@ Bhavishya-pandit

How To Run LLMS Locally

Unlock the Power of AI on Your Own Machine!

BHAVISHYA PANDIT

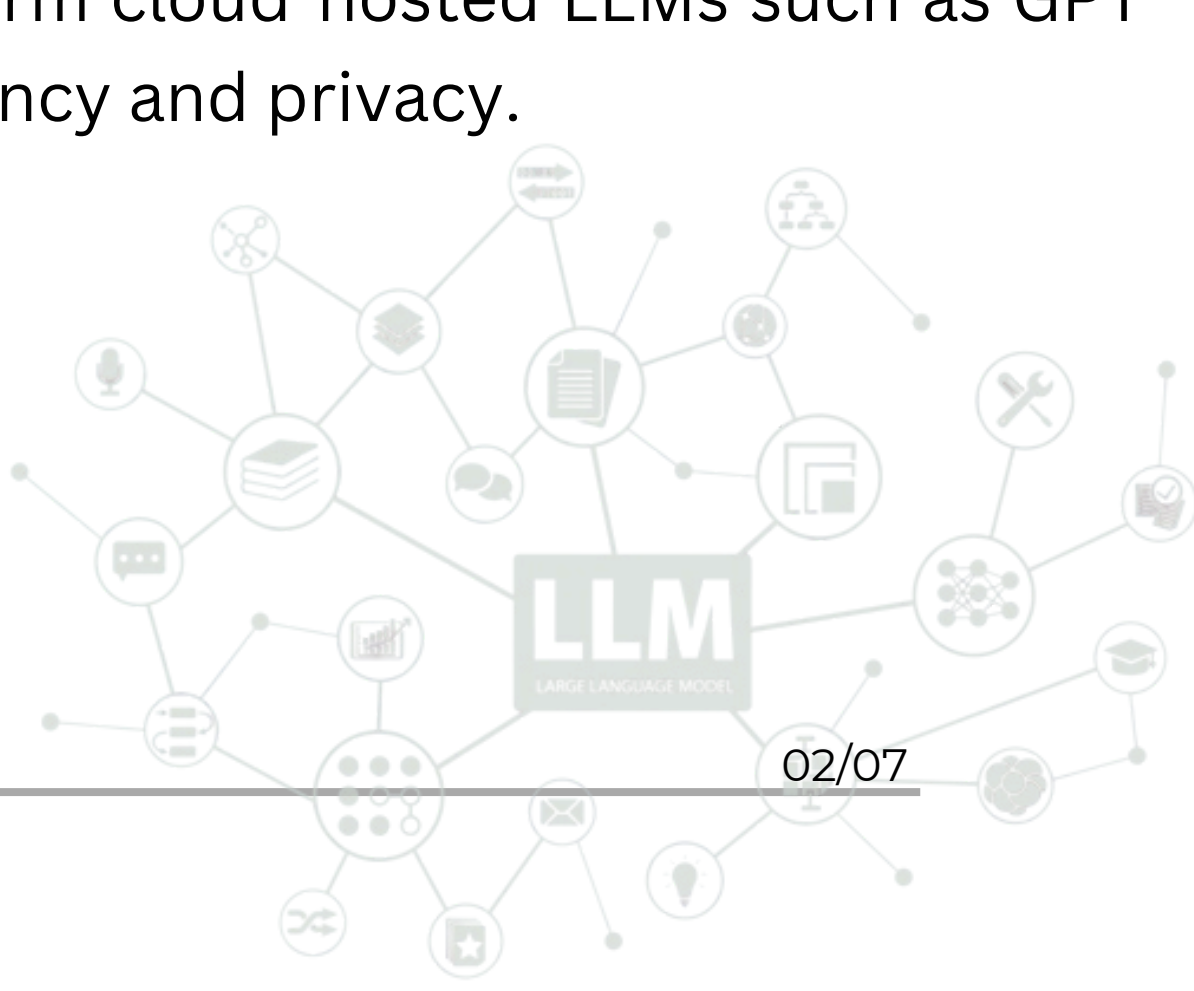
Introduction

Large Language Models (LLMs) are advanced AI models trained on vast amounts of text data to understand, generate, and manipulate human language. Running these models on our local devices has the following aspects:

Why Run LLMs Locally?

- **Privacy: Maintain control over your data.**
- **Customization: Adapt models to your unique needs.**
- **Accessibility: Operate independently of external servers.**

Locally running LLMs can outperform cloud-hosted LLMs such as GPT or Gemini in terms of speed, efficiency and privacy.



Tools: Ollama

ollama.com

Ollama is a cutting-edge tool designed to facilitate the running of large language models (LLMs) locally on personal or enterprise hardware. It allows easy access to LLMs such as Llama 3, Mistral, and Gemma through the terminal.

