LLaMA

- ollama và llama
- Ollama vs. LangChain

ollama và llama

1. LLaMA là gì?

LLaMA (*Large Language Model Meta AI*) là một dòng mô hình ngôn ngữ lớn (LLM) do **Meta (Facebook)** phát triển. Nó được thiết kế để **hiệu quả hơn GPT-3**, yêu cầu ít tài nguyên hơn để chạy nhưng vẫn có chất lượng cao.

□□ Các phiên bản chính:

- **LLaMA 1 (2023):** 7B, 13B, 30B, 65B tham số.
- **LLaMA 2 (2023):** 7B, 13B, 70B tham số, có bản `chat` để đối thoại tốt hơn.
- **LLaMA 3 (sắp ra mắt)**.

□□ Điểm mạnh:
□ Hiệu quả hơn GPT-3 (**cùng số lượng tham số nhưng thông minh hơn**).
□ Có thể chạy trên GPU yếu nếu dùng **quantization (GGUF, GPTQ, etc.)**.
□ Miễn phí sử dụng, có thể tải trên **Hugging Face**.

□□ Cách chạy LLaMA:

- **Dùng `transformers` (Hugging Face)**

  ```
  from transformers import AutoModelForCausalLM, AutoTokenizer
  model = AutoModelForCausalLM.from_pretrained("meta-llama/Llama-2-7b-chat-hf")
  tokenizer = AutoTokenizer.from_pretrained("meta-llama/Llama-2-7b-chat-hf")
  ```
- **Dùng `llama.cpp` (GGUF, tối ưu cho CPU + GPU)**
  `./main -m llama-2-7b.Q4_K.gguf --n-gpu-layers 5 -p "Hello"`

---

2. Ollama là gì?

Ollama là **một phần mềm giúp chạy LLM dễ dàng trên máy tính** (*local inference engine*). Nó hỗ trợ nhiều mô hình khác nhau (không chỉ LLaMA).

□□ Điểm mạnh của Ollama:
□ Cài đặt dễ dàng, chỉ cần `ollama run llama2`.
□ Hỗ trợ nhiều mô hình (LLaMA, Mistral, Phi-2, CodeLLaMA, v.v.).
□ Tự động tối ưu chạy trên **CPU/GPU** mà không cần cấu hình phức tạp.
□ Dùng `GGUF`, giúp chạy tốt trên máy yếu.

□□ Cách cài đặt Ollama:

- **Linux/macOS:**

  `curl -fsSL https://ollama.com/install.sh | sh`

- **Windows:**
  - Tải Ollama và cài đặt.

📌 Cách chạy LLaMA bằng Ollama:

- Chạy LLaMA-2:

```sh
ollama run llama2
```

- Chạy Mistral:

```sh
ollama run mistral
```

- □
  □□ So sánh Ollama & LLaMA

Đặc điểm	LLaMA	Ollama
Là gì?	Mô hình AI (LLM)	Phần mềm giúp chạy LLM
Ai phát triển?	Meta (Facebook)	Ollama
Chạy thế nào?	`transformers`, `llama.cpp`	`ollama run llama2`
Cấu hình?	Phải tự tải model, tối ưu GPU	Tự động tối ưu CPU/GPU
Hỗ trợ mô hình khác?	□ Chỉ LLaMA	□ Hỗ trợ LLaMA, Mistral, Phi-2, v.v.

□□ Kết luận:

- **LLaMA** = Mô hình AI do Meta phát triển.
- **Ollama** = Công cụ giúp chạy LLaMA (và nhiều mô hình khác) **dễ dàng hơn**.

Ollama vs. LangChain

Ollama và LangChain đều hỗ trợ làm việc với **mô hình ngôn ngữ lớn (LLM)**, nhưng **mục đích sử dụng** rất khác nhau.

---

1. Ollama – Chạy LLM cục bộ trên máy

□□ Tóm tắt:

- Chạy **mô hình LLM** ngay trên máy tính, **không cần internet**.
- Hỗ trợ nhiều mô hình **(Llama, Mistral, DeepSeek, Gemma, v.v.)**.
- Đơn giản, dễ dùng, **chỉ cần 1 lệnh để chạy AI**.

□□ Ví dụ sử dụng:
Cài đặt Ollama:

```
curl -fsSL https://ollama.com/install.sh | sh
```

Chạy mô hình Llama 3:

```
ollama run llama3
```

Chạy mô hình DeepSeek Coder:

```
ollama run deepseek-coder
```

□□ Ưu điểm của Ollama:
□ **Dễ sử dụng**: Chỉ cần `ollama run model_name` là chạy được AI.
□ **Không cần internet**: Phù hợp để chạy AI **ngoại tuyến**.
□ **Tối ưu GPU/CPU**: Dễ dàng chạy trên máy tính cá nhân.
□ **Hỗ trợ nhiều mô hình**: Tải và chạy **Llama, DeepSeek, Mistral, Gemma, v.v.**

□□ Nhược điểm:
□ **Không hỗ trợ tự động kết nối nhiều mô hình AI**.
□ **Không có pipeline** để kết hợp AI với dữ liệu từ nhiều nguồn.

---

2. LangChain – Xây dựng ứng dụng AI phức tạp

□□ Tóm tắt:

- Là **framework** giúp kết hợp **LLM với dữ liệu**.
- Dùng để **tạo chatbot, tìm kiếm tài liệu (RAG), tự động hóa AI**.
- Kết nối với nhiều nền tảng như **OpenAI, Ollama, Google Gemini, v.v.**

□□ Ví dụ sử dụng LangChain với Ollama:

```
from langchain_community.llms import Ollama

llm = Ollama(model="mistral")
response = llm.invoke("Giải thích Định lý Pythagoras.")
print(response)
```

□□ Ưu điểm của LangChain:
□ **Linh hoạt**: Kết nối nhiều AI khác nhau (GPT-4, Llama, Claude, v.v.).
□ **Xử lý dữ liệu từ nhiều nguồn**: Tích hợp với **PDF, website, cơ sở dữ liệu**.
□ **Tạo pipeline AI phức tạp**: Xây dựng ứng dụng như ChatGPT với dữ liệu riêng.

□□ Nhược điểm:
□ **Không thể chạy AI cục bộ** (cần kết nối Ollama hoặc OpenAI).
□ **Cấu hình phức tạp hơn Ollama**.

Chọn cái nào?

Tính năng	Ollama □□	LangChain □□
Chạy mô hình AI cục bộ	□ Có	□ Không
Hỗ trợ nhiều mô hình AI	□ Có	□ Có
Xử lý dữ liệu từ PDF, DB	□ Không	□ Có
Tạo chatbot, pipeline AI	□ Không	□ Có
Dễ sử dụng	□ Rất dễ	□ Phức tạp hơn
Không cần Internet	□ Có	□ Không

Kết luận:

- **Nếu muốn chạy AI cục bộ trên máy → Ollama**
- **Nếu muốn xây dựng ứng dụng AI phức tạp → LangChain**
- **Muốn kết hợp cả hai?** Dùng **LangChain + Ollama** để vừa có **AI cục bộ**, vừa có **pipeline AI mạnh mẽ**.