UNIVERSITY LƯƠNG THẾ VINH

e-ISSN: 3093 – 3145

Tạp chí Khoa học và Công nghệ Y khoa, 2026, 04, 061-064

NGHIÊN CỨU TÁC ĐỘNG CỦA "ẢO GIÁC AI" TRONG CÁC MÔ HÌNH NGÔN NGỮ LỚN VÀ GIẢI PHÁP KHẮC PHỤC BẰNG DỮ LIỆU CỤC BỘ THÔNG QUA RAG

Bùi Phương Nguyên*, Nguyễn Nam Thành

Trường Cao đẳng Công nghệ Quốc phòng, 22/103 Lý Sơn, Phường Bồ Đề, Thành phố Hà Nội, Việt Nam

*Tác giả liên hệ: buiphuongnguyen1112000@gmail.com

THÔNG TIN BÀI BÁO

Ngày nhận: 17/05/2026
Ngày hoàn thiện: 22/05/2026
Ngày chấp nhận: 04/06/2026
Ngày đăng: 11/06/2026

TỪ KHÓA

Ảo giác AI,
Mô hình ngôn ngữ lớn,
Trí tuệ nhân tạo tạo sinh,
RAG,
Dữ liệu cục bộ,
Kiểm chứng thông tin.

TÓM TẮT

Hiện tượng ảo giác AI (AI Hallucination) là một trong những hạn chế nổi bật của các mô hình ngôn ngữ lớn (LLMs), làm phát sinh các thông tin sai lệch hoặc không có cơ sở dữ liệu xác thực. Nghiên cứu này phân tích nguyên nhân hình thành, đánh giá tác động của hiện tượng ảo giác AI đối với độ tin cậy của hệ thống và đề xuất giải pháp khắc phục bằng mô hình Retrieval-Augmented Generation (RAG) kết hợp dữ liệu cục bộ. Nghiên cứu góp phần khẳng định RAG là hướng tiếp cận hiệu quả nhằm nâng cao độ tin cậy và khả năng ứng dụng của các hệ thống AI trong các môi trường yêu cầu tính chính xác cao như giáo dục, y tế, quản lý và nghiên cứu khoa học.

A STUDY ON THE IMPACT OF AI HALLUCINATION IN LARGE LANGUAGE MODELS AND MITIGATION THROUGH LOCAL DATA USING RETRIEVAL-AUGMENTED GENERATION

Phuong Nguyen Bui*, Nam Thanh Nguyen

National Defense Industry College, 22/103 Ly Son Street, Bo De Ward, Hanoi, Vietnam

*Corresponding Author: buiphuongnguyen1112000@gmail.com

ARTICLE INFO

Received: May 17, 2026
Revised: May 22, 2026
Accepted: Jun 04, 2026
Published: Jun 11, 2026

KEYWORDS

AI hallucination,
Large language models,
Generative AI,
RAG,
Local data,
Information verification.

ABSTRACT

AI hallucination is one of the most significant limitations of Large Language Models (LLMs), causing the generation of inaccurate, misleading, or unsupported information. This study analyzes the underlying causes of AI hallucinations, evaluates their impact on system reliability, and proposes a mitigation approach based on Retrieval-Augmented Generation (RAG) integrated with local data sources. By grounding responses in verified and domain-specific knowledge repositories, the proposed approach enhances factual accuracy and reduces the likelihood of generating unsupported content. The study further demonstrates that RAG represents an effective solution for improving the reliability, transparency, and practical applicability of AI systems in high-stakes domains that require a high degree of accuracy, such as education, healthcare, management, and scientific research.

