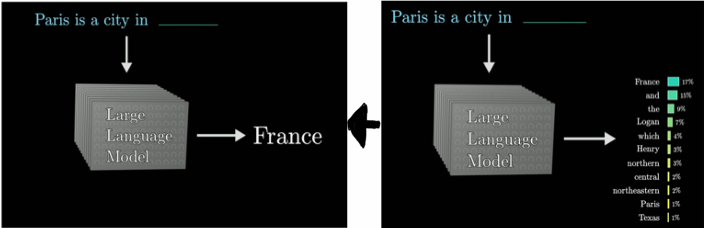
# How do LLMs reason and what are the limits?

*Cách LLMs lý luận và những giới hạn của nó?*

“A large language model is a sophisticated mathematical function that predict what word come next for any pieces of text.” – 3blue1brown.

Dịch ra tiếng Việt có nghĩa là “*Một mô hình ngôn ngữ lớn (LLMs) là một hàm toán học phức tạp dự đoán từ nào sẽ xuất hiện tiếp theo cho bất kỳ đoạn văn bản nào*.”

Hình 1. Ví dụ biểu diễn cách chọn lọc kết quả của mô hình ngôn ngữ lớn

Ở hình 1 ta có thể thấy một ví dụ đơn giản nhất về cách mà LLMs lý luận, đầu tiên nó nhận input là các văn bản đầu vào và mục tiêu của mô hình là làm sao lấy được từ tiếp theo (hoặc cuối cùng) sao cho tổng thể văn bản hợp lý nhất. Input được đưa vào mô hình ngôn ngữ lớn, trong mô hình này là vô vàn các dữ liệu đã được huấn luyện từ trước, bao gồm một tệp các đoạn văn, câu trả lời sẵn có… Từ đó thông qua nhiều tính toán để cho ra một danh sách các từ có khả năng phù hợp với văn bản đầu vào, như minh họa hình ở bên phải.

Có thể nói, LLMs không suy luận và cho câu trả lời từ kinh nghiệm thực tế, mà thực ra là một loạt các phép tính toán liên quan đến xác suất thống kê.

Đào sâu vào cách LLMs hoạt động, có 2 bước chính là Pre-training (Tiền huấn luyện) và sau đó là RLHF. Pre-training là huấn luyện một mô hình trên một tập hợp văn bản khổng lồ để học cấu trúc ngôn ngữ, ngữ pháp và kiến thức thực tế.

Và sau khi tiền huấn luyện, bước 2 là RLHF, được viết đầy đủ là Reinforcement Learning from Human Feedback. Đây là bước chính để tinh chỉnh các mô hình ngôn ngữ lớn, bởi vì nhờ xác suất thống kê, mô hình ngôn ngữ lớn có thể cho ra các câu trả lời hợp lý(trong phạm vi chúng được học), nhưng có thể nó sẽ sai lạc với ý đồ hoặc sự hài lòng của người dùng về câu trả lời. Ví dụ như khi người dùng hỏi một câu mang tính nhạy cảm hoặc cần câu trả lời có chiều sâu về ngữ cảnh, mô hình có thể đưa ra một đáp án “hợp lý về mặt ngữ pháp” nhưng lại thiếu tinh tế, thậm chí gây phản cảm. RLHF sẽ giúp mô hình học từ phản hồi của con người để biết rằng trong những tình huống đó, cần ưu tiên sự an toàn, tính hữu ích và sự phù hợp với mong đợi của người dùng hơn là chỉ dừng lại ở việc "có câu trả lời". Nhờ vậy, thay vì trả lời theo kiểu máy móc, mô hình dần trở nên tự nhiên, thân thiện và bám sát nhu cầu thực tế hơn.

Vậy giới hạn của LLMs là gì? Tôi thấy có 2 giới hạn chính của LLMs là dữ liệu, kỹ thuật huấn luyện và phần cứng. Tất cả câu trả lời đến từ các dữ liệu được dùng để huấn luyện, cho nên khi dữ liệu xấu tức là không có cấu trúc, thiên kiến, thì câu trả lời mà chúng ta nhận được sẽ không chính xác. Các câu trả lời mà LLMs cho ra đều dựa trên xác suất thống kê chứ không hề đến từ kinh nghiệm thực tế nên không thể mang tính tin cậy cao. Và nếu như gặp các vấn đề chưa từng được huấn luyện, lúc này LLMs sẽ có hiện tượng hallucinations, hay còn được gọi là ảo giác, mọi người hay đùa vui là nói dốc nói láo. Để có thể huấn luyện một mô hình ngôn ngữ lớn cần hàng nghìn GPU chuyên dụng (thường là A100 hoặc H100 của NVIDIA) chạy song song trong một cụm siêu máy tính, cùng với thời gian huấn luyện kéo dài từ vài tuần đến vài tháng, tùy vào quy mô dữ liệu và kiến trúc mô hình. Lấy ví dụ như GPT-3.0 (ra mắt năm 2020) với 175 tỷ tham số: nhóm OpenAI đã phải sử dụng tới hơn 10.000 GPU V100 và tiêu tốn hàng triệu đô la chi phí điện toán, kéo dài hàng tuần mới hoàn tất. Điều này cho thấy việc huấn luyện từ đầu một LLM là cực kỳ tốn kém, và cũng lý giải vì sao phần lớn các công ty khởi nghiệp thường chọn hướng tinh chỉnh (fine-tune) trên nền mô hình có sẵn thay vì huấn luyện lại toàn bộ từ đầu.

