Training your own AI model

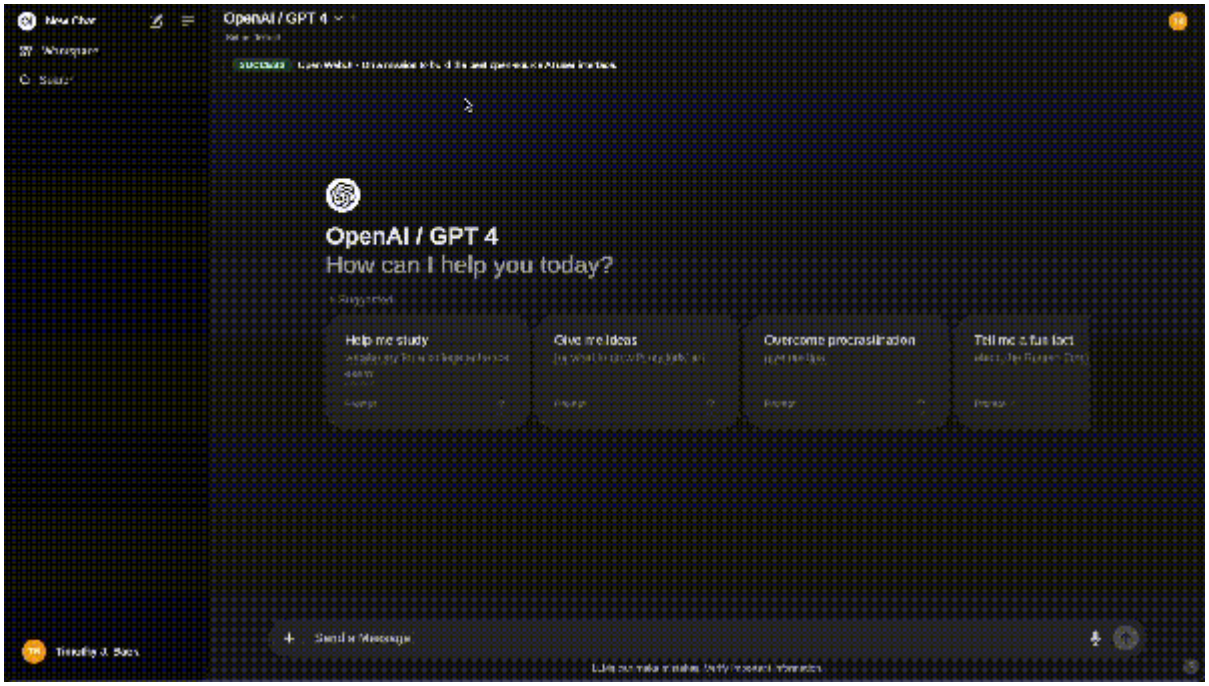
source: https://commons.wikimedia.org/wiki/File:Artificial-Intelligence.jpg

Wikipedia

> A large language model (LLM) is a type of machine learning model designed for natural language processing tasks such as language generation. LLMs are language models with many parameters, and are trained with self-supervised learning on a vast amount of text.

For some time now, I have been thinking about training an open source LLM, e.g. Ollama, and training it with all the material on my disk or in this wiki, and then publishing it on my website as a chat bot that you can ask anything about.

Ollama + Openwebui



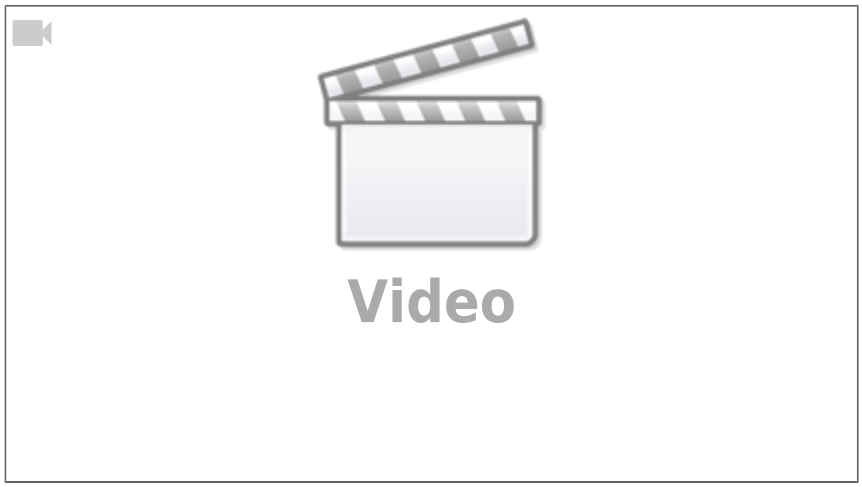
Source: https://docs.openwebui.com/assets/images/demo-d3952c8561c4808c1d447fc061c71174.gif