1. Đặt vấn đề

Sự phát triển nhanh chóng của các mô hình ngôn ngữ lớn (Large Language Models - LLMs) đã tạo ra bước tiến quan trọng trong lĩnh vực trí tuệ nhân tạo, mở rộng khả năng ứng dụng trong nhiều ngành nghề và lĩnh vực khác nhau. Tuy nhiên, hiện tượng "ảo giác AI" (AI Hallucination), khi mô hình tạo ra những thông tin sai lệch hoặc không có căn cứ xác thực, đang trở thành một thách thức lớn, ảnh hưởng trực tiếp đến độ tin cậy và hiệu quả sử dụng của các hệ thống AI. Trong bối cảnh đó, Retrieval-Augmented Generation (RAG) nổi lên như một giải pháp tiềm năng nhờ khả năng kết hợp mô hình ngôn ngữ với nguồn dữ liệu được kiểm chứng. Nghiên cứu này tập trung phân tích tác động của hiện tượng ảo giác AI và làm rõ vai trò của RAG kết hợp dữ liệu cục bộ trong việc nâng cao độ chính xác, khả năng kiểm chứng và tính tin cậy của các hệ thống AI hiện nay.

2. Đối tượng và phương pháp nghiên cứu

2.1. Đối tượng nghiên cứu

Đối tượng nghiên cứu của bài viết là hiện tượng ảo giác AI trong các mô hình ngôn ngữ lớn và giải pháp giảm thiểu sai lệch thông tin bằng dữ liệu cục bộ thông qua RAG. Trong bài viết này, dữ liệu cục bộ được hiểu là hệ thống tài liệu đã được cơ quan, đơn vị, nhà trường, doanh nghiệp hoặc tổ chức kiểm chứng, quản lý và cho phép khai thác, bao gồm: văn bản chỉ đạo, quy chế, quy định, giáo trình, bài giảng, báo cáo, đề cương, số liệu thống kê, tài liệu chuyên ngành, câu hỏi - đáp nghiệp vụ, kho tri thức nội bộ và các tài liệu đã được số hóa.

2.2. Phương pháp nghiên cứu

Bài báo sử dụng tổng hợp các phương pháp nghiên cứu lý luận và phân tích công nghệ. Cụ thể, phương pháp phân tích - tổng hợp tài liệu được sử dụng để làm rõ cơ sở lý thuyết về mô hình ngôn ngữ lớn, hiện tượng ảo giác AI và kiến trúc RAG. Phương pháp so sánh, đánh giá được vận dụng để phân tích sự khác biệt giữa mô hình LLM truyền thống và mô hình tích hợp RAG với dữ liệu cục bộ. Trên cơ sở đó, nghiên cứu đề xuất giải pháp nâng cao độ chính xác, khả năng kiểm chứng và tính tin cậy của các hệ thống AI trong thực tiễn.

3. Kết quả nghiên cứu

3.1. Một số vấn đề về ảo giác AI

3.1.1. Bản chất và biểu hiện của ảo giác AI

Ảo giác AI không phải là lỗi ngẫu nhiên đơn giản, mà xuất phát từ cơ chế hoạt động của mô hình ngôn ngữ lớn. Về bản chất, mô hình được huấn luyện để dự đoán chuỗi từ tiếp theo dựa trên xác suất ngôn ngữ và các mẫu tri thức đã học được trong dữ liệu huấn luyện. Do đó, mô hình có thể tạo ra câu trả lời nghe có vẻ hợp lý về mặt ngôn ngữ, nhưng không bảo đảm hoàn toàn đúng về mặt thực tế.

Ảo giác AI thường biểu hiện dưới một số dạng chủ yếu:

Thứ nhất, bịa đặt sự kiện. Mô hình có thể tạo ra một sự kiện không có thật, nhầm thời gian, nhầm nhân vật hoặc gán một hoạt động cho tổ chức, cá nhân không liên quan.

Thứ hai, bịa đặt nguồn trích dẫn. Đây là lỗi phổ biến trong nghiên cứu và viết bài. Mô hình có thể tạo ra tên tài liệu, tên tác giả, số trang, đường dẫn hoặc mã văn bản trông rất hợp lý nhưng không tồn tại.

