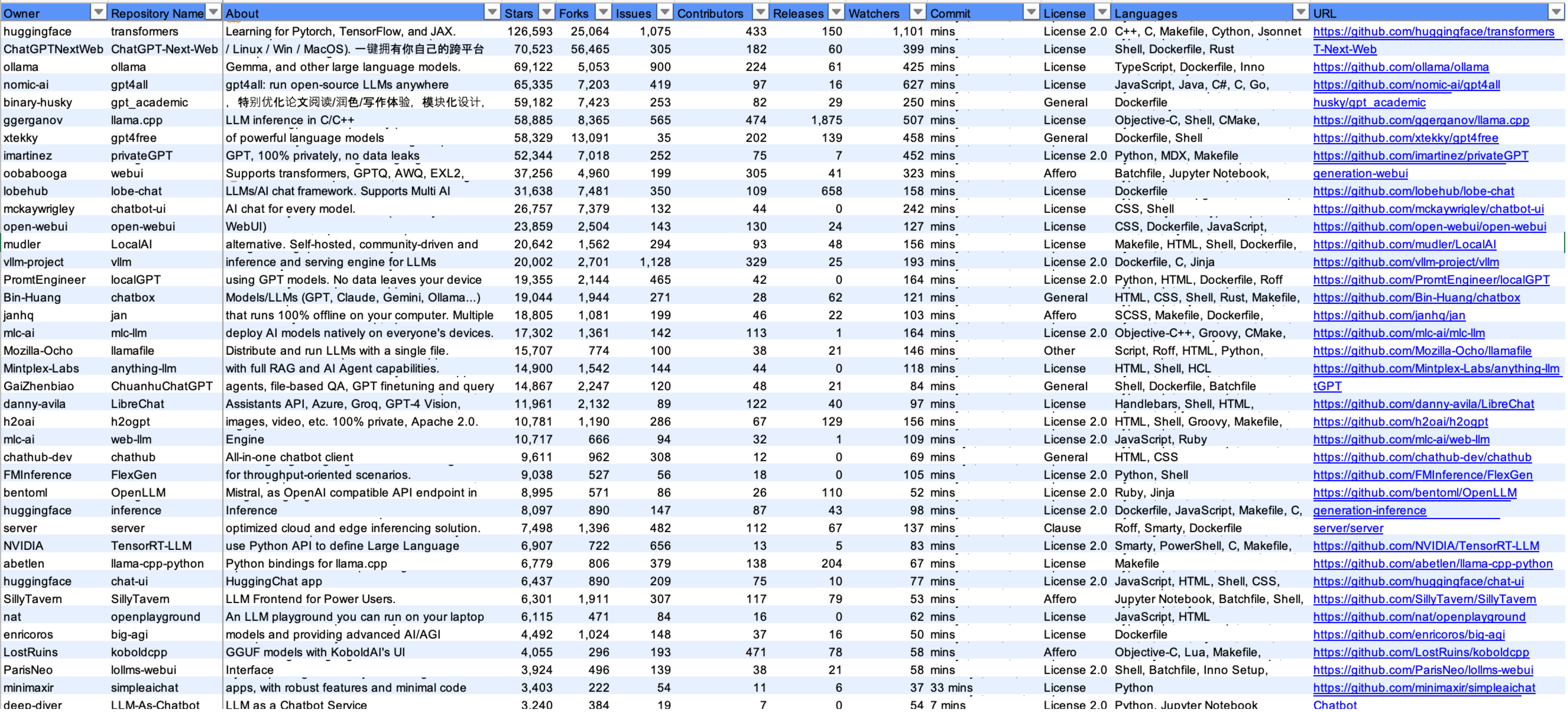
# Running LLMs Locally



| Owner | Repository Name | About | Stars | Forks | Issues | Contributors | Releases | Watchers | Commit | License | Languages | URL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| huggingface | transformers | Learning for Pytorch, TensorFlow, and JAX. | 126,593 | 25,064 | 1,075 | 433 | 150 | 1,101 | mins | License 2.0 | C++, C, Makefile, Cython, Jsonnet | https://github.com/huggingface/transformers |
| ChatGPTNextWeb | ChatGPT-Next-Web | / Linux / Win / MacOS). 一键拥有你自己的跨平台 | 70,523 | 56,465 | 305 | 182 | 60 | 399 | mins | License | Shell, Dockerfile, Rust | T-Next-Web |
| ollama | ollama | Gemma, and other large language models. | 69,122 | 5,053 | 900 | 224 | 61 | 425 | mins | License | TypeScript, Dockerfile, Inno | https://github.com/ollama/ollama |
| nomic-ai | gpt4all | gpt4all: run open-source LLMs anywhere | 65,335 | 7,203 | 419 | 97 | 16 | 627 | mins | License | JavaScript, Java, C#, C, Go, | https://github.com/nomic-ai/gpt4all |
| binary-husky | gpt_academic | ，特别优化论文阅读/润色/写作体验，模块化设计， | 59,182 | 7,423 | 253 | 82 | 29 | 250 | mins | General | Dockerfile | husky/gpt_academic |
| ggerganov | llama.cpp | LLM inference in C/C++ | 58,885 | 8,365 | 565 | 474 | 1,875 | 507 | mins | License | Objective-C, Shell, CMake, | https://github.com/ggerganov/llama.cpp |
| xtekky | gpt4free | of powerful language models | 58,329 | 13,091 | 35 | 202 | 139 | 458 | mins | General | Dockerfile, Shell | https://github.com/xtekky/gpt4free |
| imartinez | privateGPT | GPT, 100% privately, no data leaks | 52,344 | 7,018 | 252 | 75 | 7 | 452 | mins | License 2.0 | Python, MDX, Makefile | https://github.com/imartinez/privateGPT |
| oobabooga | webui | Supports transformers, GPTQ, AWQ, EXL2, | 37,256 | 4,960 | 199 | 305 | 41 | 323 | mins | Affero | Batchfile, Jupyter Notebook, | generation-webui |
| lobehub | lobe-chat | LLMs/AI chat framework. Supports Multi AI | 31,638 | 7,481 | 350 | 109 | 658 | 158 | mins | License | Dockerfile | https://github.com/lobehub/lobe-chat |
| mckaywrigley | chatbot-ui | AI chat for every model. | 26,757 | 7,379 | 132 | 44 | 0 | 242 | mins | License | CSS, Shell | https://github.com/mckaywrigley/chatbot-ui |
| open-webui | open-webui | WebUI) | 23,859 | 2,504 | 143 | 130 | 24 | 127 | mins | License | CSS, Dockerfile, JavaScript, | https://github.com/open-webui/open-webui |