# How do we make sure that LLMs generate factually correct output?

*Làm sao để biết LLMs tạo sinh ra output đúng?*

Như đã đề cập ở câu hỏi đầu tiên, việc câu trả lời được tạo ra từ xác suất thống kê nên câu trả lời có khả năng sẽ gặp tình trạng hallucinations. Để ngăn chặn việc này xảy ra, tôi có bốn giải pháp chính. Thứ nhất, dùng RLHF để mô hình học tăng cường có sự phản hồi, giám sát từ con người. Thứ hai, huấn luyện với dữ liệu được chuẩn bị tốt (có cấu trúc rõ ràng, sạch sẽ), nên mua từ các công ty có tiếng về dữ liệu để có độ tin cậy cao, huấn luyện liên tục để mô hình được học kiến thức mới và loại bỏ các kiến thức lỗi thời. Thứ ba, nên giới hạn lĩnh vực sử dụng của mô hình, ví dụ như trong môi trường y tế sẽ có các dữ liệu được chọn lọc riêng chỉ ngành y, các cách tinh chỉnh hay câu trả lời mẫu cũng nằm trong lĩnh vực y, từ đó giúp mô hình nhẹ hơn nhưng vẫn cho người dùng một câu trả lời hợp lý. Và cuối cùng cũng là điều quan trọng nhất, luôn luôn tự kiểm tra câu trả lời nhận được, DYOR (Do Your Own Research), câu trả lời không đúng hoàn toàn nhưng người dùng nên dựa vào đó để tự kiểm chứng các thông tin mà họ nhận được để từ đó có thể quyết định xem có nên tin hay không tin vào câu trả lời mà mô hình đưa ra.

# How do we fairly compensate the people who create the data that is used to train LLMs?

*Làm sao để trả quyền lợi một cách công bằng cho các tác giả của những dữ liệu được dùng để huấn luyện mô hình?*

Trước hết, ta cần hiểu rõ rằng “tác giả” ở đây có thể là bất kỳ ai: một nhà văn, một nhà nghiên cứu, một người truyền cảm hứng, một nhà toán học, một lập trình viên, hay thậm chí là một cá nhân bình thường chia sẻ nội dung trên mạng xã hội, bởi dữ liệu dùng để huấn luyện mô hình ngôn ngữ lớn là vô cùng rộng lớn. Những bài viết, những đoạn code, hay các sản phẩm sáng tạo của họ có thể bị lấy đi để đưa vào tập huấn luyện mà chính tác giả còn không hề hay biết. Ví dụ như mô hình ngôn ngữ lớn Grok đã sử dụng dữ liệu từ các tweet (bài viết) của người dùng trên ứng dụng X và hệ sinh thái xung quanh nó. Hay gần đây, vào tháng 4 năm 2025, đội ngũ phát triển ChatGPT đã ra mắt một mô hình có khả năng “nhái lại” phong cách hội họa Ghibli – nét vẽ nổi tiếng gắn liền với đạo diễn Miyazaki Hayao – và điều này đã trở thành “giọt nước tràn ly”, thổi bùng lên làn sóng tranh cãi gay gắt về AI và tác quyền. Chính vì thế, tác giả của những dữ liệu được sử dụng để huấn luyện AI cần được bảo vệ quyền lợi, từ việc chia sẻ lợi nhuận, ghi credit để tôn trọng bản quyền, cho đến cơ chế thuê hoặc xin cấp phép chính thức đối với các tác phẩm đã đăng ký. Chỉ khi đảm bảo sự công bằng này, AI mới có thể phát triển bền vững mà không chà đạp lên giá trị sáng tạo của con người.

# How do we use LLMs in learning, so human learning is not compromised?

*Chúng ta nên sử dụng các mô hình ngôn ngữ lớn (LLMs) trong học tập như thế nào để không làm ảnh hưởng hay suy giảm quá trình học tập của con người?*

Ta nên sử dụng LLMs như công cụ hỗ trợ để mở rộng hiểu biết, giải thích vấn đề, và khơi gợi ý tưởng, chứ không nên dùng để làm thay hoặc học thụ động. Hãy chủ động tư duy, phản biện và kiểm tra lại thông tin, cũng như tận dụng LLMs để tạo môi trường học tập sáng tạo, không để quá trình học của con người bị ảnh hưởng tiêu cực