I/ Tóm tắt nội dung tuần 1 của LLM

Các tệp này có nội dung liên quan đến kỹ thuật làm việc với mô hình ngôn ngữ lớn (LLM), cụ thể là:

Giới thiệu & Ứng dụng LLM

- **LLM Engineering**: Cách triển khai nhanh chóng các mô hình ngôn ngữ lớn.
- **GPT & Transformer**: Khái niệm cốt lõi và thực hành.
- **Ứng dụng AI trong NLP & Chatbot**: Xây dựng hệ thống giao tiếp thông minh.
- **Công cụ & Nền tảng**: OpenAI, Hugging Face – hỗ trợ phát triển mô hình.
- **Bảo mật API**: Sử dụng .env để bảo vệ API Keys.
- **Fine-tuning mô hình**: Hugging Face & PyTorch.
- **Inference**: Dự đoán từ mô hình đã huấn luyện.
- **Thiết lập môi trường Python**: Hướng dẫn thực hành.
- **Thử nghiệm GPT**: Kiểm tra khả năng xử lý câu hỏi.
- **Đánh giá Claude của Anthropic**: So sánh với các mô hình khác.

So sánh các Mô hình AI hàng đầu

- **Google Gemini**: Chính xác nhưng chưa tinh tế, hài hước kém hơn GPT.
- **Cohere Command+**: Chuyên sâu, có cấu trúc rõ ràng, chi tiết hơn Claude.
- **Meta AI (Llama 3)**: Khá tốt nhưng chưa mạnh bằng các mô hình khác, có khả năng tạo ảnh nhưng hạn chế.
- **Perplexity**: Không phải LLM mà là công cụ tìm kiếm, mạnh về thông tin thời sự.

Phát triển Kiến trúc Transformer

- **Nguồn gốc & Lịch sử**:
  - *Attention is All You Need* (2017) → GPT-4 Turbo (2024).
  - Transformer là nền tảng của nhiều mô hình AI hiện đại.
- **Cách hoạt động**:
  - Self-Attention giúp dự đoán từ tiếp theo từ dữ liệu lớn.
- **Tranh luận về AI**:
  - *AI chỉ là bộ máy thống kê?* (Stochastic Parrot).
  - *AI có trí tuệ nổi lên?* (Emergent Intelligence).
- **Ứng dụng thực tế**:
  - Thử nghiệm khả năng AI trên *edward.com*, mời người dùng đánh giá.

Xu hướng AI mới nhất

- **Prompt Engineering**: Từng lương cao ($500K) nhưng giảm dần do tự động hóa.
- **Custom GPTs**: GPT Store phổ biến nhưng dần bão hòa.
- **Copilots AI**: Microsoft Copilot & GitHub Copilot được tích hợp rộng rãi.
- **Agentic AI**: LLMs kết hợp giải quyết vấn đề phức tạp, có khả năng ghi nhớ và tự chủ.

Tham số (Weights) trong LLMs

- **Mô hình truyền thống**: Chỉ có 20–200 tham số.
- **Mô hình LLM hiện đại**:
  - GPT-1: 117 triệu
  - GPT-2: 1.5 tỷ
  - GPT-3: 175 tỷ
  - GPT-4: 1.76 nghìn tỷ
  - **Frontier models**: Lên đến 10 nghìn tỷ tham số.
- **Mô hình mã nguồn mở**:
  - Gemini: 2 tỷ tham số
  - Llama: 2B, 8B, 70B, 405B

Tokenization và Xử Lý Văn Bản

- **Tokenization** là quá trình chia văn bản thành các đơn vị nhỏ hơn để mô hình xử lý.
- Các phương pháp **tokenization**:
    - **Character-based**
    - **Word-based**
    - **Subword-based (BPE, SentencePiece, WordPiece)**

=> Giúp cân bằng giữa kích thước từ vựng và hiệu suất xử lý.

Context Windows và Giới Hạn Token

- **Context window** là số lượng token mà mô hình AI có thể xử lý trong một lần, ảnh hưởng đến khả năng ghi nhớ.
- **Giới hạn token** của các mô hình:
    - **GPT-4:** 128K tokens
    - **Claude:** 200K tokens
    - **Gemini 1.5 Flash:** 1 triệu tokens

Chi Phí API và Subscription Plan

Có hai dạng chi phí:

1. **Đăng ký** (Subscription) → ChatGPT Pro ($20/tháng).
2. **Trả phí API** → Tính theo số lượng token đầu vào/đầu ra.

Các mô hình phổ biến: OpenAI API, Claude API, Gemini API.

Có **rate limiting** để kiểm soát mức sử dụng.