| mudler | LocalAI | alternative. Self-hosted, community-driven and | 20,642 | 1,562 | 294 | 93 | 48 | 156 | mins | License | Makefile, HTML, Shell, Dockerfile, | https://github.com/mudler/LocalAI |
| vllm-project | vllm | inference and serving engine for LLMs | 20,002 | 2,701 | 1,128 | 329 | 25 | 193 | mins | License 2.0 | Dockerfile, C, Jinja | https://github.com/vllm-project/vllm |
| PromtEngineer | localGPT | using GPT models. No data leaves your device | 19,355 | 2,144 | 465 | 42 | 0 | 164 | mins | License 2.0 | Python, HTML, Dockerfile, Roff | https://github.com/PromtEngineer/localGPT |
| Bin-Huang | chatbox | Models/LLMs (GPT, Claude, Gemini, Ollama...) | 19,044 | 1,944 | 271 | 28 | 62 | 121 | mins | General | HTML, CSS, Shell, Rust, Makefile, | https://github.com/Bin-Huang/chatbox |
| janhq | jan | that runs 100% offline on your computer. Multiple | 18,805 | 1,081 | 199 | 46 | 22 | 103 | mins | Affero | SCSS, Makefile, Dockerfile, | https://github.com/janhq/jan |
| mlc-ai | mlc-llm | deploy AI models natively on everyone's devices. | 17,302 | 1,361 | 142 | 113 | 1 | 164 | mins | License 2.0 | Objective-C++, Groovy, CMake, | https://github.com/mlc-ai/mlc-llm |
| Mozilla-Ocho | llamafile | Distribute and run LLMs with a single file. | 15,707 | 774 | 100 | 38 | 21 | 146 | mins | Other | Script, Roff, HTML, Python, | https://github.com/Mozilla-Ocho/llamafile |
| Mintplex-Labs | anything-llm | with full RAG and AI Agent capabilities. | 14,900 | 1,542 | 144 | 44 | 0 | 118 | mins | License | HTML, Shell, HCL | https://github.com/Mintplex-Labs/anything-llm |
| GaiZhenbiao | ChuanhuChatGPT | agents, file-based QA, GPT finetuning and query | 14,867 | 2,247 | 120 | 48 | 21 | 84 | mins | General | Shell, Dockerfile, Batchfile | tGPT |
| danny-avila | LibreChat | Assistants API, Azure, Groq, GPT-4 Vision, | 11,961 | 2,132 | 89 | 122 | 40 | 97 | mins | License | Handlebars, Shell, HTML, | https://github.com/danny-avila/LibreChat |
| h2oai | h2ogpt | images, video, etc. 100% private, Apache 2.0. | 10,781 | 1,190 | 286 | 67 | 129 | 156 | mins | License 2.0 | HTML, Shell, Groovy, Makefile, | https://github.com/h2oai/h2ogpt |
| mlc-ai | web-llm | Engine | 10,717 | 666 | 94 | 32 | 1 | 109 | mins | License 2.0 | JavaScript, Ruby | https://github.com/mlc-ai/web-llm |
| chathub-dev | chathub | All-in-one chatbot client | 9,611 | 962 | 308 | 12 | 0 | 69 | mins | General | HTML, CSS | https://github.com/chathub-dev/chathub |
