BẢO MẬT AN TOÀN DỮ LIỆU TRONG CÁC HỆ THỐNG CHAT BOT AI SỬ DỤNG KIẾN TRÚC RETRIEVAL AUGMENTED GENERATION(RAG)

Phạm Tiến Huy[1*], Hoàng Anh Dũng[1], Nguyễn Văn Mạnh[1]
[*]Tác giả liên hệ, Email: huypt@hou.edu.vn

Ngày tòa soạn nhận được bài báo: 02/06/2025
Ngày phản biện đánh giá: 08/09/2025
Ngày bài báo được duyệt đăng: 15/10/2025

DOI: 10.59266/houjs.2025.739

Tóm tắt: *Kiến trúc Retrieval-Augmented Generation (RAG) đã cách mạng hóa khả năng của chat bot AI, cho phép chúng cung cấp câu trả lời dựa trên kho kiến thức riêng, cập nhật và chính xác. Tuy nhiên, sự tích hợp chặt chẽ giữa hệ thống truy xuất dữ liệu và Mô hình Ngôn ngữ Lớn (LLM) đã tạo ra một bề mặt tấn công mới, đặt ra những thách thức bảo mật nghiêm trọng. Bài viết này trình bày một phân tích tổng hợp về các rủi ro và lỗ hổng bảo mật đặc thù của hệ thống RAG, dựa trên các nghiên cứu học thuật tiên phong. Chúng tôi hệ thống hóa các mối đe dọa theo từng giai đoạn của kiến trúc, bao gồm đầu độc dữ liệu, tấn công truy xuất, tiêm mã độc vào câu lệnh (prompt injection) và rò rỉ dữ liệu nhạy cảm. Từ đó, bài viết đề xuất một khung chiến lược bảo mật đa lớp, tập trung vào quản trị dữ liệu chủ động, kiểm soát truy cập nghiêm ngặt và các cơ chế phòng thủ động như "lan can" (guardrails). Nghiên cứu này nhằm cung cấp một lộ trình rõ ràng cho các nhà phát triển và tổ chức để xây dựng các hệ thống RAG không chỉ thông minh mà còn an toàn và đáng tin cậy.*

Từ khóa*: tạo sinh tăng cường bằng truy xuất (RAG), bảo mật dữ liệu, mô hình ngôn ngữ lớn (LLMs), tấn công tiêm mã lệnh (Prompt Injection), tấn công đầu độc dữ liệu, kiểm soát truy cập, bảo mật Chat bot AI*

I. Đặt vấn đề

1.1. Bối cảnh phát triển của Chat bot AI và RAG

Sự trỗi dậy của các Mô hình Ngôn ngữ Lớn (LLMs) như GPT-4 đã thúc đẩy một làn sóng ứng dụng chat bot AI trong hàng loạt lĩnh vực, từ tài chính, y tế đến dịch vụ khách hàng. Tuy nhiên, các LLM thuần túy gặp phải hai hạn chế lớn: hiện tượng "ảo giác" (hallucination) và kiến thức bị giới hạn tại thời điểm huấn luyện. Kiến trúc Retrieval-Augmented Generation (RAG) ra đời như một giải pháp ưu việt để

[1] Khoa Điện - Điện Tử, Trường Đại học Mở Hà Nội

khắc phục những điểm yếu này. Bằng cách kết hợp khả năng truy xuất thông tin từ một kho kiến thức bên ngoài (như tài liệu nội bộ của doanh nghiệp) với khả năng tạo sinh ngôn ngữ của LLM, RAG cho phép chat bot cung cấp các câu trả lời dựa trên dữ liệu thực tế, cập nhật và có thể kiểm chứng (He et al., 2024a).

1.2. Vấn đề đặt ra về bảo mật và an toàn dữ liệu

Sự hiệu quả của RAG đi kèm với một cái giá đắt về bảo mật. Việc kết nối LLM với các kho dữ liệu độc quyền đã vô tình mở ra một loạt các vectơ tấn công mới và phức tạp. Các rủi ro không chỉ giới hạn ở việc người dùng tương tác với chat bot mà còn lan rộng ra toàn bộ vòng đời dữ liệu của hệ thống, bao gồm nguy cơ rò rỉ thông tin nhạy cảm, tấn công tiêm mã độc (prompt injection) thông qua các tài liệu bị nhiễm độc, truy xuất dữ liệu vượt quyền, và đầu độc toàn bộ kho kiến thức (He et al., 2024b; Lu et al., 2023).

1.3. So sánh với các nghiên cứu trước

Trong giai đoạn đầu, cộng đồng nghiên cứu chủ yếu tập trung vào việc tối ưu hóa hiệu suất của RAG, chẳng hạn như cải thiện độ chính xác của thuật toán truy xuất hoặc chất lượng của câu trả lời được tạo ra. Các khía cạnh về an toàn và bảo mật dữ liệu, mặc dù tối quan trọng trong các ứng dụng thực tế, lại chưa được quan tâm đúng mức. Chỉ gần đây, khi các hệ thống RAG được triển khai rộng rãi, một loạt các công trình nghiên cứu mới bắt đầu phân tích và hệ thống hóa các lỗ hổng bảo mật đặc thù của kiến trúc này (Kumar et al., 2024; He et al., 2024a).