I tested this toolkit and unfortunately but all the necessary files with which we would like to teach AI have to be sent via the web panel to the model we are training which is terribly laborious. Then the model has to read it all which makes it take even longer. This is not the best solution for such an application as I mentioned in the introduction. There is also no possibility of connecting a file directory so that the AI can index everything and then answer questions according to this knowledge.

The advantage of this solution is that it is a web-based programme that can be opened anywhere and has the possibility of connecting different AI service providers in one place, which allows the results of different AI models to be compared.

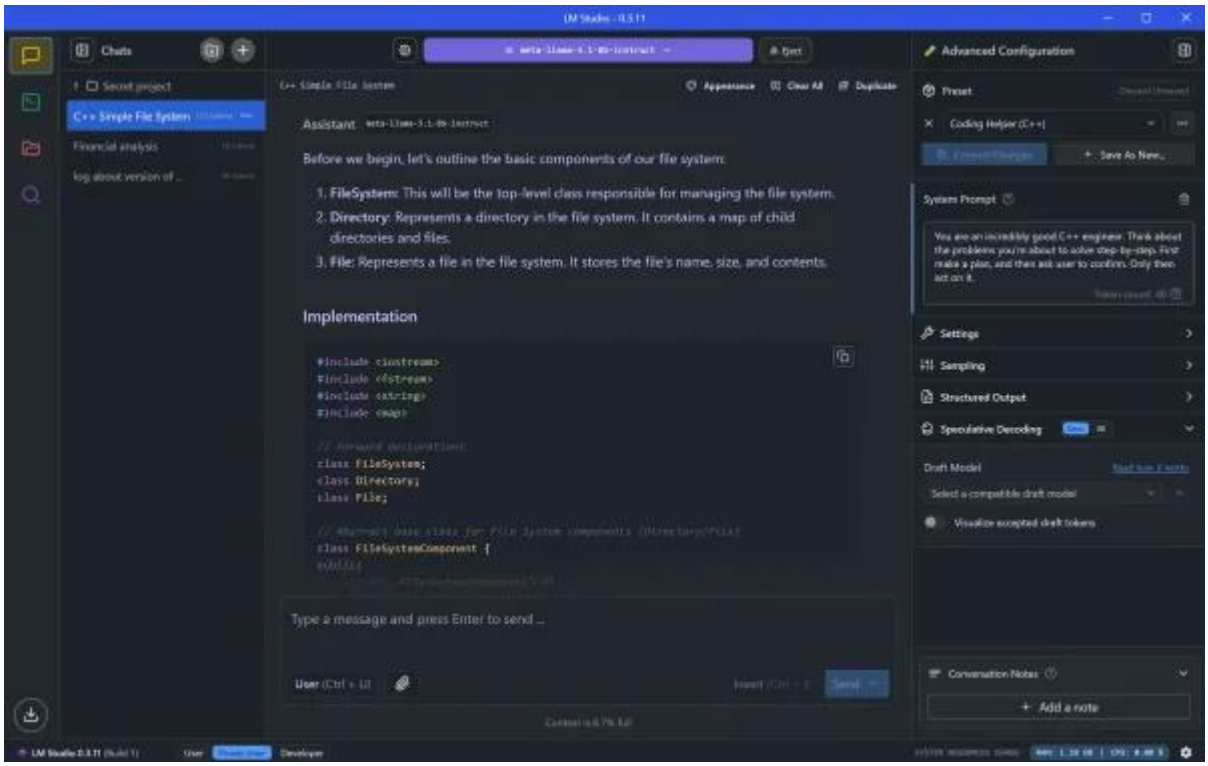
Also, it has cool settings for model permissions and prompts.

I occasionally use this tool myself because it is more cost-effective to pay for individual requests to the OpenAI API than to pay for the entire ChatGPT PLUS or PRO subscription.



