

Khóa Học Ngày Thứ Hai: Giới Thiệu & Nội Dung

1. Tổng kết ngày thứ nhất

Chuẩn bị những gì:

- Chạy mô hình cục bộ với Ollama.
- Viết mã gọi các mô hình tiên tiến của OpenAI.
- Phân biệt giữa Prompt của Hệ thống và Người dùng.
- Thực hành tóm tắt văn bản.

2. Nội dung học hôm nay

- Hiểu các yếu tố cơ bản để thành thạo LLM.
- Chuẩn bị cho sự thành công khi làm việc với AI.
- Nhận diện các mô hình tiên tiến hàng đầu và cách sử dụng chúng.

LLM Engineering và các công cụ hỗ trợ

Ba khía cạnh chính trong **LLM Engineering**:

- Hiểu về các mô hình LLM.
- Sử dụng công cụ hỗ trợ như Hugging Face, LangChain, Gradio, Weights & Biases, Modal.
- Áp dụng AI vào giải quyết vấn đề kinh doanh.

3. Định hướng khóa học

- Phù hợp với nhiều mức độ kỹ năng, từ cơ bản đến trung cấp.
- Khuyến khích xem lại kiến thức nền tảng về tokens và Python cơ bản.
- Cung cấp hướng dẫn trong tuần 1 để nâng cao khả năng lập trình Python.
- Khuyến học viên:
 - **Thực hành theo mã mẫu.**
 - **Tự debug khi gặp lỗi.**
 - **Chia sẻ mã nguồn trên GitHub để xây dựng danh mục dự án.**
- Khuyến khích sáng tạo, phát triển dự án cá nhân và đóng góp cho cộng đồng mã nguồn mở.

4. Giới thiệu các mô hình AI

Mô hình đóng nguồn (Super Scalers)

- **GPT (OpenAI):** Ra mắt cuối năm 2022, mạnh mẽ và phổ biến.
- **Claude (Anthropic):** Được ưa chuộng trong cộng đồng khoa học dữ liệu.
- **Gemini (Google) & Command R (Cohere):** Các mô hình cạnh tranh mạnh mẽ.

Mô hình mã nguồn mở

- **Llama (Meta):** Một trong những mô hình nổi bật nhất.
- **Mistral:** Mô hình hỗn hợp từ nhiều chuyên gia nhỏ.
- **Quen (Alibaba Cloud), Gemma (Google), Fi (Microsoft):** Các mô hình nguồn mở đáng chú ý.

Các phương thức sử dụng mô hình AI

- Chat Interfaces (Giao diện trò chuyện): ChatGPT, chạy trên đám mây.
- Cloud APIs (Giao diện lập trình API đám mây): Gọi API của mô hình (ví dụ: OpenAI API).
- Managed AI Cloud Services (Dịch vụ AI đám mây quản lý): Sử dụng dịch vụ AI của Amazon, Google, Microsoft.

So sánh sự khác biệt

- Giao diện trò chuyện: Miễn phí hoặc có phí đăng ký.
- Cloud APIs: Không phí đăng ký, nhưng tính phí theo yêu cầu API.
- Managed AI Cloud Services: Dịch vụ AI đám mây linh hoạt, hỗ trợ nhiều mô hình.

5. Cách sử dụng mã nguồn mở và cài đặt mô hình AI

- **Hugging Face:** Làm việc trực tiếp với mô hình qua Python (Google Colab).
- **Ollama:** Chạy mô hình AI cục bộ mà không cần mã nguồn gốc.

So sánh sự khác biệt

- Hugging Face: Điều khiển chi tiết hơn.
- Ollama: Chạy mô hình cục bộ mà không cần tinh chỉnh sâu.

6. Hướng dẫn khởi động môi trường làm việc

- Mở Anaconda Prompt (Windows) hoặc Terminal (Mac).
- Kích hoạt môi trường conda hoặc virtualenv.
- Mở JupyterLab và điều hướng đến thư mục bài tập.

7. Bài tập thực hành ngày thứ hai

- Gọi mô hình LLaMA từ mã Python.
- Cập nhật dự án tóm tắt từ ngày 1 để sử dụng Ollama thay vì OpenAI.
- Lợi ích của LLaMA:

Không mất phí API.

Dữ liệu không rời khỏi máy tính.

- Hạn chế: Mô hình mã nguồn mở không mạnh bằng mô hình thương mại.

Hướng dẫn cài đặt và chạy Ollama

1. Tải về từ trang web Ollama.
2. Chạy lệnh `ollama serve` để khởi động dịch vụ.
3. Kiểm tra trạng thái tại `localhost:11434`.
4. Debug lỗi nếu gặp sự cố.

Gọi LLaMA từ Python

1. Thiết lập API URL, model (```*****```), dữ liệu đầu vào.
2. Gửi yêu cầu HTTP bằng thư viện ````.
3. Trích xuất và hiển thị kết quả từ phản hồi JSON.

Sử dụng thư viện Ollama để gọi mô hình dễ dàng hơn

Thay vì gửi HTTP request thủ công, có thể dùng

*****ollama.chat().

Lưu ý: Kết quả có thể thay đổi mỗi lần gọi do tính ngẫu nhiên của mô hình.

8. Giới thiệu nội dung buổi học tiếp theo

- Thực hành với sáu mô hình Frontier.
- So sánh hiệu suất giữa các mô hình.
- Khám phá điểm mạnh và yếu của từng mô hình.