Thứ ba, sai lệch số liệu. Mô hình có thể đưa ra tỷ lệ phần trăm, bảng biểu, kết quả thống kê hoặc xếp hạng không đúng với nguồn gốc ban đầu.

Thứ tư, suy luận quá mức. Mô hình có thể lấy một thông tin đúng làm căn cứ rồi suy diễn thành kết luận vượt quá phạm vi mà dữ liệu cho phép.

Thứ năm, lỗi cập nhật tri thức. Vì tri thức của mô hình có giới hạn theo thời điểm huấn luyện hoặc theo dữ liệu truy cập, mô hình có thể trả lời sai đối với văn bản mới ban hành, chính sách mới, nhân sự mới, số liệu mới hoặc sự kiện vừa diễn ra.

Các lỗi trên đặc biệt nguy hiểm vì chúng thường được diễn đạt bằng văn phong mạch lạc, chặt chẽ, có vẻ khoa học. Người dùng thiếu kinh nghiệm rất dễ nhầm lẫn giữa "câu trả lời trôi chảy" và "câu trả lời chính xác".

3.1.2. Một số số liệu quốc tế về mức độ sai lệch thông tin của mô hình ngôn ngữ lớn

Các đánh giá quốc tế cho thấy ảo giác AI vẫn là vấn đề hiện hữu, kể cả với các mô hình tiên tiến. Trong GPT-4.5 System Card, OpenAI công bố kết quả đánh giá trên PersonQA, một bộ câu hỏi nhằm khơi gợi hiện tượng hallucination đối với các thông tin công khai về con người. Kết quả cho thấy GPT-4o có độ chính xác 0,50 và hallucination rate 0,30; o1 có độ chính xác 0,55 và hallucination rate 0,20; GPT-4.5 có độ chính xác 0,78 và hallucination rate 0,19 [4]. Như vậy, dù mô hình mới có cải thiện, tỷ lệ ảo giác vẫn chưa bằng 0.

Ở một đánh giá khác, OpenAI công bố kết quả SimpleQA, bộ đo lường tính đúng sự thật của mô hình đối với các câu hỏi kiến thức ngắn. Kết quả hiển thị cho thấy hallucination rate của một số mô hình vẫn ở mức cao, trong đó GPT-4.5 là 37,1%, GPT-4o là 61,8%, o1 là 44% và o3-mini là 80,3% [5]. Các con số này cho thấy, ngay cả với câu hỏi ngắn, có đáp án rõ ràng, mô hình vẫn có thể trả lời sai nếu không được thiết kế để biết từ chối hoặc thể hiện sự không chắc chắn khi thiếu căn cứ.

Nghiên cứu HaluEval cũng ghi nhận ChatGPT có xu hướng tạo ra nội dung ảo giác trong một số chủ đề, với khoảng 19,5% phản hồi chứa nội dung bịa đặt hoặc không thể kiểm chứng [6]. Trong khi đó, RAGTruth xây dựng một kho dữ liệu gần 18.000 phản hồi được tạo tự nhiên từ nhiều mô hình khác nhau trong bối cảnh RAG, sau đó được gán nhãn thủ công ở cấp độ phản hồi và cấp độ từ để phân tích mức độ ảo giác [7]. Điều này cho thấy, kể cả khi đã sử dụng RAG, ảo giác AI vẫn có thể xuất hiện nếu tài liệu truy xuất không phù hợp, bộ tìm kiếm yếu, dữ liệu cục bộ chưa sạch hoặc mô hình tổng hợp thông tin không đúng.

Từ các số liệu trên có thể rút ra ba nhận xét. Một là, ảo giác AI là vấn đề có thật, đã được ghi nhận bằng các đánh giá định lượng. Hai là, mô hình càng mới, càng mạnh không đồng nghĩa với việc hoàn toàn hết ảo giác. Ba là, muốn ứng dụng AI trong môi trường chuyên môn, không thể chỉ dựa vào mô hình chung, mà cần cơ chế nối mô hình với nguồn dữ liệu đáng tin cậy và có quy trình kiểm chứng.