There is a nice video discussing OpenWebUi on YT

LMstudio

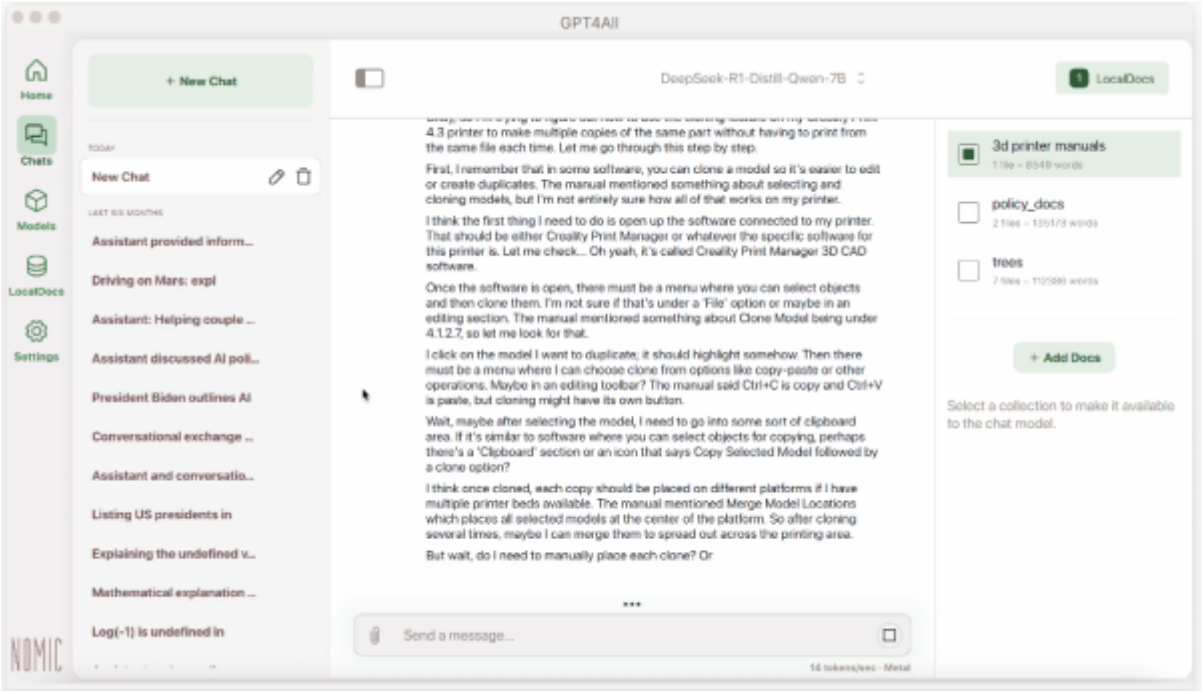


źródło:https://lmstudio.ai/_next/image?url=%2F_next%2Fstatic%2Fmedia%2Fhero-windows.2a9fa20d.webp&w=3840&q=75

This is the tool I installed on my main workstation. It's a nice program if you want to do some AI on your computer on some confidential topics. Here again is the same problem as the one with OpenWebUi, the tool behaves like a Web application converted into a desktop application using the Elektron package. This has its drawbacks, uploading files works the same way as in OpenWebUi, it causes the same problem, i.e. we can't upload 100 files because the whole application crashes, there is also no possibility to plug in a directory so that AI can index it.

This is a nice application if you want to easily experiment with LLM models on your own computer.

GPT4ALL



source: https://www.nomic.ai/gpt4all

Right away disclaimer I have not yet tested this tool. However, from the page we can read that:

> Chat with Your Files Privately: Introducing LocalDocs Grant your local LLM access to your private, sensitive documents with LocalDocs. Your documents stay secure and private. Your local LLM can access your documents without an internet connection.

So this might be exactly what I have in mind.

In the documentation on the website we can read:

> How It Works. A LocalDocs collection uses Nomic AI's free and fast on-device embedding models to index your folder into text snippets that each get an embedding vector. These vectors allow us to find snippets from your files that are semantically similar to the questions and prompts you enter in your chats. We then include those semantically similar snippets in the prompt to the LLM. To try the embedding models yourself, we recommend using the Nomic Python SDK

It promises to be good, from the documentation it appears that there is a special tool from Nomic AI which indexes the folders the files are converted into snippets and each of them has a nesting vector, then these vectors allow the AI to find snippets from files which are semantically similar to the questions and prompts we give to the model. And then these semantically similar snipets are added to the answers.

From what we have been able to read this is the closest thing to what I wanted to achieve.

In one of the next posts we'll be doing some testing of this solution and looking at how it handles more data and whether it's suitable for anything at all.

— Kacper Ostrowski 2025/05/09 11:08