

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?	<code>transformers</code> , <code>llama.cpp</code>	<code>ollama run llama2</code>
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**.