

Meta phát hành bộ công cụ nguồn mở bảo đảm an toàn AI

10:43 10/12/2023

Dự án Purple Llama của Meta nhằm mục đích tạo ra bộ công cụ nguồn mở giúp các nhà phát triển xây dựng các mô hình AI tạo sinh một cách có trách nhiệm.



Ảnh minh họa (Ảnh: internet)

Meta đã giới thiệu Purple Llama, một dự án dành riêng cho việc tạo ra các công cụ nguồn mở để các nhà phát triển đánh giá và nâng cao độ tin cậy và an toàn của các mô hình trí tuệ nhân tạo (AI) tạo sinh trước khi chúng được đưa vào sử dụng.

Meta nhấn mạnh sự cần thiết của những nỗ lực hợp tác trong việc đảm bảo an toàn cho AI, đồng thời cho rằng các thách thức về AI không thể được giải quyết một cách biệt lập.

Công ty cho biết mục tiêu của Purple Llama là thiết lập một nền tảng chung để phát triển AI tạo sinh an toàn hơn khi mối lo ngại ngày càng tăng về các mô hình ngôn ngữ lớn và các công nghệ AI khác.

Gareth Lindahl-Wise, Giám đốc An ninh thông tin tại Công ty an ninh mạng Ontinue, gọi Purple Llama là “một bước đi tích cực và chủ động” hướng tới công nghệ AI an toàn hơn.

Dự án Purple Llama hợp tác với các nhà phát triển AI; dịch vụ đám mây như AWS và Google Cloud; các công ty bán dẫn như Intel, AMD và Nvidia; và các công ty phần mềm, gồm cả Microsoft.

Bộ công cụ đầu tiên được phát hành thông qua Purple Llama là CyberSecEval - công cụ đánh giá rủi ro an ninh mạng trong phần mềm do AI tạo ra.

Nó có mô hình ngôn ngữ xác định văn bản không phù hợp hoặc có hại, bao gồm các cuộc thảo luận về bạo lực hoặc hoạt vi bất hợp pháp.

Các nhà phát triển có thể sử dụng CyberSecEval để kiểm tra mô hình AI của họ có xu hướng tạo mã không an toàn hoặc hỗ trợ các cuộc tấn công mạng hay không.

Nghiên cứu của Meta đã phát hiện ra rằng các mô hình ngôn ngữ lớn thường đề xuất mã dễ bị tấn công, nêu bật tầm quan trọng của việc thử nghiệm và cải tiến liên tục đối với bảo mật AI.

Llama Guard là một bộ công cụ khác - một mô hình ngôn ngữ lớn được đào tạo để xác định ngôn ngữ có khả năng gây hại hoặc xúc phạm.

Các nhà phát triển có thể sử dụng Llama Guard để kiểm tra xem mô hình của họ có tạo ra hoặc chấp nhận nội dung không an toàn hay không, giúp lọc ra các lời nhắc có thể dẫn đến kết quả đầu ra không phù hợp.

Hà Trần (t/h)