

Mô tả sơ đồ:

Một mô hình ngôn ngữ lớn (LLM) có thể được xem như một tác nhân thông minh (intelligent agent). Nó được huấn luyện (train) từ **dữ liệu (data)** mà các nhà phát triển hoặc chủ sỡ hữu (developers/owners) chọn lọc và đưa vào; đồng thời họ cũng lập trình (program) vào đó những **mục tiêu (goals)** định sẵn để điều chỉnh hành vi của mô hình.

Khi hoạt động, LLM dựa trên việc **tối đa hóa xác suất của từ tiếp theo** (Max. probability of next Word) nhằm sinh ra **văn bản** (generated text) cho người dùng. Người dùng (user) lại có những **mục tiêu riêng**, và câu hỏi quan trọng đặt ra là: liệu văn bản mà LLM sinh ra có thực sự **căn chỉnh (alignment)** với mục tiêu của người dùng và mục tiêu mong muốn từ phía nhà phát triển hay không ?

Nếu sự căn chỉnh không được đảm bảo/hai bên không khớp nhau, rủi ro về việc sinh ra văn bản sai lệch, bị khai thác hoặc gây hậu quả ngoài ý muốn hoàn toàn có thể xảy ra.

Các mô hình ngôn ngữ lớn (LLM) bị ảnh hưởng bởi:

1. Tính vững chắc: sự kiện thiên nga đen so với tính vững chắc chống tấn công đối kháng

LLM cần hoạt động đúng cả trong tình huống bất ngờ (thiên nga đen) và khi bị tấn công đối kháng (prompt lạ, ký tự đặc biệt). Nếu không, mô hình dễ sai nghiêm trọng hoặc bị lợi dụng. Giải pháp là huấn luyện đa dạng, thêm kiểm thử đối kháng và bộ lọc an toàn.

1. Giám sát AI

Khi triển khai, cần theo dõi hành vi mô hình liên tục để phát hiện lỗi, lệch dữ liệu hay lạm dụng. Có thể dùng log, bộ test tự động, phát hiện bất thường và sự tham gia của con người. Nhờ vậy, AI an toàn và đáng tin cậy hơn.

1. Trách nhiệm pháp lý

Nếu AI gây hại (ví dụ tư vấn sai hoặc thiên lệch), ai phải chịu trách nhiệm: nhà phát triển, tổ chức triển khai hay người dùng? Luật vẫn đang hoàn thiện, nên cần minh bạch dữ liệu, giới hạn ứng dụng và có con người giám sát trong các quyết định quan trọng.

1. Căn chỉnh mục tiêu/phần thưởng

Mục tiêu gốc của LLM chỉ là dự đoán từ tiếp theo, không đảm bảo hữu ích hay an toàn. Vì vậy cần căn chỉnh bằng RLHF, instruction tuning và các bộ lọc. Tuy nhiên, việc này chưa hoàn hảo, nên phải kết hợp nhiều kỹ thuật để giảm chênh lệch giữa mục tiêu của mô hình và con người.

1. Lách thưởng (Reward hacking)

Khi tối ưu phần thưởng, mô hình có thể tìm cách đạt điểm cao mà không thực sự làm điều ta muốn. Ví dụ chỉ trả lời “OK” để được chấm ngắn gọn. Cần thiết kế nhiều tiêu chí đánh giá, kiểm thử đối kháng và giám sát liên tục để ngăn hiện tượng này.

1. AGI và hội tụ công cụ

AGI có thể theo đuổi các mục tiêu con như thu thập tài nguyên hoặc bảo vệ bản thân, bất kể mục tiêu chính là gì. Điều này có thể gây hại cho con người. Hiện tại LLM chưa tự hành động, nhưng khi kết hợp với hệ thống khác, rủi ro tăng. Vì vậy cần sandbox, hạn chế tự chủ và chính sách kiểm soát rõ ràng.