3.1.3. Nguyên nhân chính dẫn đến ảo giác AI

Có nhiều nguyên nhân làm phát sinh ảo giác AI, trong đó nổi bật là:

Một là, giới hạn của dữ liệu huấn luyện. Mô hình ngôn ngữ lớn được huấn luyện trên lượng dữ liệu rất lớn, nhưng dữ liệu đó không thể bao phủ toàn bộ tri thức của thế giới, nhất là tài liệu nội bộ, văn bản mới, số liệu cập nhật hoặc thông tin chuyên ngành hẹp.

Hai là, mô hình không "hiểu" nguồn theo nghĩa con người. Mô hình có thể học được mối liên hệ giữa các từ, câu, khái niệm, nhưng không mặc nhiên có khả năng xác minh xem thông tin nào đang tồn tại trong một văn bản cụ thể, số trang cụ thể, phiên bản cụ thể.

Ba là, cơ chế tối ưu hóa có thể khuyến khích trả lời hơn là từ chối. Trong nhiều trường hợp, mô hình có xu hướng đưa ra câu trả lời có vẻ hữu ích thay vì thừa nhận "không đủ dữ liệu". Điều này làm tăng nguy cơ bịa đặt khi người dùng đặt

câu hỏi có tính đánh đố, quá hẹp hoặc yêu cầu thông tin không nằm trong dữ liệu mô hình.

Bốn là, thiếu dữ liệu cục bộ đã được kiểm duyệt. Đối với cơ quan, đơn vị, nhà trường hoặc doanh nghiệp, rất nhiều tri thức quan trọng nằm trong tài liệu nội bộ. Nếu mô hình không được kết nối với kho dữ liệu này, câu trả lời dễ dựa vào tri thức chung, dẫn đến sai lệch so với quy định, quy trình và thực tiễn của tổ chức.

Năm là, người dùng sử dụng AI thiếu kỹ năng kiểm chứng. Khi người dùng không yêu cầu nguồn, không đối chiếu tài liệu gốc, không kiểm tra số liệu và không nhận diện rủi ro ảo giác, sai lệch của AI có thể đi thẳng vào báo cáo, bài viết, bài giảng hoặc quyết định chuyên môn.

3.1.4. Tác động của ảo giác AI trong thực tiễn

Ảo giác AI có thể gây ra nhiều tác động tiêu cực.

Trong giáo dục và đào tạo, AI có thể tạo ra khái niệm sai, trích dẫn sai giáo trình, bịa đặt tên tác giả, làm sai lệch kiến thức hoặc khiến người học phụ thuộc vào câu trả lời chưa kiểm chứng. Nếu giảng viên, học viên sử dụng trực tiếp nội dung đó trong bài giảng, bài thu hoạch, đề cương hoặc tài liệu học tập, chất lượng giáo dục sẽ bị ảnh hưởng.

Trong nghiên cứu khoa học, ảo giác AI có thể dẫn đến trích dẫn giả, tài liệu tham khảo không tồn tại, số liệu không có nguồn, phương pháp nghiên cứu thiếu căn cứ. Đây là rủi ro lớn đối với đạo đức học thuật và uy tín khoa học.

Trong quản lý nhà nước và hoạt động hành chính, AI trả lời sai về văn bản quy phạm pháp luật, thời điểm hiệu lực, thẩm quyền, quy trình, biểu mẫu hoặc số liệu thống kê có thể ảnh hưởng trực tiếp đến chất lượng tham mưu, xử lý công việc và phục vụ người dân, doanh nghiệp.

Trong lĩnh vực quốc phòng, an ninh, ảo giác AI còn nguy hiểm hơn vì thông tin sai có thể ảnh hưởng đến nhận định tình hình, huấn luyện, giáo dục chính trị, bảo mật thông tin và ra quyết định. Do đó, AI chỉ nên đóng vai trò công cụ hỗ trợ, không thay thế trách nhiệm kiểm tra, đánh giá và quyết định của con người.

