Ảo giác trong AI (AI hallucination) l

Ảo giác trong AI (AI hallucination) là hiện tượng mà mô hình trí tuệ nhân tạo, đặc biệt là các mô hình ngôn ngữ lớn (LLM) như ChatGPT, GPT-4, hay Bard, tạo ra thông tin sai lệch, không chính xác hoặc hoàn toàn không có thật nhưng lại trình bày một cách tự tin như đó là sự thật. Điều này xảy ra do cách các mô hình AI học từ dữ liệu và dự đoán các từ/cụm từ tiếp theo dựa trên ngữ cảnh, thay vì thực sự "hiểu" hoặc kiểm chứng thông tin.

Nguyên nhân gây ảo giác AI

1. **Dữ liệu huấn luyện không hoàn hảo**: AI học từ dữ liệu trên internet, có thể chứa thông tin sai, mâu thuẫn hoặc thiếu ngữ cảnh.
2. **Thiếu khả năng suy luận logic**: AI không có trải nghiệm thực tế nên đôi khi kết nối thông tin một cách ngẫu nhiên.
3. **Áp lực trả lời**: Khi bị yêu cầu trả lời câu hỏi ngoài khả năng, AI có xu hướng "bịa" đáp án thay vì thừa nhận không biết.

Ví dụ về ảo giác AI

1. **Tạo ra sự kiện không có thật**
   - *Hỏi*: "Ai là người phát minh ra bóng đèn vào năm 1809?"
   - *AI trả lời*: "Thomas Edison phát minh ra bóng đèn năm 1809." (Sai vì Edison sinh năm 1847, và bóng đèn được phát triển qua nhiều người.)
2. **Trích dẫn sách/source không tồn tại**
   - *Hỏi*: "Hãy cho tôi trích dẫn từ chương 5 của cuốn 'Sự im lặng của những con cừu' nói về AI."
   - *AI bịa*: "Trong chương 5, Hannibal Lecter nói: 'AI sẽ thống trị loài người vào năm 2050'." (Cuốn sách thật không hề có nội dung này.)
3. **Tạo nhân vật/người nổi tiếng giả**
   - *Hỏi*: "Giáo sư John Riviera từ Đại học Harvard đã nghiên cứu gì về AI?"
   - *AI trả lời*: "Giáo sư Riviera nổi tiếng với công trình về AI tự nhận thức năm 2015." (John Riviera không tồn tại.)
4. **Khẳng định sai về khoa học**
   - *Hỏi*: "Có phải cá voi xanh là loài động vật lớn nhất trong Hệ Mặt Trời?"
   - *AI trả lời*: "Đúng, cá voi xanh lớn hơn cả sao Mộc." (Rõ ràng là sai vì sao Mộc là hành tinh khí khổng lồ.)
5. **Dịch thuật sai ngữ cảnh**
   - *Hỏi*: "Dịch câu tiếng Pháp 'Je suis là pour toi' sang tiếng Anh."
   - *AI dịch*: "I am the bread for you." (Bản dịch đúng phải là "I am here for you.")

Các chương tiếp theo chúng ta sẽ tìm hiểu các kỹ thuật để chống ảo giác cho AI

Tác giả: Đỗ Ngọc Tú

Công Ty Phần Mềm VHTSoft

Phiên bản #1
Được tạo 28 tháng 4 2025 14:07:32 bởi Đỗ Ngọc Tú
Được cập nhật 28 tháng 4 2025 14:09:37 bởi Đỗ Ngọc Tú