llamafile is the new best way to run a LLM on your own computer

29th November 2023

Mozilla’s innovation group and Justine Tunney just released llamafile, and I think it’s now the single best way to get started running Large Language Models (think your own local copy of ChatGPT) on your own computer.

A llamafile is a single multi-GB file that contains both the model weights for an LLM and the code needed to run that model—in some cases a full local server with a web UI for interacting with it.

The executable is compiled using Cosmopolitan Libc, Justine’s incredible project that supports compiling a single binary that works, unmodified, on multiple different operating systems and hardware architectures.

Here’s how to get started with LLaVA 1.5, a large multimodal model (which means text and image inputs, like GPT-4 Vision) fine-tuned on top of Llama 2. I’ve tested this process on an M2 Mac, but it should work on other platforms as well (though be sure to read the Gotchas section of the README, and take a look at Justine’s list of supported platforms in a comment on Hacker News).

1. Download the 4.26GB `llamafile-server-0.1-llava-v1.5-7b-q4` file from Justine’s repository on Hugging Face.

   ```
   curl -LO https://huggingface.co/jartine/llava-v1.5-7B-GGUF/resolve/main/llava-v1.5-7b-q4-server.llamafile
   ```

2. Make that binary executable, by running this in a terminal:

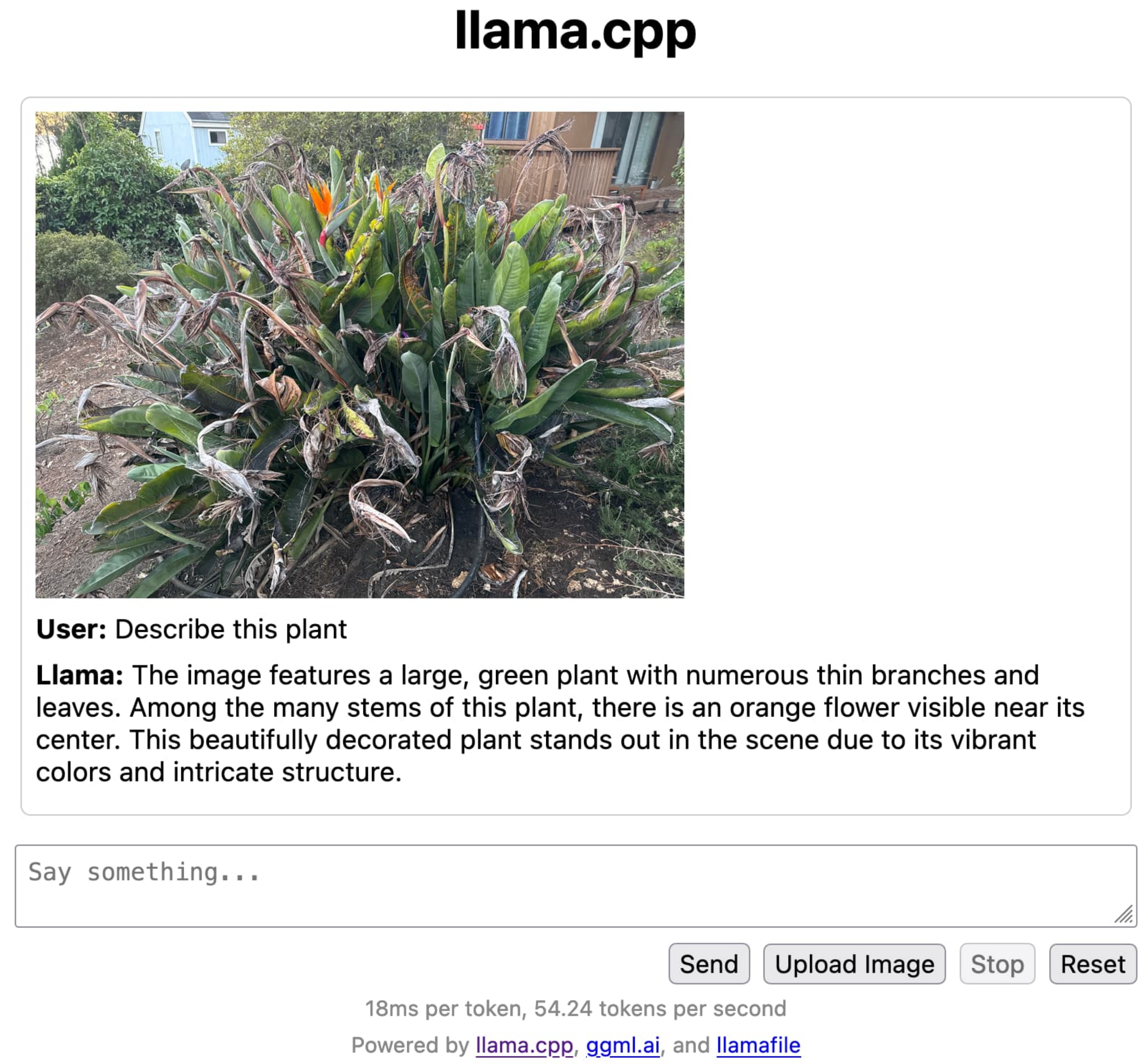
   ```
   chmod 755 llava-v1.5-7b-q4-server.llamafile
   ```