| FMInference | FlexGen | for throughput-oriented scenarios. | 9,038 | 527 | 56 | 18 | 0 | 105 | mins | License 2.0 | Python, Shell | https://github.com/FMInference/FlexGen |
| bentoml | OpenLLM | Mistral, as OpenAI compatible API endpoint in | 8,995 | 571 | 86 | 26 | 110 | 52 | mins | License 2.0 | Ruby, Jinja | https://github.com/bentoml/OpenLLM |
| huggingface | inference | Inference | 8,097 | 890 | 147 | 87 | 43 | 98 | mins | License 2.0 | Dockerfile, JavaScript, Makefile, C, | generation-inference |
| server | server | optimized cloud and edge inferencing solution. | 7,498 | 1,396 | 482 | 112 | 67 | 137 | mins | Clause | Roff, Smarty, Dockerfile | server/server |
| NVIDIA | TensorRT-LLM | use Python API to define Large Language | 6,907 | 722 | 656 | 13 | 5 | 83 | mins | License 2.0 | Smarty, PowerShell, C, Makefile, | https://github.com/NVIDIA/TensorRT-LLM |
| abetlen | llama-cpp-python | Python bindings for llama.cpp | 6,779 | 806 | 379 | 138 | 204 | 67 | mins | License | Makefile | https://github.com/abetlen/llama-cpp-python |
| huggingface | chat-ui | HuggingChat app | 6,437 | 890 | 209 | 75 | 10 | 77 | mins | License 2.0 | JavaScript, HTML, Shell, CSS, | https://github.com/huggingface/chat-ui |
| SillyTavern | SillyTavern | LLM Frontend for Power Users. | 6,301 | 1,911 | 307 | 117 | 79 | 53 | mins | Affero | Jupyter Notebook, Batchfile, Shell, | https://github.com/SillyTavern/SillyTavern |
| nat | openplayground | An LLM playground you can run on your laptop | 6,115 | 471 | 84 | 16 | 0 | 62 | mins | License | JavaScript, HTML | https://github.com/nat/openplayground |
| enricoros | big-agi | models and providing advanced AI/AGI | 4,492 | 1,024 | 148 | 37 | 16 | 50 | mins | License | Dockerfile | https://github.com/enricoros/big-agi |
| LostRuins | koboldcpp | GGUF models with KoboldAI's UI | 4,055 | 296 | 193 | 471 | 78 | 58 | mins | Affero | Objective-C, Lua, Makefile, | https://github.com/LostRuins/koboldcpp |
| ParisNeo | lollms-webui | Interface | 3,924 | 496 | 139 | 38 | 21 | 58 | mins | License 2.0 | Shell, Batchfile, Inno Setup, | https://github.com/ParisNeo/lollms-webui |
| minimaxir | simpleaichat | apps, with robust features and minimal code | 3,403 | 222 | 54 | 11 | 6 | 37 | 33 mins | License | Python | https://github.com/minimaxir/simpleaichat |
| deep-diver | LLM-As-Chatbot | LLM as a Chatbot Service | 3,240 | 384 | 19 | 7 | 0 | 54 | 7 mins | License 2.0 | Python, Jupyter Notebook | Chatbot |