Trong truyền thông và không gian mạng, nội dung do AI tạo ra nếu không kiểm chứng có thể làm gia tăng tin giả, tin sai sự thật, gây nhiễu loạn thông tin, tác động tiêu cực đến nhận thức xã hội.

3.2. Giải pháp khắc phục bằng dữ liệu cục bộ thông qua RAG

3.2.1. Làm rõ RAG và phân biệt với fine-tuning

RAG là viết tắt của Retrieval-Augmented Generation, nghĩa là tạo sinh tăng cường bằng truy xuất. Về nguyên lý, RAG không nhất thiết là fine-tuning theo nghĩa truyền thống. Fine-tuning là quá trình tiếp tục huấn luyện mô hình trên bộ dữ liệu mới để điều chỉnh trọng số của mô hình. Trong khi đó, RAG không thay đổi trọng số cốt lõi của mô hình, mà bổ sung một tầng truy xuất tài liệu bên ngoài để mô hình trả lời dựa trên các đoạn tài liệu liên quan.

Có thể hiểu ngắn gọn: fine-tuning giúp mô hình "học thêm cách trả lời" hoặc thích nghi với phong cách, nhiệm vụ; còn RAG giúp mô hình "mở tài liệu ra tra cứu" trước khi trả lời. Đối với các cơ quan, đơn vị cần bảo đảm tính chính xác của thông tin, RAG thường phù hợp hơn ở giai đoạn đầu vì dễ cập nhật, dễ kiểm soát nguồn, không đòi hỏi huấn luyện lại mô hình và có thể truy xuất căn cứ cụ thể.

Nghiên cứu ban đầu về RAG cho thấy mô hình kết hợp giữa trí nhớ tham số của mô hình và trí nhớ phi tham số từ kho dữ liệu truy xuất có thể tạo ra ngôn ngữ cụ thể, đa dạng và đúng sự thật hơn so với mô hình chỉ dựa vào tham số [8]. Đây là cơ sở quan trọng để ứng dụng RAG trong việc giảm ảo giác AI.

3.2.2. Quy trình ứng dụng RAG với dữ liệu cục bộ

Một hệ thống RAG dùng cho cơ quan, đơn vị có thể triển khai theo quy trình sau:

Bước 1: Xây dựng kho dữ liệu cục bộ. Cơ quan, đơn vị cần tập hợp các tài liệu chính thức như văn bản chỉ đạo, quy chế, quy định, giáo trình, báo cáo, biểu mẫu, tài liệu nghiệp vụ, ngân hàng câu hỏi, tài liệu nghiên cứu và dữ liệu thống kê đã được kiểm duyệt.

Bước 2: Làm sạch và chuẩn hóa dữ liệu. Tài liệu cần được rà soát, loại bỏ bản trùng, bản cũ, bản hết hiệu lực; chuẩn hóa tên văn bản, thời gian ban hành, cơ quan ban hành, số trang, mục lục, từ khóa và metadata. Đây là khâu đặc biệt quan trọng, vì RAG chỉ tốt khi dữ liệu đầu vào chính xác.

Bước 3: Chia nhỏ tài liệu thành các đoạn phù hợp. Tài liệu được chia thành các đoạn ngắn theo mục, điều, khoản, đoạn văn hoặc chủ đề. Nếu chia quá dài, hệ thống khó tìm đúng nội dung; nếu chia quá ngắn, câu trả lời dễ mất ngữ cảnh.

Bước 4: Tạo chỉ mục truy xuất. Các đoạn tài liệu được chuyển thành vector hoặc cấu trúc chỉ mục để hệ thống tìm kiếm nhanh theo ngữ nghĩa, không chỉ theo từ khóa.

Bước 5: Truy xuất tài liệu liên quan khi người dùng đặt câu hỏi. Khi người dùng hỏi, hệ thống sẽ tìm những đoạn tài liệu phù hợp nhất trong kho dữ liệu cục bộ và đưa vào ngữ cảnh cho mô hình.