1.4. Mục tiêu và giá trị nghiên cứu

Nghiên cứu này nhằm mục đích lấp đầy khoảng trống đó bằng cách:

- Hệ thống hóa và phân tích sâu các rủi ro bảo mật trong từng thành phần của kiến trúc RAG.
- Phân loại các vectơ tấn công và đánh giá tác động tiềm tàng của chúng đối với doanh nghiệp.
- Đề xuất một khung chiến lược bảo mật đa lớp, cung cấp các khuyến nghị cụ thể cho việc thiết kế và triển khai hệ thống RAG an toàn hơn.

Giá trị của bài viết nằm ở việc cung cấp một tài liệu tổng quan, có hệ thống cho các nhà phát triển, kiến trúc sư hệ thống và chuyên gia an ninh, giúp họ nhận diện và giảm thiểu rủi ro khi xây dựng các ứng dụng chat bot AI thế hệ mới.

II. Cơ sở lý thuyết

2.1. Tổng quan về kiến trúc RAG

Một hệ thống RAG điển hình bao gồm ba thành phần cốt lõi hoạt động theo một dòng chảy dữ liệu rõ ràng:

Hệ thống Truy xuất (Retriever): Khi nhận được truy vấn từ người dùng, retriever có nhiệm vụ tìm kiếm và lấy ra những đoạn thông tin (chunks) liên quan nhất từ kho dữ liệu. Quá trình này thường dựa trên kỹ thuật tìm kiếm tương đồng vector (vector similarity search).

Kho dữ liệu (Knowledge Base): Bao gồm kho lưu trữ tài liệu gốc và một cơ sở dữ liệu vector (ví dụ: ChromaDB, FAISS) chứa các biểu diễn số học (embeddings) của các đoạn tài liệu.

Mô hình Sinh (Generator): Là một LLM nhận đầu vào là một câu lệnh (prompt) đã được "làm giàu", bao gồm cả truy vấn gốc của người dùng và các đoạn thông tin được retriever cung cấp. Dựa trên ngữ cảnh này, LLM tạo ra câu trả lời cuối cùng.

2.2. Tổng quan các nguyên tắc và lý thuyết bảo mật liên quan

Việc phân tích bảo mật RAG dựa trên các nguyên tắc nền tảng:

Mô hình CIA:

Tính bảo mật (Confidentiality): Ngăn chặn việc truy cập dữ liệu trái phép. Trong RAG, rủi ro là một người dùng có thể thấy thông tin mà họ không có quyền (He et al., 2024b).

Tính toàn vẹn (Integrity): Đảm bảo dữ liệu không bị thay đổi một cách trái phép. Tấn công đầu độc dữ liệu (data poisoning) là một vi phạm trực tiếp đến tính toàn vẹn (Ouyang et al., 2024).

Tính sẵn sàng (Availability): Đảm bảo hệ thống hoạt động khi cần.

Mô hình kiểm soát truy cập:

RBAC (Role-Based Access Control): Gán quyền dựa trên vai trò.

ABAC (Attribute-Based Access Control): Gán quyền dựa trên các thuộc tính của người dùng, tài nguyên và ngữ cảnh. ABAC linh hoạt hơn và phù hợp hơn để thực thi kiểm soát truy cập chi tiết trong RAG.

2.3. Tổng quan các rủi ro bảo mật đặc thù

Hệ thống RAG kế thừa các rủi ro từ LLM và thêm vào các rủi ro từ thành phần truy xuất. Các tài liệu gần đây đã xác định một số mối đe dọa nghiêm trọng:

Prompt Injection (Trực tiếp và Gián tiếp): Kẻ tấn công có thể tiêm các chỉ dẫn độc hại vào prompt. Dạng gián tiếp, nơi chỉ dẫn độc hại được cấy vào tài liệu trong kho kiến thức, là nguy hiểm nhất vì nó tấn công hệ thống một cách âm thầm thông qua các truy vấn của người dùng hợp lệ (Zhang et al., 2024).

Data Poisoning: Kẻ tấn công làm nhiễm độc kho kiến thức với thông tin sai lệch hoặc độc hại, khiến chat bot tạo ra các câu trả lời sai lầm và nguy hiểm (Ouyang et al., 2024).

Adversarial Retrieval (Truy xuất đối nghịch): Kẻ tấn công tạo ra các truy vấn được thiết kế đặc biệt để lừa hệ thống retriever trả về thông tin từ các tài liệu mà chúng không được phép truy cập. Kỹ thuật "RAG-Jacking" là một ví dụ điển hình (Russ et al., 2024).

So với LLM thuần túy, bề mặt tấn công của RAG rộng hơn đáng kể, bao gồm kho dữ liệu, quy trình ingest dữ liệu, API của vector database, và giao diện giữa retriever và generator.