3. Run your new executable, which will start a web server on port 8080:

   ```
   ./llava-v1.5-7b-q4-server.llamafile
   ```

4. Navigate to `http://127.0.0.1:8080/` to start interacting with the model in your browser.

That’s all there is to it. On my M2 Mac it runs at around 55 tokens a second, which is *really* fast. And it can analyze images—here’s what I got when I uploaded a photograph and asked “Describe this plant”:



llama.cpp

User: Describe this plant

Llama: The image features a large, green plant with numerous thin branches and leaves. Among the many stems of this plant, there is an orange flower visible near its center. This beautifully decorated plant stands out in the scene due to its vibrant colors and intricate structure.

18ms per token, 54.24 tokens per second

Powered by llama.cpp, ggml.ai, and llamafile

How this works

There are a number of different components working together here to make this work.

- The LLaVA 1.5 model by Haotian Liu, Chunyuan Li, Yuheng Li and Yong Jae Lee is described in this paper, with further details on llava-vl.github.io.
- The models are executed using llama.cpp, and in the above demo also use the `llama.cpp` server example to provide the UI.
- Cosmopolitan Libc is the magic that makes one binary work on multiple platforms. I wrote more about that in a TIL a few months ago, Catching up with the Cosmopolitan ecosystem.

Trying more models

The llamafile README currently links to binaries for `Mistral-7B-Instruct`, `LLaVA 1.5` and `WizardCoder-Python-13B`.

You can also download a much smaller `llamafile` binary from their releases, which can then execute any model that has been compiled to GGUF format:

I grabbed `llamafile-server-0.1` (4.45MB) like this:

```
curl -LO https://github.com/Mozilla-Ocho/llamafile/releases/download/0.1/llamafile-se
chmod 755 llamafile-server-0.1
```

Then ran it against a 13GB `llama-2-13b.Q8_0.gguf` file I had previously downloaded:

```
./llamafile-server-0.1 -m llama-2-13b.Q8_0.gguf
```

This gave me the same interface at `http://127.0.0.1:8080/` (without the image upload) and let me talk with the model at 24 tokens per second.

One file is all you need

I think my favourite thing about llamafile is what it represents. This is a single binary file which you can download and then use, forever, on (almost) any computer.

You don’t need a network connection, and you don’t need to keep track of more than one file.

Stick that file on a USB stick and stash it in a drawer as insurance against a future apocalypse. You’ll never be without a language model ever again.

Posted 29th November 2023 at 8:54 pm · Follow me on Mastodon or Twitter or subscribe to my newsletter

More recent articles

- Datasette Enrichments: a new plugin framework for augmenting your data - 1st December 2023
- Prompt injection explained, November 2023 edition - 27th November 2023
- I'm on the Newsroom Robots podcast, with thoughts on the OpenAI board - 25th November 2023
- Weeknotes: DevDay, GitHub Universe, OpenAI chaos - 22nd November 2023
- Deciphering clues in a news article to understand how it was reported - 22nd November 2023
- Exploring GPTs: ChatGPT in a trench coat? - 15th November 2023
- Financial sustainability for open source projects at GitHub Universe - 10th November 2023
- ospeak: a CLI tool for speaking text in the terminal via OpenAI - 7th November 2023
- DALL-E 3, GPT4All, PMTiles, sqlite-migrate, datasette-edit-schema - 30th October 2023

Part of series **LLMs on personal devices**

7. Leaked Google document: "We Have No Moat, And Neither Does OpenAI" - May 4, 2023, 4:05 p.m.
8. My LLM CLI tool now supports self-hosted language models via plugins - July 12, 2023, 2:24 p.m.
9. Run Llama 2 on your own Mac using LLM and Homebrew - Aug. 1, 2023, 6:56 p.m.
10. **llamafile is the new best way to run a LLM on your own computer** - Nov. 29, 2023, 8:54 p.m.

mozilla 45 cosmopolitan 6 generativeai 328 llama 43 ai 371 llms 296

Next: Datasette Enrichments: a new plugin framework for augmenting your data

Previous: Prompt injection explained, November 2023 edition