https://docs.google.com/spreadsheets/d/1Xv38p90V3GiJXjq0a3qc24056Vicn1I5MG6QiFE6nVE/edit#gid=0

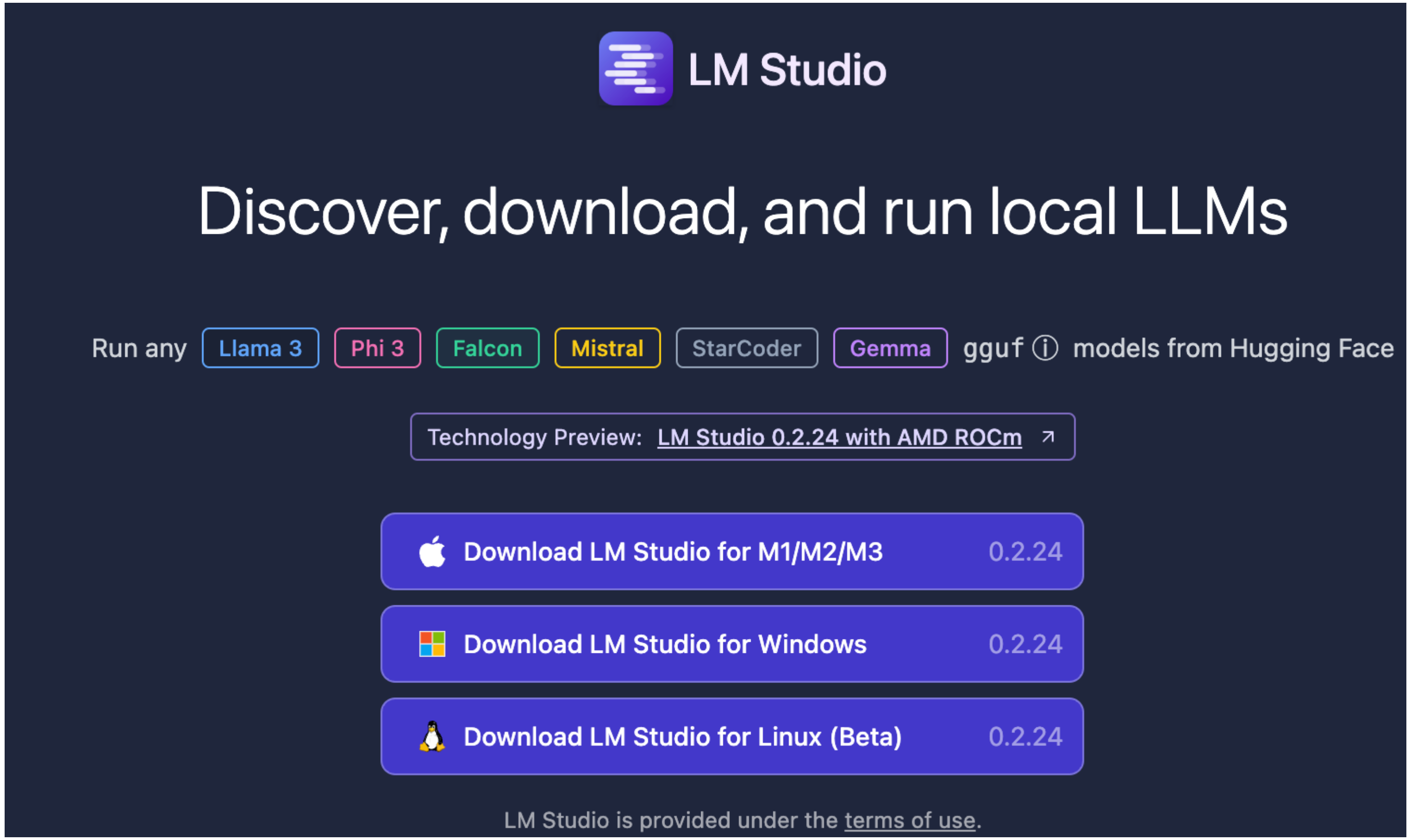
# Open-source Tools

1. All-in-one desktop solutions for accessibility
2. LLM inference via the CLI and backend API servers
3. Front-end UIs for connecting to LLM backends

Ollama

https://medium.com/thedeephub/50-open-source-options-for-running-llms-locally-db1ec6f5a54f

# LM Studio



LM Studio

Discover, download, and run local LLMs

Run any Llama 3 Phi 3 Falcon Mistral StarCoder Gemma gguf ⓘ models from Hugging Face

