMPT
- SERVING AS A REST API
- LOCAL EXECUTION & OFFLINE ACCESS
- EXTENSIVE PRE-TRAINED MODEL LIBRARY
- SEAMLESS INTEGRATION
- CUSTOMIZATION (CREATE OUR OWN MODELS AND APPLY USER FRIENDLY-INTERFACES)

FUNCTIONALITY

DIGGIN' DIP INTO IT

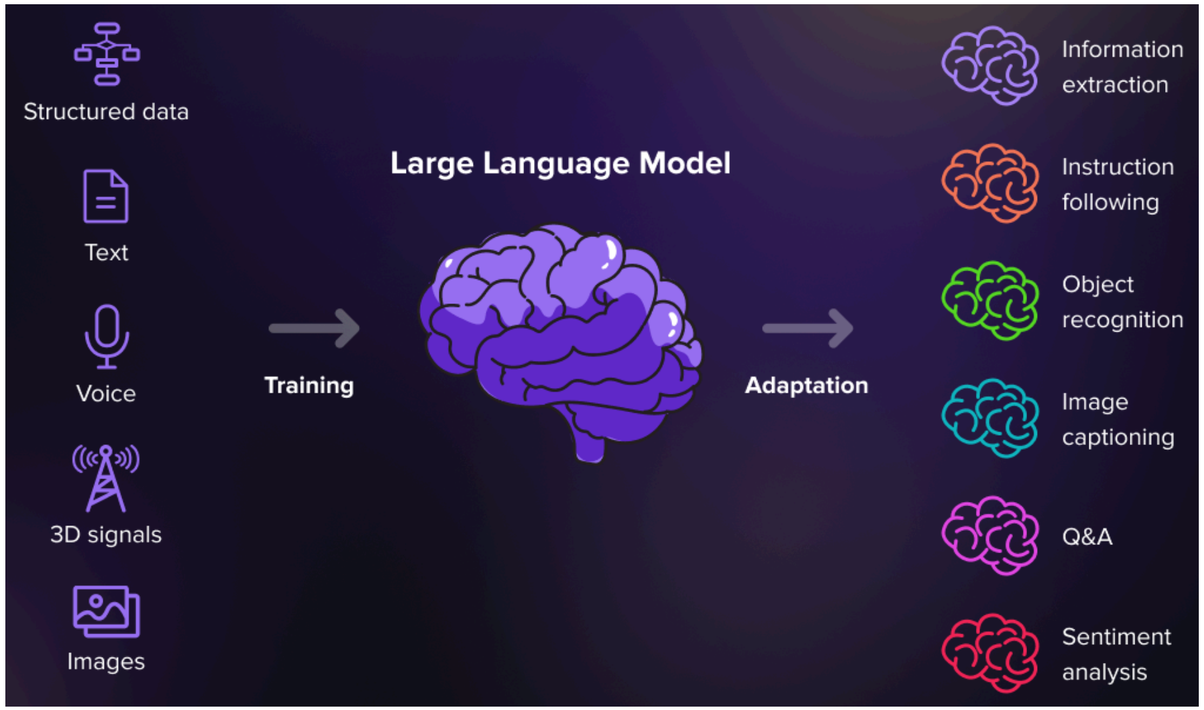
How is it possible that Ollama can run LLMs without internet?, How does it work?, Is my data compromised?

LARGE LANGUAGE MODELS

The main component of Ollama.

Are **pre-trained** AI models with big amounts of human-like text to learn patterns and regularities, and to understand and generate human language. This models are **based on** GPT, BERT and ELM transformer architectures.

It's **what's behind the scenes** of all AI chatbots and AI writing generators.



FUNCTIONALITY

OLLAMA MODELS CATEGORIES

TEXT & CHAT MODELS

Designed for conversational interactions and text generation.

CODE COMPLETION

Trained on vast amounts of code, generating, completing, and understanding code.

