

Tin tặc có thể sử dụng AI tạo sinh để thao túng các cuộc trò chuyện trực tiếp

10:00 09/04/2024

Các nhà nghiên cứu bảo mật của IBM đã chứng minh rằng, tin tặc có thể sử dụng trí tuệ nhân tạo (AI) tạo sinh và công nghệ deepfake âm thanh để chiếm quyền điều khiển và thao túng các cuộc hội thoại trực tiếp.

Tin tặc có thể sử dụng AI tạo sinh để thao túng các cuộc trò chuyện trực tiếp

Các nhà nghiên cứu đã dễ dàng sử dụng kỹ thuật chỉnh âm thanh để chặn âm thanh của người nói và thay thế các đoạn giọng nói thật bằng một đoạn âm thanh deepfake. Họ cho biết: “Thay vì sử dụng AI tạo sinh để tạo giọng nói giả cho toàn bộ cuộc trò chuyện, một việc tương đối dễ bị phát hiện, chúng tôi đã phát hiện ra cách chặn cuộc trò chuyện trực tiếp và thay thế từ khóa dựa trên ngữ cảnh”.

Tất cả những gì các nhà nghiên cứu cần để sao chép giọng nói là ba giây âm thanh.

Các nhà nghiên cứu đã đưa ra một mô hình ngôn ngữ lớn (LLM) xử lý âm thanh từ hai nguồn trong một cuộc trò chuyện trực tiếp trên điện thoại và chú ý các từ khóa và cụm từ cụ thể - trong trường hợp này là cụm từ "tài khoản ngân hàng". Khi mô hình phát hiện cụm từ này, nó sẽ thay thế số tài khoản ngân hàng thật bằng số tài khoản giả.

LLM đóng vai trò là người ở giữa, theo dõi cuộc trò chuyện trực tiếp. Các nhà nghiên cứu đã sử dụng tính năng chuyển giọng nói thành văn bản để giúp mô hình hiểu được ngữ cảnh của cuộc trò chuyện. “Nó giống như việc biến những người trong cuộc trò chuyện thành những con rối giả, và do ngữ cảnh ban đầu được giữ nguyên nên rất khó phát hiện”, họ nói.

Mối đe dọa không chỉ ở việc thao túng tài chính, trong đó tin tặc có thể lừa nạn nhân gửi hàng tỷ USD vào tài khoản của chúng. Kỹ thuật này có thể được sử dụng để kiểm duyệt thông tin, hướng dẫn phi công sửa đổi đường bay và thay đổi nội dung trong các chương trình phát sóng tin tức trực tiếp và các bài phát biểu chính trị trong thời gian thực.

Các nhà nghiên cứu cho biết, việc phát triển hệ thống AI để thực hiện nhiệm vụ đặt ra rất ít thách thức, ngay cả khi việc thực hiện một cuộc tấn công yêu cầu tội phạm phải có kỹ năng lừa đảo và kỹ thuật xã hội. Tuy nhiên, việc sử dụng kỹ thuật chỉnh âm thanh một cách dễ dàng cũng cần phần lớn thời gian để tìm ra cách thu âm thanh từ micrô và cung cấp âm thanh cho AI tổng hợp.

Các nhà nghiên cứu đã gặp phải một số rào cản ảnh hưởng đến tính thuyết phục của cuộc tấn công.

Thứ nhất, giọng nói nhân bản cần phải tính đến âm điệu và tốc độ để hòa vào cuộc trò chuyện thực.

Thứ hai, độ trễ trong GPU gây ra sự chậm trễ trong cuộc trò chuyện, do cần truy cập LLM và API chuyển văn bản thành giọng nói từ xa. Tuy nhiên, các nhà nghiên cứu đã giải quyết vấn đề bằng cách xây dựng các khoảng dừng nhân tạo. Họ đã huấn luyện mô hình cách sử dụng các cụm từ cầu nối để lấp đầy bất kỳ khoảng trống nào được tạo ra bằng cách yêu cầu mô hình xử lý thao tác.

Các nhà nghiên cứu cho biết: “Trong lúc PoC đang kích hoạt khi nghe từ khóa 'tài khoản ngân hàng' và lấy tài khoản ngân hàng khác để chèn vào cuộc trò chuyện, độ trễ được che giấu bởi các cụm từ bắc cầu như ‘Tất nhiên, đợi tôi một giây’.”

Họ cũng nói thêm rằng tin tặc sẽ cần có sẵn một năng lực tính toán “đáng kể” để có thể thực hiện và mở rộng những cuộc tấn công này.

T. Hương (Nguồn: <https://antoanthongtin.vn/>)