**How do Large Language Models fit into the AI Framework in this Course?**

**Các Mô Hình Ngôn Ngữ Lớn phù hợp với Khung AI trong khóa học này như thế nào?**

**Think like a human?** – Suy nghĩ như con người (mô phỏng quá trình nhận thức).

**Act like a human?** – Hành động như con người (cách Turing Test).

**Think rationally?** – Suy nghĩ hợp lý (dựa trên logic).

**Act rationally?** – Hành động hợp lý (mô hình agent hợp lý).

**“LLMs làm gì? Chúng có hành động hợp lý không?”**

Hãy hỏi chatbot xem nó có hành động hợp lý không.

Và liệu nó có phải là một tác nhân thông minh (intelligent agent) không.

**Kết luận:**

LLMs không thực sự suy nghĩ như con người (không có ý thức, cảm xúc).

LLMs có thể hành động giống con người ở mức độ nào đó (trả lời hội thoại tự nhiên).

LLMs không luôn suy nghĩ hợp lý (có thể sai logic, "ảo giác").

Nhưng LLMs có thể hành động hợp lý trong nhiều tình huống (giúp trả lời câu hỏi, giải quyết vấn đề theo mục tiêu).

* **Vì vậy, trong 4 hướng tiếp cận AI, LLMs thường gần với “Act Rationally” – Hành động hợp lý nhất, nhưng chưa hoàn hảo.**

**Turing Test: Large Language Models (LLMs)**

Minh họa: 1 người đang cố gắng phân biệt giữa người và AI

**Liệu một LLM hiện đại có vượt qua được Turing Test không?**

* **Bạn có bị “lừa” không?**

Có thể bị lừa trong ngắn hạn, tuy nhiên, nếu đi sâu vào ngữ cảnh chuyên biệt, chuyên môn, hoặc các vấn đề thiên về cá nhân hóa, trải nghiệm thì sẽ phát hiện được.

* **Tại sao LLM vượt qua hoặc không vượt qua bài kiểm tra?**

Vượt qua: duy trì ngữ cảnh hội thoại, mô phỏng cảm xúc dựa trên lượng dữ liệu khổng lồ mà LLM được huấn luyện.

Không vượt qua: LLM không thật sự “hiểu”, thiếu lý luận sâu và trải nghiệm thực tế.

* **Điều này có ý nghĩa gì với AGI (trí tuệ nhân tạo tổng quát) hay chỉ là narrow AI (AI hẹp, chuyên biệt)?**

LLMs là Narrow AI: giỏi xử lý ngôn ngữ, không hiểu biết rộng hay ý thức tự chủ. Để đạt lên AGI thì cần tổng quát, đa nhiệm, suy luận như con người và thích ứng linh hoạt.

**Đánh giá hiệu suất LLMs?**

****

LLM leaderboard cùng với các tiêu chí đánh giá

**Lập luận “Căn phòng Trung Quốc” (Chinese Room Argument – John Searle, 1980)**

Minh họa: Một người trong phòng chỉ nhận ký hiệu (input) và tra cứu quy tắc (rule ledger) để trả về ký hiệu khác (output).

**Ý nghĩa:**

Người trong phòng không thực sự “hiểu” tiếng Trung, chỉ đơn thuần thao tác theo luật.

Tương tự, LLM có thể giả lập trí tuệ (tạo ra câu trả lời mạch lạc) mà không thực sự hiểu.

**Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (Sắc lệnh hành pháp của Tổng thống Mỹ về phát triển và sử dụng AI an toàn, đáng tin cậy).**

**Nội dung:**

* **AI phải an toàn và bảo mật.**
* **Khuyến khích đổi mới, cạnh tranh và hợp tác có trách nhiệm.**
* **Bảo vệ quyền riêng tư và quyền dân sự của người dân Mỹ.**

**Câu hỏi được đặt ra:**

1. **Có nên quản lý (regulate) việc sử dụng LLMs không?**

* Có. Vì nếu không quản lý, LLM có thể bị lạm dụng để tạo tin giả, vi phạm bản quyền, lan truyền thông tin sai lệch, hoặc gây rủi ro cho quyền riêng tư và việc làm.

1. **Làm thế nào để quản lý?**

* Thiết lập tiêu chuẩn an toàn và minh bạch cho các mô hình.
* Yêu cầu gắn nhãn nội dung AI-generated (nội dung do AI tạo ra).
* Giới hạn hoặc kiểm soát việc truy cập mô hình mạnh (đặc biệt là mô hình open-source có thể bị lợi dụng).
* Cơ chế kiểm định độc lập để đánh giá rủi ro trước khi phát hành rộng rãi.
* Hợp tác quốc tế để có chuẩn chung, tránh lỗ hổng pháp lý.

1. **Còn vấn đề bản quyền (copyright) thì sao?**

* Dữ liệu huấn luyện LLM thường lấy từ Internet, dễ chứa nội dung có bản quyền → cần cơ chế trả phí/đền bù cho tác giả khi tác phẩm bị sử dụng.
* Nội dung AI tạo ra có thể khó xác định tác giả thật sự → cần khung pháp lý rõ ràng: quyền sở hữu thuộc người dùng, công ty AI, hay không ai cả?
* Cần phát triển công cụ truy vết dữ liệu huấn luyện để minh bạch hơn.