Download Ollama for your device

Install and Run

On terminal run:
$ ollama run llama3

gpt4all

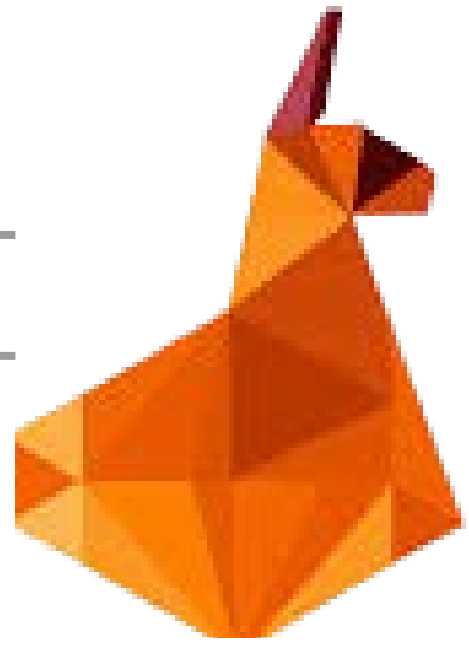
Tools: GPT4All

nomic.ai/gpt4all

GPT4All runs large language models (LLMs) privately on desktops & laptops. It helps you chat with your local files while maintaining your privacy allowing you to run LLMs on CPUs and GPUs. It fully supports Mac M Series chips, AMD, and NVIDIA GPUs.

Get the installer → Download a model of choice → Select model and chat with it

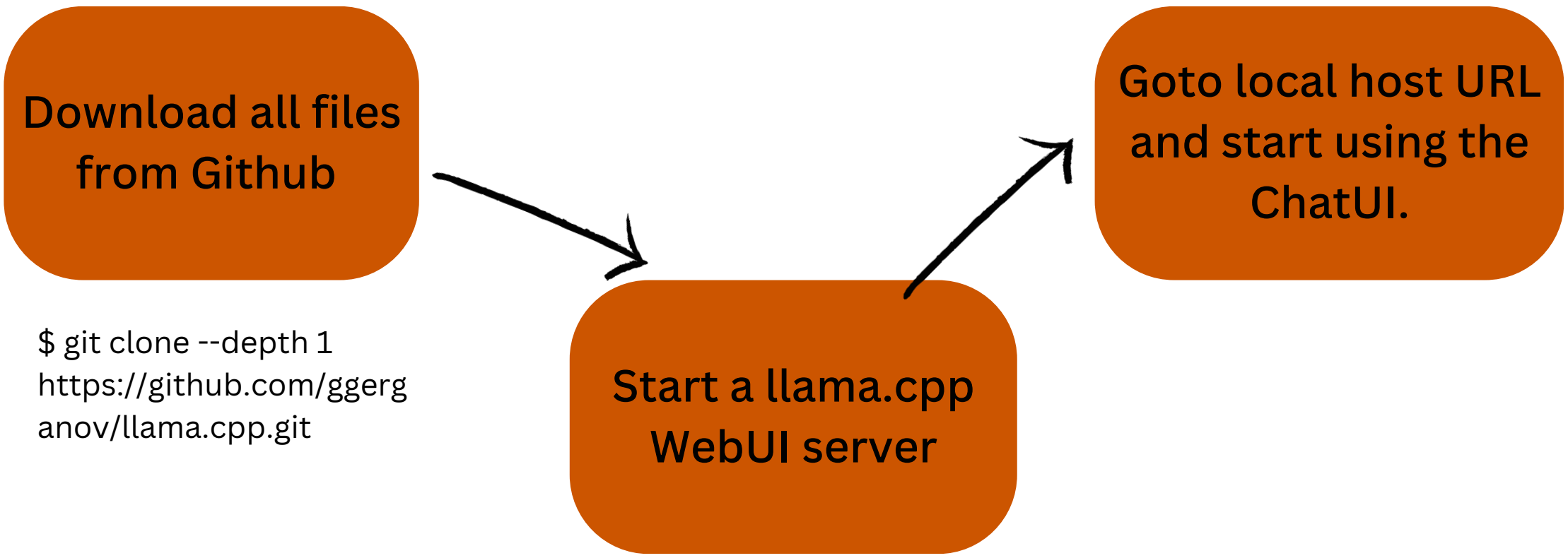
@ Bhavishya pandit



Tools: llama.cpp

github.com/ggerganov/llama.cpp

llama.cpp is an open-source LLM framework that requires minimal setup and gives state-of-the-art performance on a wide variety of hardware - locally and in the cloud. It is completely in C/C++ for faster and more efficient inference.



@ Bhavishya pandit

Tools: LM Studio

lmstudio.ai

LM Studio is similar to GPT4All however, it doesn’t allow connecting a local folder to generate context-aware answers. But it can run LLMs completely offline through in-app ChatUI or an OpenAI-compatible local server, allowing users to discover new LLMS from HuggingFcae Repos.

1. Get the installer
2. Download a model using search
3. Select model and chat with it using the ChatUI.

@ Bhavishya pandit

Which tool do you find most user-friendly for running LLMs locally and why?

JOIN THE CONVERSATION IN THE COMMENTS!

@ Bhavishya-pandit

Follow for more
AI/ML posts

SHARE YOUR THOUGHTS

SAVE FOR LATER

LIKE THIS POST