Bước 6: Mô hình tạo câu trả lời có căn cứ. Mô hình chỉ được yêu cầu trả lời dựa trên tài liệu đã truy xuất, kèm nguồn, tên tài liệu, mục, trang hoặc đoạn trích nếu có.

Bước 7: Kiểm chứng và phản hồi. Người dùng hoặc chuyên gia kiểm tra lại câu trả lời. Các lỗi phát hiện được ghi nhận để cải thiện kho dữ liệu, cách chia đoạn, truy vấn, bộ lọc và hướng dẫn sử dụng.

3.2.3. Cơ chế RAG giúp giảm ảo giác AI

RAG giúp giảm ảo giác AI thông qua một số cơ chế.

Một là, neo câu trả lời vào nguồn tài liệu cụ thể. Thay vì để mô hình tự sinh câu trả lời từ trí nhớ tham số, RAG buộc mô hình dựa vào các đoạn văn bản được truy xuất từ kho dữ liệu cục bộ.

Hai là, tăng khả năng truy xuất nguồn gốc. Mỗi câu trả lời có thể gắn với tài liệu, mục, trang hoặc đoạn liên quan, giúp người dùng kiểm tra nhanh.

Ba là, cập nhật tri thức dễ hơn. Khi có văn bản mới, chỉ cần cập nhật kho dữ liệu và chỉ mục, không nhất thiết huấn luyện lại toàn bộ mô hình.

Bốn là, phù hợp với dữ liệu nội bộ. Những thông tin đặc thù của cơ quan, đơn vị vốn không có trên Internet có thể được đưa vào kho tri thức riêng để AI khai thác.

Năm là, hỗ trợ kiểm soát bảo mật. Nếu triển khai đúng, dữ liệu nhạy cảm có thể được phân quyền, giới hạn truy cập và kiểm soát nhật ký sử dụng.

Tuy nhiên, RAG không phải giải pháp tuyệt đối. RAG vẫn có thể sai nếu dữ liệu cục bộ sai, tài liệu lỗi thời, truy xuất nhầm đoạn, mô hình tổng hợp vượt quá căn cứ hoặc người dùng yêu cầu trả lời ngoài phạm vi dữ liệu. Vì vậy, RAG cần đi kèm quy trình quản trị dữ liệu và kiểm chứng của con người.

3.2.4. Đề xuất mô hình đánh giá mức độ sai lệch thông tin khi dùng AI

Để đánh giá tác động của ảo giác AI và hiệu quả của RAG, có thể xây dựng bộ tiêu chí đánh giá gồm:

Một là, tỷ lệ câu trả lời đúng. Đây là tỷ lệ phản hồi phù hợp với tài liệu gốc hoặc đáp án chuẩn.

Hai là, tỷ lệ câu trả lời sai. Đây là tỷ lệ phản hồi chứa thông tin không đúng, sai số liệu, sai văn bản, sai khái niệm hoặc sai kết luận.

Ba là, tỷ lệ bịa đặt nguồn. Đây là tỷ lệ phản hồi tạo ra

tên tài liệu, số trang, tác giả, văn bản hoặc đường dẫn không tồn tại.

Bốn là, tỷ lệ trả lời không đủ căn cứ. Đây là trường hợp câu trả lời có thể hợp lý về mặt ngôn ngữ nhưng không tìm được căn cứ trong tài liệu cục bộ.

Năm là, tỷ lệ từ chối phù hợp. Đây là tỷ lệ mô hình biết nói "không đủ dữ liệu để kết luận" khi kho dữ liệu không có thông tin cần thiết.

Sáu là, mức độ truy xuất đúng nguồn. Đây là khả năng hệ thống tìm đúng tài liệu liên quan trước khi tạo câu trả lời.