Kỹ Thuật Prompting

- **One-shot prompting:** Cung cấp một ví dụ duy nhất để hướng dẫn AI.
- **Multi-shot prompting:** Dùng nhiều ví dụ để tăng độ chính xác.
- **tructured Outputs:** Yêu cầu AI phản hồi dưới dạng JSON để đảm bảo dữ liệu có tổ chức.

Ứng Dụng AI trong Kinh Doanh & Web Scraping

- **Tạo brochure tiếp thị bằng AI**

Sử dụng OpenAI API để thu thập và tổng hợp nội dung từ web.

- **Web Scraping bằng JupyterLab**

Dùng BeautifulSoup, Scrapy, Python Requests để trích xuất dữ liệu.

- **Markdown & Streaming Optimization**

Tối ưu hiển thị phản hồi theo thời gian thực bằng Markdown & streaming mode.

Phát Triển AI Tutor & Hệ Thống Hỗ Trợ

- **Xây dựng trợ lý học tập cá nhân hóa**

=> Kết hợp GPT-4 và Llama.

- **So sánh chất lượng phản hồi giữa các mô hình**
- **Gợi ý phát triển UI với Gradio**

=> Tạo hệ thống hỗ trợ khách hàng đa phương tiện.

II/ Các từ khóa quan trọng liên quan đến khóa học LLM

1. LLM (Large Language Models) là gì?

LLM (Mô hình ngôn ngữ lớn) là một dạng trí tuệ nhân tạo được huấn luyện trên lượng dữ liệu văn bản khổng lồ để hiểu và tạo nội dung ngôn ngữ tự nhiên. Các mô hình tiêu biểu gồm GPT-4, BERT, LLaMA.

2. GPT (Generative Pre-trained Transformer)

GPT là một mô hình Transformer được huấn luyện trước (pre-trained) trên dữ liệu lớn, có khả năng sinh văn bản tự nhiên, trả lời câu hỏi và hỗ trợ chatbot thông minh.

3. Transformer hoạt động như thế nào?

Transformer sử dụng cơ chế **self-attention**, giúp hiểu mối quan hệ giữa các từ trong câu mà không phụ thuộc vào vị trí. Đây là nền tảng của nhiều mô hình AI như GPT, BERT.

4. Ứng dụng của AI và Chatbot

AI và chatbot được sử dụng trong nhiều lĩnh vực:

- **Chăm sóc khách hàng**: Tự động phản hồi, hỗ trợ 24/7.
- **Dịch thuật tự động**: Google Translate, DeepL.
- **Phân tích ngôn ngữ**: Hỗ trợ viết, tổng hợp nội dung.

5. Công nghệ hỗ trợ phát triển LLM

- **Hugging Face**: Thư viện mã nguồn mở giúp tải và huấn luyện mô hình NLP.
- **OpenAI**: Nhà phát triển GPT, cung cấp API chatbot & AI.
- **Anaconda**: Môi trường Python hỗ trợ quản lý thư viện và mô hình ML.

6. Machine Learning (Học máy) là gì?

Machine Learning (ML) là một nhánh của AI, giúp máy tính học từ dữ liệu để đưa ra dự đoán hoặc quyết định mà không cần lập trình cứng.

7. Môi trường lập trình trong phát triển AI & LLM

Thiết lập môi trường lập trình giúp tối ưu quá trình phát triển mô hình AI, từ việc cài đặt thư viện đến kiểm thử mô hình.

8. API (Application Programming Interface)

API là giao diện cho phép các phần mềm giao tiếp với nhau thông qua yêu cầu và phản hồi dữ liệu.

9. .env File - Tệp cấu hình bảo mật API Keys

.env được sử dụng để lưu trữ các biến môi trường, giúp bảo vệ API Keys và thông tin nhạy cảm khỏi bị lộ.

10. Bảo mật trong AI & lập trình

Bảo vệ dữ liệu, mã nguồn và API Keys khỏi rò rỉ hoặc bị tấn công.

11. Fine-tuning - Tinh chỉnh mô hình AI

Fine-tuning giúp điều chỉnh mô hình AI theo nhu cầu cụ thể, tăng độ chính xác cho các bài toán chuyên biệt.

12. Claude (Anthropic)

Claude là mô hình AI của Anthropic, được thiết kế để an toàn và dễ kiểm soát hơn so với các AI truyền thống. Phiên bản mới nhất **Claude 3.5 Sonnet** được đánh giá cao về hiệu suất.