MULTIMODAL OR VISION MODELS

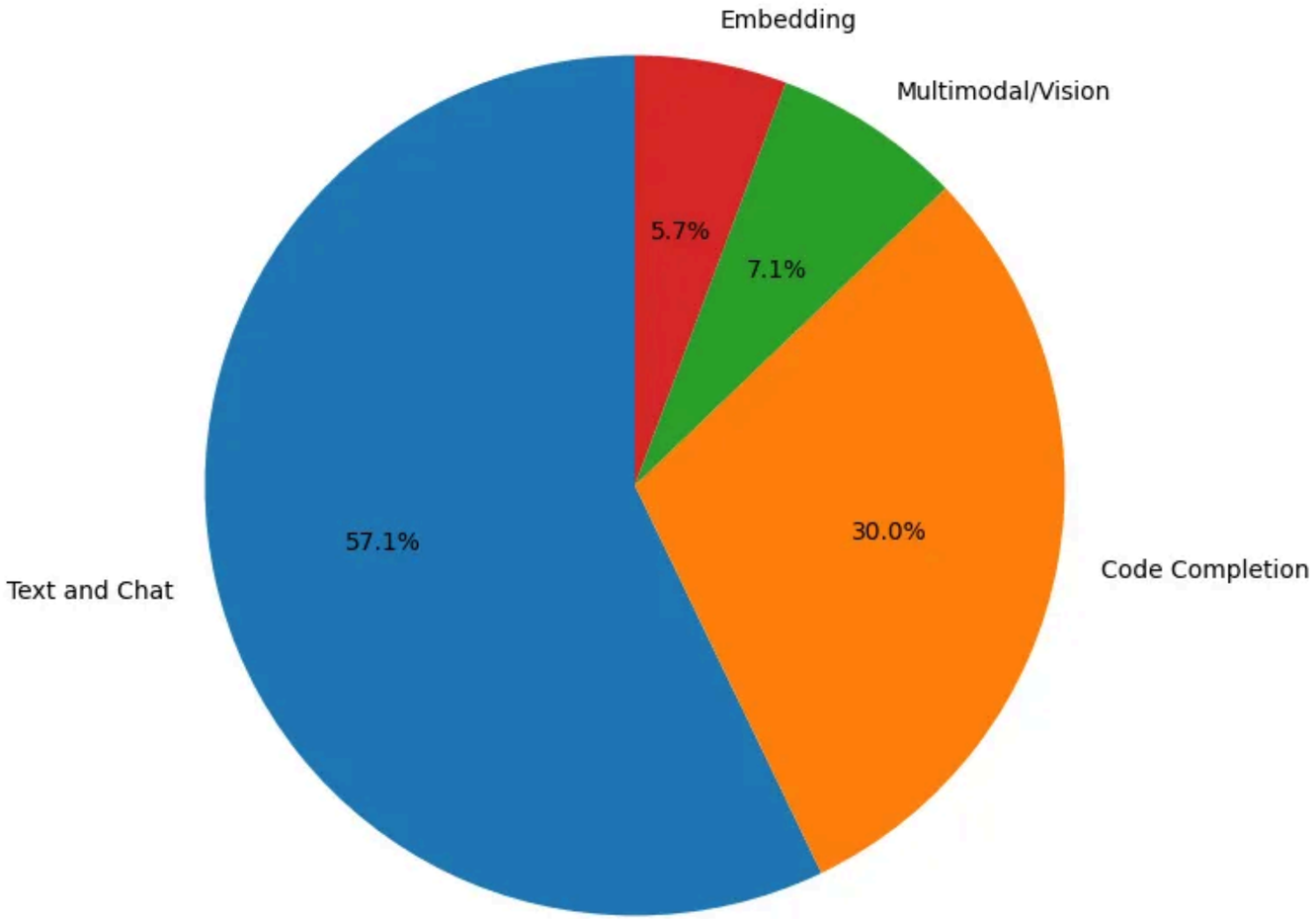
Integrate image understanding, analyzing both textual and visual information.

EMBEDDING MODELS

Convert text into numerical representations, facilitating tasks like similarity search and information retrieval.

FUNCTIONALITY

DISTRIBUTION OF OPEN-SOURCE LLM CATEGORIES



TOP LLM MODELS

PHI3

Parameters: 3B
Size: 2.2GB
Context length: 4K tokens

LLAMA3.2

Parameters: 1B - 3B
Size: 4.7GB
Context length: 131K tokens

QWEN2.5

Parameters: 0.5B - 72B
Size: 4.7GB
Context length: 128K tokens

MISTRAL

Parameters: 7B
Size: 4.1 GB
Context length: 10K tokens

MORE MODELS IN HUGGINGFACE!

PROS

LATENCY

Cloud-based models often suffer from network latency. With Ollama, the model **runs on your local machine**, eliminating this issue.

DATA TRANSFER

With cloud-based solutions, you have to send your data over the internet. **Ollama keeps it local**, offering a more **secure environment** for your sensitive data.

MODEL INFERENCE

Ollama can reduce your model **inference time** by up to 50%, depending on your hardware configuration.

CUSTOMIZATION

Ollama gives you the freedom to tweak the **models as per your needs**, something that's often restricted in cloud-based platforms.

CAUTIONS

OFFLINE ACCESS (A DOUBLE-EDGED SWORD)

May not have access to real-time data or the ability to learn from new experiences when disconnected from the internet. In certain situations, they **might not provide the most accurate or up-to-date information.**

TEXT-BASED MODEL

Can’t access to our files stored in our computers because most of them are text-base models. Don’t have that ability to process any file like images or PDFs.

CONTEXT-LENGTH AND PARAMETERS LIMITATIONS

SESSION DURATION RESTRICTION

```
**Mistral-4B**: The maximum session duration is around 8-12 hours.
**phi3-XL**: The maximum session duration is around 4-6 hours.
**qwen2.5**: Unfortunately, I couldn't find specific information on
```