Technology Preview: LM Studio 0.2.24 with AMD ROCm ↗

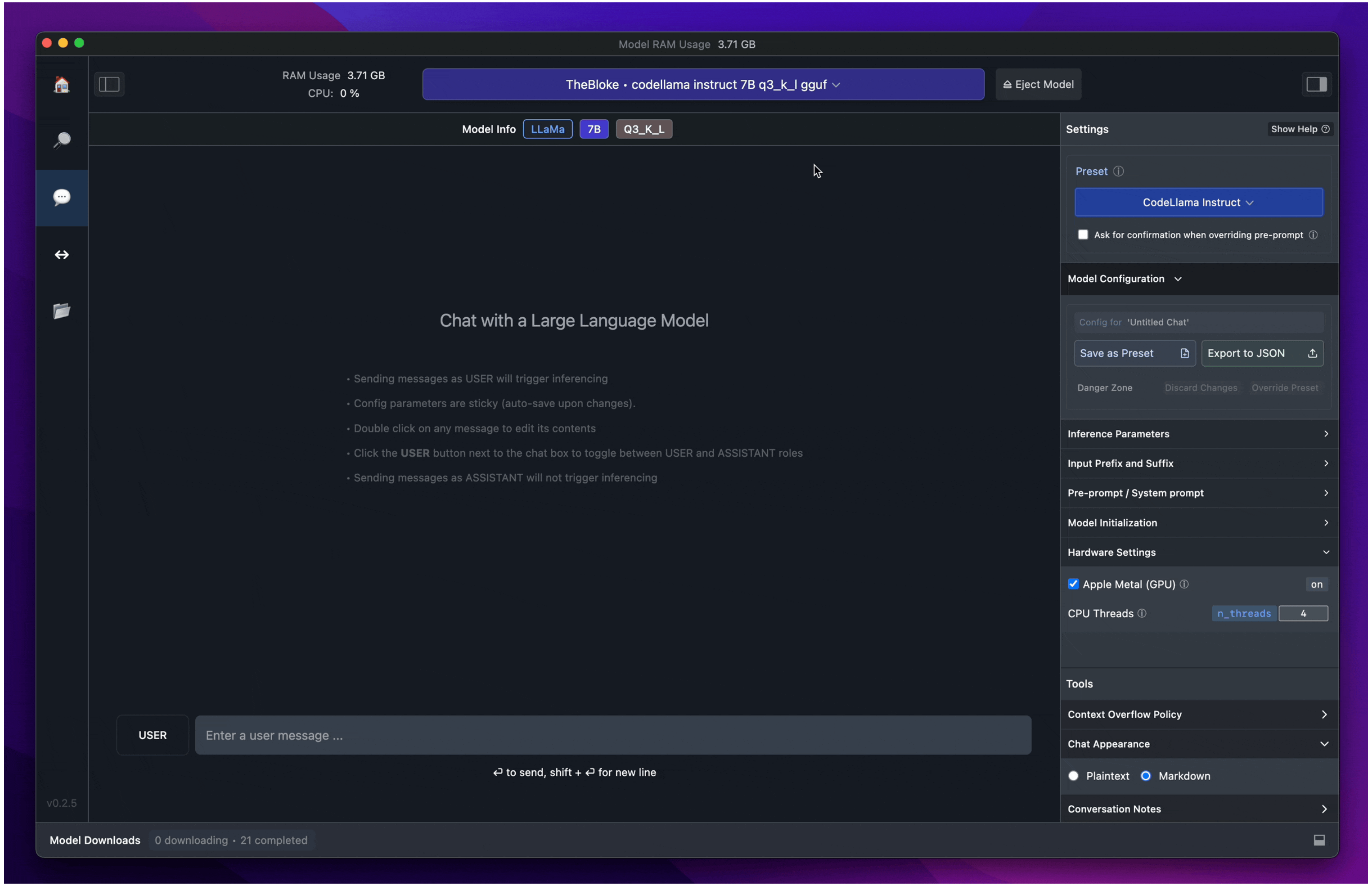
Download LM Studio for M1/M2/M3 0.2.24

Download LM Studio for Windows 0.2.24

Download LM Studio for Linux (Beta) 0.2.24

LM Studio is provided under the terms of use.

# LM Studio



Model RAM Usage 3.71 GB

RAM Usage 3.71 GB
CPU: 0 %

TheBloke • codellama instruct 7B q3_k_l gguf

Eject Model

Model Info LLaMa 7B Q3_K_L

Chat with a Large Language Model

- Sending messages as USER will trigger inferencing
- Config parameters are sticky (auto-save upon changes).
- Double click on any message to edit its contents
- Click the **USER** button next to the chat box to toggle between USER and ASSISTANT roles
- Sending messages as ASSISTANT will not trigger inferencing

USER

Enter a user message ...

↵ to send, shift + ↵ for new line

v0.2.5

Model Downloads 0 downloading • 21 completed

Settings Show Help

Preset: CodeLlama Instruct

□ Ask for confirmation when overriding pre-prompt

Model Configuration

Config for 'Untitled Chat'

Save as Preset

Export to JSON

Danger Zone Discard Changes Override Preset

Inference Parameters

Input Prefix and Suffix

Pre-prompt / System prompt

Model Initialization

Hardware Settings

☑ Apple Metal (GPU) on

CPU Threads n_threads 4

Tools

Context Overflow Policy

Chat Appearance

○ Plaintext ● Markdown

Conversation Notes

# LLama.cpp

## LLM Inference

- **Core Implementation**: Pure C/C++ with no external dependencies.
- **Hardware Optimization**:
  - **Apple Silicon**
  - **x86 Architectures**
- **Quantization**: Supports 1.5-bit to 8-bit quantization for efficient inference and reduced memory footprint.
- **GPU Support**:
  - **NVIDIA GPUs**: Custom CUDA kernels.
  - **AMD GPUs**: HIP compatibility.
- **Hybrid Inference**: CPU+GPU collaboration for handling models larger than GPU memory limits.

