Jai Guru!

19 September 2024

Ollama

* Open-source platform that allows users to run LLMs locally on their own hardware
* Designed to be easy to use and set up
* Supports a variety of LLMs from different sources

Features

* Local deployment: allows users to run LLMs w/o the need for a constant internet connection. This helps to secure sensitive data and gives users complete control over the models
* Model support: supports a variety of LLMs, including Llama-2, Mistral, Phi-3, and Llama-3
* Customization: users can customize Ollama’s behavior or use specific models by modifying the Docker cmd.
* Modelfile: bundles model weights, configurations, and data into a single package, defined by a Modelfile
* CLI cmds: CLI offers a range of cmds and options to help the users interact with the LLM, such as adjusting the temp param or setting the top k-sampling value

Purpose

* Caters to specific use cases by offering a range of pre-built templates and tools designed for specialized language tasks. Users can leverage these resources to streamline model development for applications (sentiment analysis, text summarization, question-answering)

What is Ollama and how to use it?

* Exposes a local API, allowing developers to seamlessly integrate LLMs into their applications and workflows. This API facilitates efficient communication between the application and the LLM, enabling you to send prompts, receive responses, and leverage the full potential of these powerful AI models.
* Is an application you can download onto your computer or server to run open-souce gen AI LLMs (Meta’s Llama-3 -> competitive with OpenAI’s GPT3.5 and 4)

Is Ollama better than Chat GPT?

* ChatGPT runs the models in the cloud and hence, incurs significant costs, esp for large-scale or long-running tasks. However, Since Ollama leverages local hardware, it can be more cost-effective, particularly for user who already have access to powerful computing resources.

Cons

* Data privacy concerns due to external data handling
* Network latency can affect performance
* Ongoing operational costs

Pros

* Since it runs AI models directly on your device, Ollama eliminates the need to send your data to external servers. This local processing approach offers the advantage of enhanced privacy as your data never leaves your device and ensures complete confidentiality

Ollama commands

* Pull – used to pull a model from the Ollama model hub
* Rm – used to remove the already downloaded model from the local computer
* Cp – used to make a copy of the model
* List – used to see the list of downloaded models