5 WAYS TO RUN LLMS LOCALLY WITHOUT INTERNET

Bhavishya Pandit

1.GPT4ALL

GPT4ALL supports **1000+** open source LLM models.

The GPT4All Desktop Application allows you to download and run large language models (LLMs) locally & privately on your device.

With GPT4All, you can chat with models, turn your local files into information sources for models (LocalDocs), or browse models available online to download onto your device.

Follow this 3 step process to get started:

Download the desktop client

Download the model e.g. Llama 3

Start Chatting

2. OLLAMA

103K Github stars ⭐

Ollama is another easy to use tool using which you can run LLM models from the terminal itself.

Just download Ollama from **https://ollama.com/download** and you are ready to go.

command to run Llama 3 using ollama:

```
ollama run llama3
```

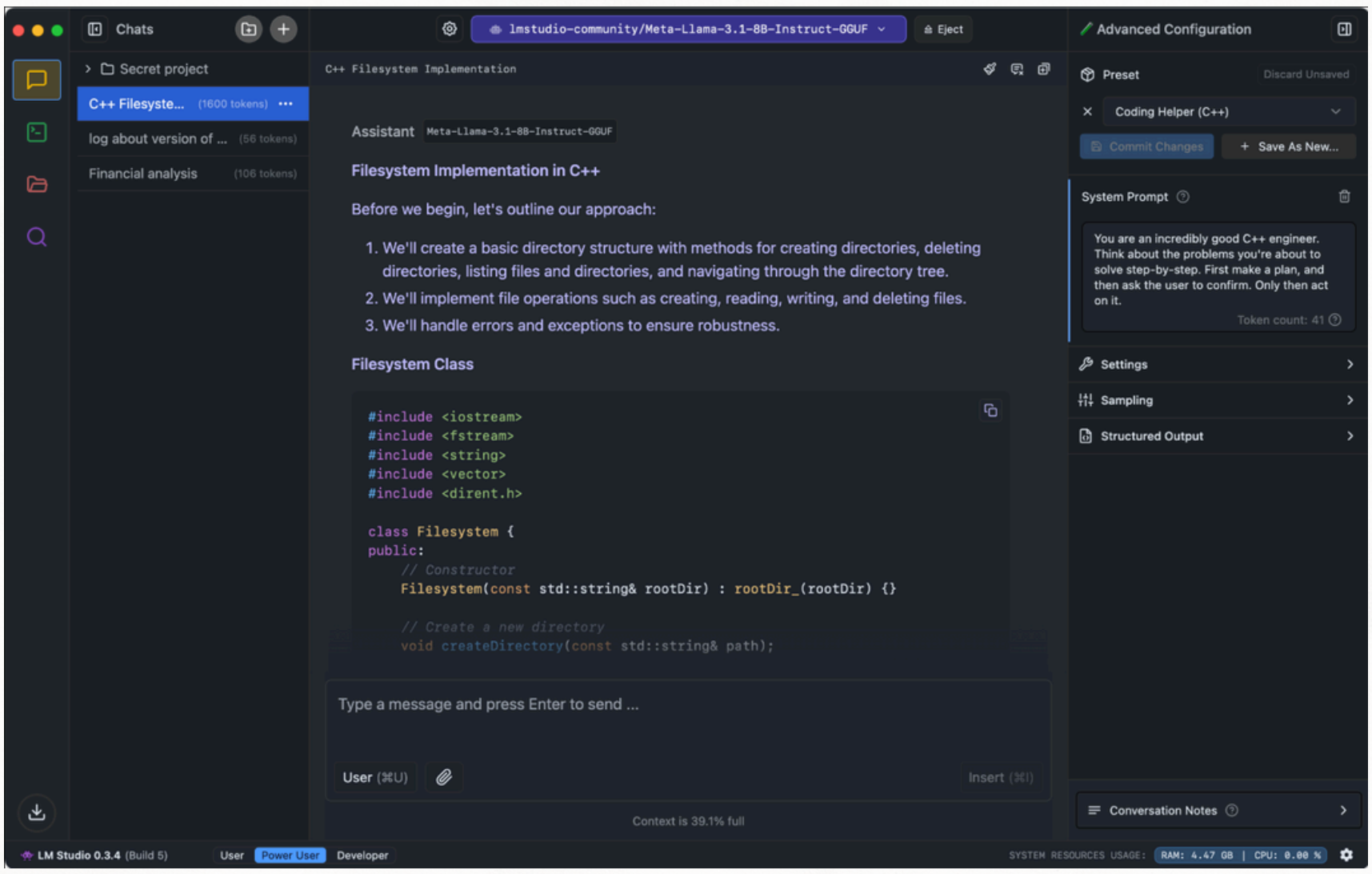
commands to run different models below:

Model	Parameters	Size	Download
Llama 3	8B	4.7GB	`ollama run llama3`
Llama 3	70B	40GB	`ollama run llama3:70b`
Mistral	7B	4.1GB	`ollama run mistral`
Dolphin Phi	2.7B	1.6GB	`ollama run dolphin-phi`
Phi-2	2.7B	1.7GB	`ollama run phi`
Neural Chat	7B	4.1GB	`ollama run neural-chat`
Starling	7B	4.1GB	`ollama run starling-lm`
Code Llama	7B	3.8GB	`ollama run codellama`
Llama 2 Uncensored	7B	3.8GB	`ollama run llama2-uncensored`
Llama 2 13B	13B	7.3GB	`ollama run llama2:13b`
Llama 2 70B	70B	39GB	`ollama run llama2:70b`
Orca Mini	3B	1.9GB	`ollama run orca-mini`

3. LM Studio

LM Studio is similar to GPT4All however unlike GPT4ALL it doesn't allow connecting a local folder. Setting up LM Studio is pretty easy:

1. **Installation:** Download installer from https://lmstudio.ai/
2. **Download model:** You can download any model of your choice from Hugging Face using search function.
3. **Generate Response:** We can select our downloaded model from the dropdown menu and chat with it .