# llama.cpp (https://github.com/ggerganov/llama.cpp)

```
git clone https://github.com/ggerganov/llama.cpp.git
cd llama.cpp
make
```

# Download model

https://huggingface.co/mys/ggml_llava-v1.5-7b

```
wget https://huggingface.co/mys/ggml_llava-v1.5-7b/resolve/main/ggml-model-q4_k.gguf
```

```
wget https://huggingface.co/mys/ggml_llava-v1.5-7b/resolve/main/mmproj-model-f16.gguf
```

# Running Model

```
../llama.cpp/llava-cli -m ggml-model-q4_k.gguf --mmproj mmproj-model-f16.gguf  --image test1.png
```

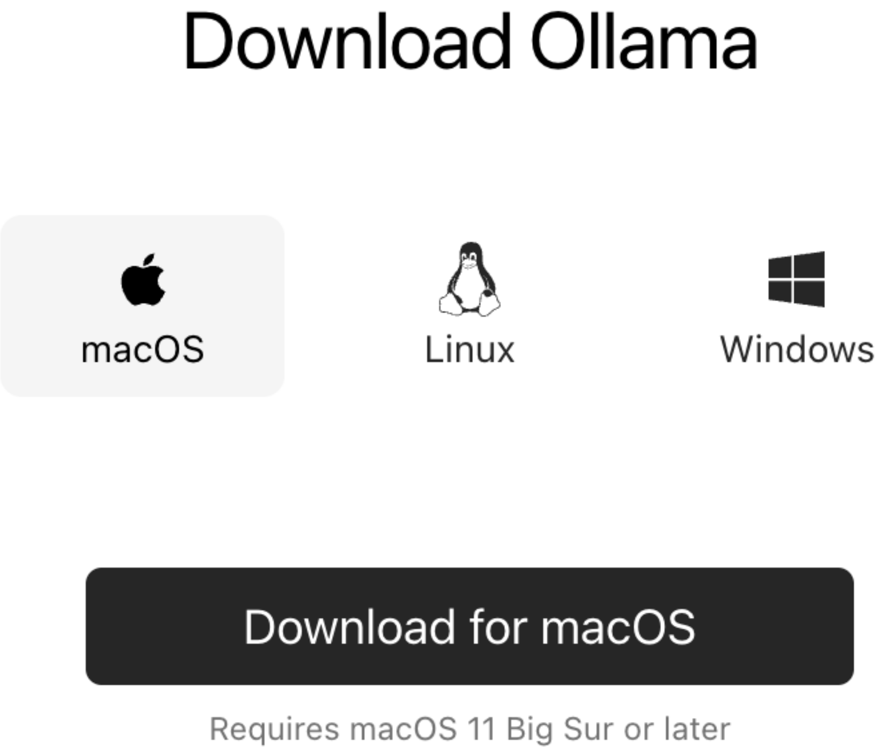
Output:

```
 The image showcases a woman in a lab setting, likely working on a project related to the technology industry. She appears to be focused on a task, using a computer as the center of her workspace.

There are several people in the image, including the main woman, with some individuals located near the left and right sides of the frame. Other people can be seen in the background, likely fellow colleagues or collaborators.

The scene is set in a well-equipped work environment, with several keyboards visible in the image. These keyboards might be used for controlling the various machines in the lab or for data processing and analysis.
```