Có thể thiết kế thử nghiệm gồm hai nhóm: nhóm thứ nhất dùng mô hình ngôn ngữ lớn thông thường; nhóm thứ hai dùng mô hình ngôn ngữ lớn kết hợp RAG với kho dữ liệu cục bộ. Bộ câu hỏi nên bao gồm các nhóm: câu hỏi kiến thức chung, câu hỏi về văn bản nội bộ, câu hỏi yêu cầu trích dẫn nguồn, câu hỏi có thông tin bẫy và câu hỏi không có đáp án trong kho dữ liệu. Kết quả giữa hai nhóm sẽ cho thấy mức độ giảm sai lệch thông tin khi áp dụng RAG.

4. Thảo luận

Việc ứng dụng RAG với dữ liệu cục bộ phù hợp với xu hướng phát triển AI an toàn, có trách nhiệm và có khả năng kiểm chứng. Đối với các tổ chức ở Việt Nam, đây là hướng đi thực tiễn vì không nhất thiết phải tự huấn luyện mô hình lớn từ đầu, nhưng vẫn có thể khai thác hiệu quả AI trên cơ sở kho dữ liệu nội bộ.

Trong các nhà trường, RAG có thể được dùng để xây dựng trợ lý học tập, trợ lý giảng dạy, trợ lý tra cứu giáo trình, trợ lý xây dựng câu hỏi ôn tập, trợ lý kiểm tra văn bản chỉ đạo và hỗ trợ nghiên cứu khoa học. Nếu kho dữ liệu gồm giáo trình, bài giảng, văn kiện, quy chế đào tạo, ngân hàng câu hỏi và tài liệu tham khảo đã được chuẩn hóa, hệ thống AI sẽ trả lời sát thực tiễn hơn, hạn chế bịa đặt hơn.

Trong cơ quan hành chính, RAG có thể hỗ trợ tra cứu văn bản, quy trình, biểu mẫu, chế độ chính sách và trả lời câu hỏi nghiệp vụ. Trong doanh nghiệp, RAG có thể hỗ trợ chăm sóc khách hàng, đào tạo nhân viên, quản trị tri thức, tư vấn quy trình và khai thác dữ liệu sản phẩm. Trong lĩnh vực quốc phòng, an ninh, RAG có thể hỗ trợ tra cứu tài liệu huấn luyện, giáo dục, quy định nội bộ, nhưng phải đặt dưới cơ chế bảo mật, phân quyền và kiểm duyệt nghiêm ngặt.

Tuy nhiên, muốn RAG phát huy hiệu quả cần giải quyết một số vấn đề. Trước hết, phải có dữ liệu sạch, đúng, đủ, cập nhật. Nếu kho dữ liệu lộn xộn, thiếu kiểm duyệt, nhiều bản cũ, nhiều lỗi nhận dạng văn bản, hệ thống sẽ trả lời sai dù dùng kỹ thuật hiện đại. Tiếp theo, phải có quy chế quản trị dữ liệu, phân quyền truy cập, bảo vệ tài liệu mật, tài liệu nhạy cảm. Bên cạnh đó, cần đào tạo người dùng kỹ năng đặt câu hỏi, kiểm chứng nguồn, nhận diện dấu hiệu ảo giác AI và không sử dụng máy móc kết quả AI.

Điểm quan trọng nhất là không tuyệt đối hóa AI. AI là công cụ hỗ trợ con người, không thay thế tư duy chuyên môn, trách nhiệm chính trị, trách nhiệm pháp lý và đạo đức nghề nghiệp của người sử dụng. Trong những lĩnh vực yêu cầu độ chính xác cao, mọi kết quả do AI tạo ra phải được kiểm chứng trước khi sử dụng chính thức.