A standout feature of LM Studio is its ability to run and **host multiple models** simultaneously. This enables users to compare outputs across models and apply them to various use cases.

However, running multiple sessions requires a GPU with high VRAM capacity.

4. llama.cpp

~70K Github stars ⭐

llama.cpp is purely written in C/C++ making it fast and efficient.

Step instructions to get started with llama.cpp:

1. Downloading llama.cpp:

```
git clone --depth 1 https://github.com/ggerganov/llama.cpp.git
```

2. Build the project:

```
cmake -B build
cmake --build build --config Release
```

2. Start llama.cpp's WebUI Server:

```
llama-server -m model.gguf --port 8080
```

5. Jan

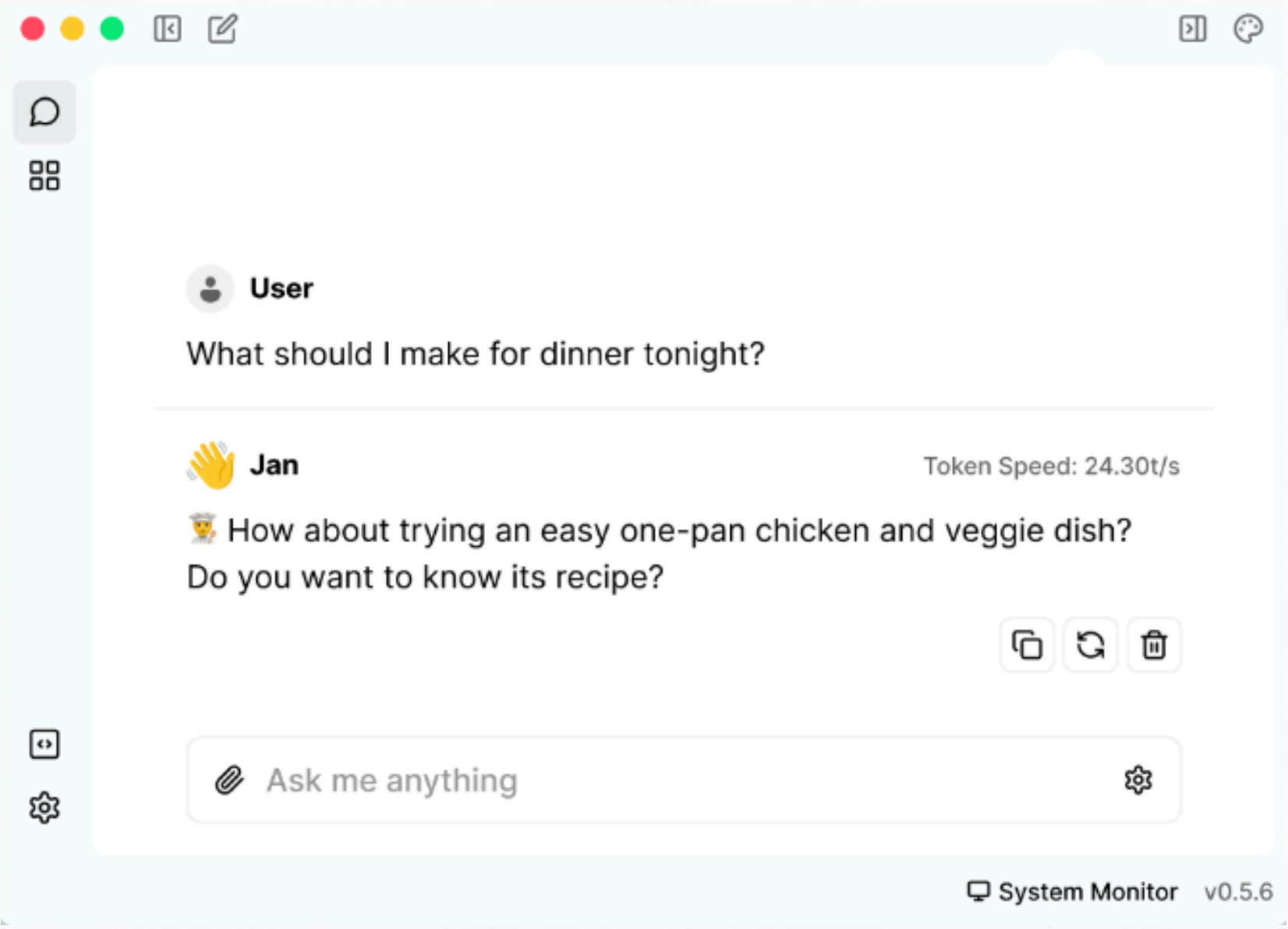
2.1 million downloads

“Jan is an open source ChatGPT-alternative that runs 100% offline.”

Setting up Jan is pretty easy:

1. Install Jan from **https://jan.ai/**
2. Download a model by selecting from the drop down menu or pasting Hugging Face model ID.
3. Start chatting with the model.

Jan also offers feature to connect to a Remote API i.e. access to models hosted on external servers.



Follow to stay updated on Generative AI

LIKE

COMMENT

REPOST