# Ollama



Download Ollama

macOS

Linux

Windows

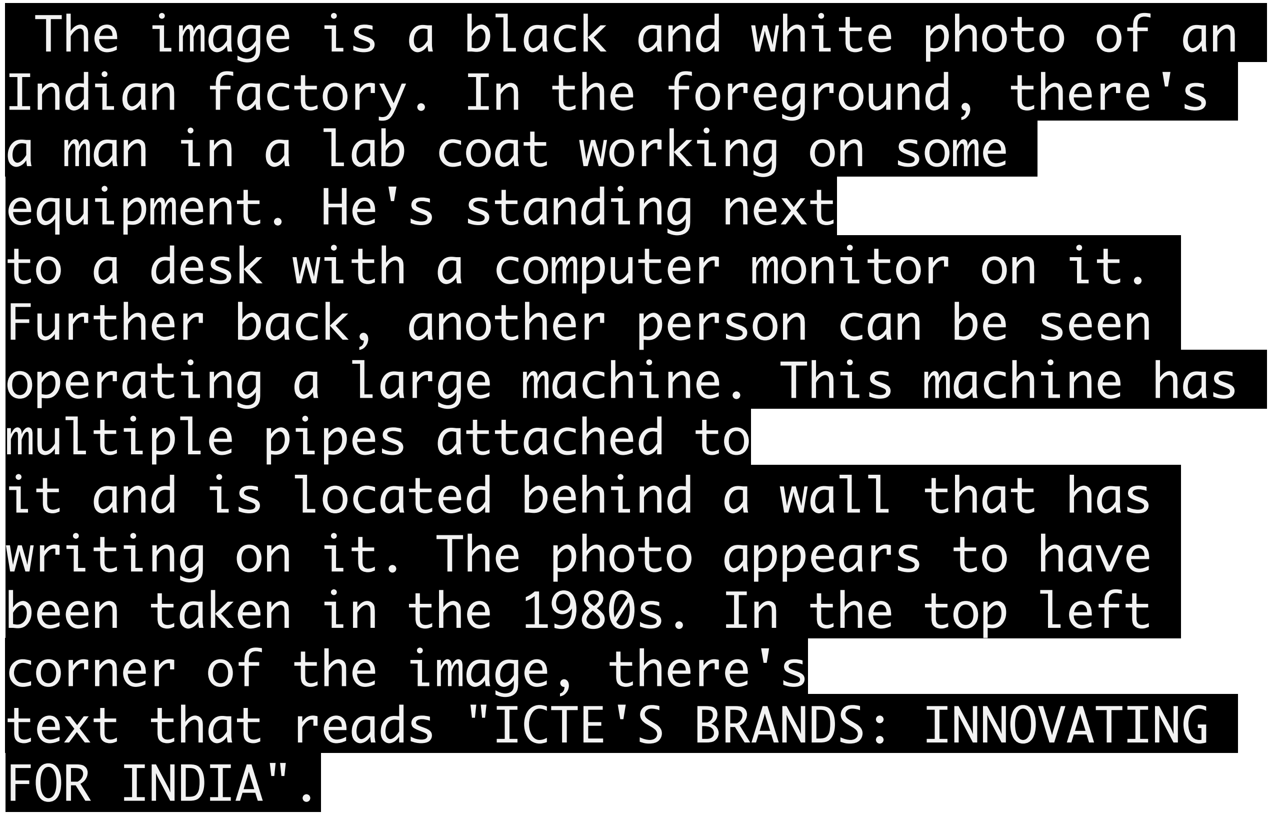
Download for macOS

Requires macOS 11 Big Sur or later

# Running model

```
ollama run llava-phi3 "describe this image: ./test1.png"
```

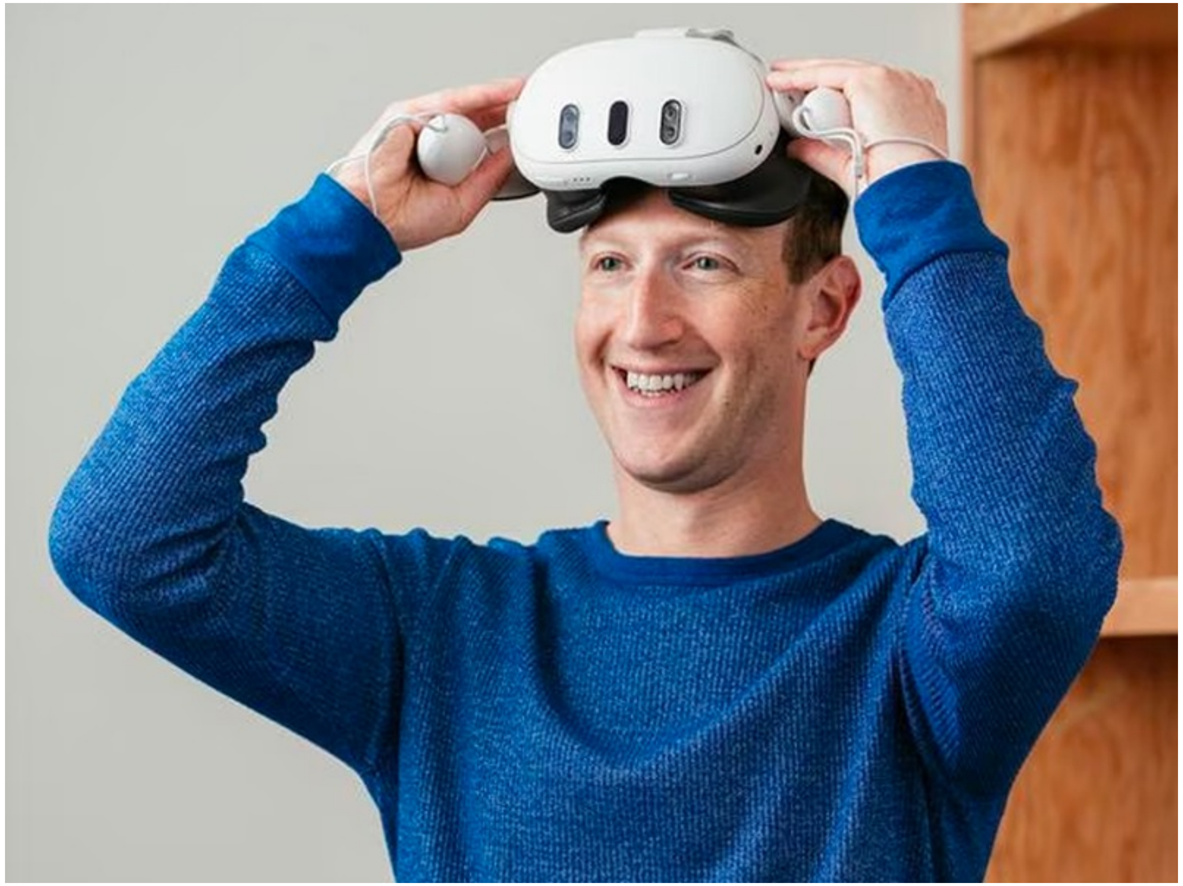
Output:

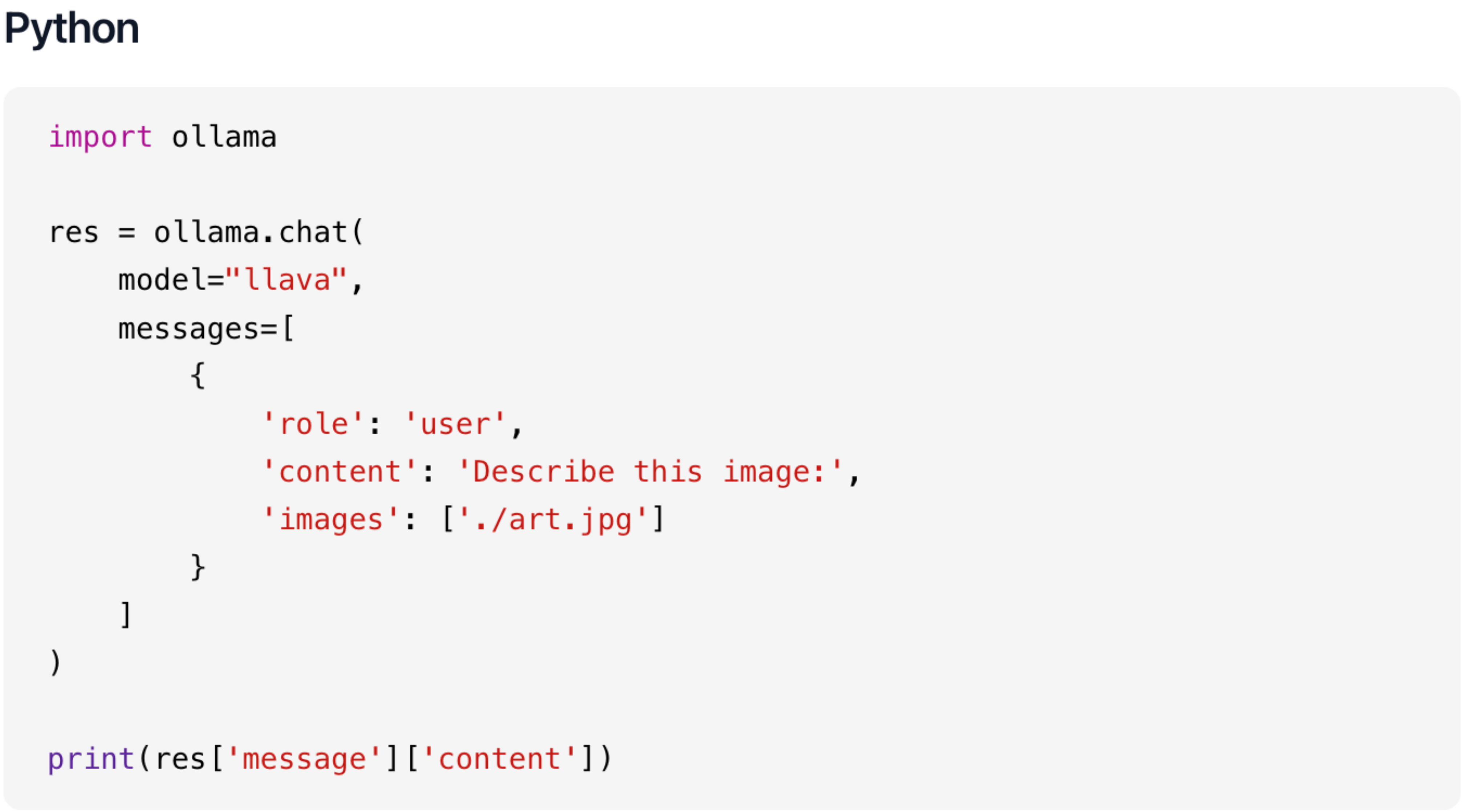


```
 The image is a black and white photo of an
Indian factory. In the foreground, there's
a man in a lab coat working on some
equipment. He's standing next
to a desk with a computer monitor on it.
Further back, another person can be seen
operating a large machine. This machine has
multiple pipes attached to
it and is located behind a wall that has
writing on it. The photo appears to have
been taken in the 1980s. In the top left
corner of the image, there's
text that reads "ICTE'S BRANDS: INNOVATING
FOR INDIA".
```

https://ollama.com/blog/vision-models

```
ollama run llava-phi3 "tell me what do you see in this picture? ./objectdetection.jpg"
```





## Python

```python
import ollama

res = ollama.chat(
    model="llava",
    messages=[
        {
            'role': 'user',
            'content': 'Describe this image:',
            'images': ['./art.jpg']
        }
    ]
)

print(res['message']['content'])
```