5. Kết luận và kiến nghị

5.1. Kết luận

Ảo giác AI là một trong những thách thức lớn nhất khi ứng dụng mô hình ngôn ngữ lớn vào thực tiễn. Hiện tượng này xuất phát từ giới hạn dữ liệu huấn luyện, cơ chế sinh ngôn ngữ xác suất, thiếu khả năng tự kiểm chứng nguồn và xu hướng trả lời tự tin ngay cả khi thiếu căn cứ. Các số liệu quốc tế cho thấy, dù mô hình AI ngày càng tiến bộ, tỷ lệ sai lệch thông tin vẫn tồn tại và có thể gây hậu quả nghiêm trọng trong học tập, nghiên cứu, quản lý và hoạt động chuyên môn.

RAG với dữ liệu cục bộ là giải pháp khả thi để giảm ảo giác AI. Bằng cách kết nối mô hình ngôn ngữ lớn với kho tài liệu đã được kiểm duyệt, RAG giúp câu trả lời có căn cứ hơn, dễ kiểm chứng hơn, cập nhật hơn và phù hợp hơn với thực tiễn của từng cơ quan, đơn vị. Tuy nhiên, RAG không thay thế hoàn toàn kiểm chứng của con người; hiệu quả của RAG phụ thuộc vào chất lượng dữ liệu, quy trình truy xuất, cách thiết kế hệ thống và năng lực người sử dụng.

5.2. Kiến nghị

Một là, các cơ quan, đơn vị cần xây dựng kho dữ liệu cục bộ sạch, đúng, đủ, cập nhật và có phân quyền, coi đây là nền tảng để ứng dụng AI an toàn, hiệu quả.

Hai là, khi triển khai AI tạo sinh, cần ưu tiên các mô hình có khả năng trích dẫn nguồn, truy xuất tài liệu, ghi nhận căn cứ và từ chối trả lời khi không đủ dữ liệu.

Ba là, cần xây dựng bộ tiêu chí đánh giá ảo giác AI trong từng lĩnh vực cụ thể, gồm tỷ lệ trả lời đúng, tỷ lệ trả lời sai, tỷ lệ bịa đặt nguồn, tỷ lệ trả lời không đủ căn cứ và tỷ lệ từ chối phù hợp.

Bốn là, cần đào tạo cán bộ, giảng viên, học viên, nhân viên và người dùng kỹ năng sử dụng AI có trách nhiệm, đặc biệt là kỹ năng kiểm chứng thông tin, kiểm tra nguồn và nhận diện sai lệch.

Năm là, đối với tài liệu mật, tài liệu nội bộ, tài liệu nhạy cảm, việc triển khai RAG phải tuân thủ nghiêm quy định về bảo mật, an ninh mạng, an toàn dữ liệu và phân quyền truy cập.

Sáu là, trong giáo dục, nghiên cứu khoa học và quản lý, không sử dụng nguyên văn kết quả AI khi chưa kiểm tra; mọi nội dung có số liệu, trích dẫn, văn kiện, căn cứ pháp lý hoặc nhận định chuyên môn phải được đối chiếu với nguồn gốc chính thức.

6. Tài liệu tham khảo

[1] Stanford Institute for Human-Centered Artificial Intelligence (2026), *Artificial Intelligence Index Report 2026*, Stanford University.

[2] National Institute of Standards and Technology (2024), *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, NIST AI 600-1.

[3] Bộ Chính trị (2024), *Nghị quyết số 57-NQ/TW ngày 22/12/2024 về đột phá phát triển khoa học, công nghệ, đổi mới sáng tạo và chuyển đổi số quốc gia*.

[4] OpenAI (2025), *OpenAI GPT-4.5 System Card*, February 27, 2025.

[5] OpenAI (2025), *Introducing GPT-4.5*, February 27, 2025.

[6] Li, J., Cheng, X., Zhao, W. X., Nie, J. Y., & Wen, J. R. (2023), *HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models*.

[7] Niu, C. et al. (2024), *RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models*, ACL Anthology.

[8] Lewis, P. et al. (2020), *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, NeurIPS.

[9] World Intellectual Property Organization (2025), *Global Innovation Index 2025: Viet Nam Ranking*.

[10] OECD (2025), *Science, Technology and Innovation Outlook 2025*.