13. AI Benchmarking

Là quá trình đo lường và so sánh hiệu suất của các mô hình AI dựa trên các tiêu chí:

- **Độ chính xác**
- **Tốc độ xử lý**
- **Khả năng hiểu ngữ cảnh**
- **Khả năng suy luận**

14. Ứng dụng thực tế của AI

AI được ứng dụng trong giáo dục, y tế, tài chính, thương mại điện tử, marketing, lập trình, và nhiều lĩnh vực khác.

15. AI Model Evaluation (Benchmarking)

Các bài kiểm tra phổ biến:

- **MMLU** (Massive Multitask Language Understanding)
- **GSM8K** (bài toán toán học cấp tiểu học)

16. Ứng dụng AI trong kinh doanh

Doanh nghiệp ứng dụng AI để **tự động hóa quy trình**, **phân tích dữ liệu**, **tạo nội dung**, **viết code**, và nhiều mục đích khác.

17. Các mô hình AI hàng đầu

- **Gemini AI (Google)**: Mô hình đa phương thức (văn bản, hình ảnh, âm thanh).
- **Cohere Command+**: Tập trung vào khai thác tri thức và truy vấn thông tin.
- **Meta AI (Llama)**: Mô hình của Meta (Facebook).
- **Perplexity AI**: Chuyên tìm kiếm và tổng hợp thông tin chính xác.

18. So sánh các mô hình AI

Các tiêu chí so sánh:

- **Độ chính xác**
- **Khả năng hiểu ngữ cảnh**
- **Phong cách trả lời**

19. Inference Mode - Chế độ suy luận

Inference ảnh hưởng đến cách AI xử lý thông tin và đưa ra câu trả lời.

20. Multimodal AI

AI có khả năng xử lý nhiều loại dữ liệu như văn bản, hình ảnh, âm thanh.

21. Các khái niệm quan trọng trong AI

- **Tokenization**: Chia văn bản thành các đơn vị nhỏ hơn để mô hình xử lý.
- **Context Windows**: Số lượng token AI có thể xử lý trong một lần.
- **API Pricing**: Chi phí API của OpenAI, Claude, Gemini.
- **Prompt Engineering**: Kỹ thuật viết prompt để tối ưu AI.
- **Structured Outputs**: Định dạng đầu ra của AI (JSON, XML).
- **AI Marketing**: Ứng dụng AI trong tiếp thị và quảng cáo.
- **Web Scraping**: Thu thập dữ liệu từ web bằng AI.
- **Multi-shot Prompting**: Cung cấp nhiều ví dụ để AI tăng độ chính xác.
- **LLM Tutor**: Xây dựng trợ lý học tập bằng AI.
- **Markdown & Streaming Responses**: Tối ưu phản hồi AI theo thời gian thực.

III/ Các công nghệ được đề cập trong tuần 1 khóa học LLM

1. Mô hình ngôn ngữ lớn (LLM - Large Language Models)

- **GPT (Generative Pre-trained Transformer)**
- **Transformer**
- **Machine Learning (ML)**
- **AI (Artificial Intelligence - Trí tuệ nhân tạo)**

2. Frameworks & Thư viện hỗ trợ

- **PyTorch** (Framework Machine Learning)
- **Hugging Face** (Thư viện NLP)
- **OpenAI** (Cung cấp API cho GPT)

3. Các khái niệm quan trọng

- **Inference** (Dự đoán kết quả từ mô hình đã huấn luyện)
- **API** (Giao tiếp với mô hình AI)
- **Python** (Ngôn ngữ lập trình chính)
- **.env File** (Bảo mật API Keys)

4. Các mô hình AI nổi bật

- **Claude 3.5 Sonnet** (Anthropic)
- **GPT-4** (OpenAI)
- **Gemini AI** (Google)
- **LLaMA** (Meta AI)

5. Đánh giá & Ứng dụng AI
- **AI Model Evaluation (Benchmarking)**
- **Business AI Applications** (Ứng dụng AI trong doanh nghiệp)

6. Các mô hình AI hàng đầu khác
- **Cohere Command Plus**
- **Meta AI**
- **Perplexity AI**
- **Multimodal AI** (Xử lý văn bản, hình ảnh, âm thanh)

7. Công nghệ Tokenization
- **BPE (Byte Pair Encoding)**
- **SentencePiece**
- **WordPiece**

8. API & Pricing
- **OpenAI API**
- **Claude API**
- **Gemini API**

9. Công cụ & Ứng dụng AI
- **JupyterLab** (Môi trường lập trình Python)
- **BeautifulSoup** (Thu thập dữ liệu từ web - Web Scraping)
- **Python Requests** (Thư viện gọi API)
- **Gradio** (Giao diện AI đơn giản)
- **AI Agent** (Tác tử AI thực hiện nhiệm vụ tự động)