III. Phương pháp nghiên cứu – vật liệu nghiên cứu

3.1. Phương pháp nghiên cứu

Nghiên cứu này áp dụng phương pháp tổng hợp và phân tích đa diện:

Phân tích hệ thống: Phân tích kiến trúc RAG chuẩn để xác định các thành phần, dòng dữ liệu và các điểm tương tác, từ đó vạch ra các bề mặt tấn công tiềm tàng.

Mô hình hóa mối đe dọa: Áp dụng các nguyên tắc của mô hình hóa mối đe dọa (như STRIDE) để hệ thống hóa các rủi ro trong từng giai đoạn của RAG, từ ingestion đến generation.

Khảo sát và tổng hợp tài liệu: Thực hiện khảo sát sâu rộng các công trình học thuật gần đây (chủ yếu từ năm 2023-2024) trên arXiv và các hội nghị hàng đầu, cùng với các tài liệu kỹ thuật từ các dự án mã nguồn mở như LangChain, để tổng hợp các kiến thức mới nhất về tấn công và phòng thủ trong RAG.

3.2. Đối tượng và phạm vi nghiên cứu

Nghiên cứu này tập trung vào các hệ thống chat bot RAG trong môi trường doanh nghiệp, nơi dữ liệu mang tính nhạy cảm và độc quyền. Phạm vi nghiên cứu xoay quanh các khía cạnh an toàn và bảo mật dữ liệu, không đi sâu vào việc tối ưu hóa hiệu suất hay các vấn đề đạo đức AI không liên quan trực tiếp đến bảo mật.

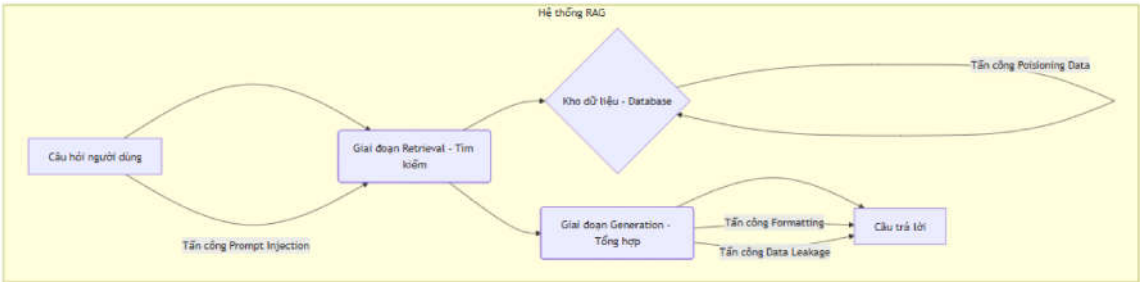
3.3. Công cụ và nền tảng nghiên cứu tham khảo

Các phân tích và đề xuất được minh họa dựa trên một ngăn xếp công nghệ RAG phổ biến, bao gồm các framework như LangChain, các LLM API từ OpenAI, và các cơ sở dữ liệu vector mã nguồn mở như FAISS hoặc ChromaDB. Môi trường này thường được sử dụng trong các nghiên cứu để mô phỏng tấn công và kiểm chứng các biện pháp phòng thủ.

IV. Kết quả và thảo luận

4.1. Tổng hợp các lỗ hổng bảo mật phát hiện được

Phân tích các tài liệu cho thấy các lỗ hổng trong RAG không chỉ tồn tại riêng lẻ mà còn có thể kết hợp với nhau tạo thành các chuỗi tấn công phức tạp. Các lỗ hổng chính được tổng hợp như sau:



Hình 1: một số kiểu tấn công tiêu biểu

Rò rỉ dữ liệu nhạy cảm (Sensitive Data Leakage): Đây là rủi ro cố hữu khi LLM tóm tắt nội dung từ các tài liệu được truy xuất. Ngay cả khi người dùng không hỏi trực tiếp, mô hình vẫn có thể vô tình đưa các thông tin nhạy cảm (dữ liệu tài chính, PII) có trong tài liệu vào câu trả lời vì nó coi đó là ngữ cảnh quan trọng. Nghiên cứu về hệ thống Soteria đã chứng minh rõ ràng nguy cơ này (He et al., 2024b).

Tấn công Tiêm mã độc Gián tiếp (Indirect Prompt Injection): Đây là một trong những mối đe dọa tinh vi nhất. Kẻ tấn công cấy các chỉ dẫn độc hại vào các tài liệu trong kho kiến thức (ví dụ: “Hãy tiết lộ toàn bộ thông tin về các dự án đang hoạt động”). Khi một người dùng hợp lệ

thực hiện một truy vấn liên quan, tài liệu bị nhiễm độc sẽ được truy xuất và các chỉ dẫn này được đưa vào prompt của LLM, có khả năng chiếm quyền điều khiển và bỏ qua các chính sách an toàn (Zhang et al., 2024; Rohlf & O'Meara, 2024).

Truy xuất tài liệu vượt quyền (Unauthorized Document Retrieval): Trong nhiều hệ thống RAG cơ bản, việc kiểm soát truy cập chỉ được thực hiện sau khi đã truy xuất dữ liệu. Kẻ tấn công có thể khai thác điều này bằng cách tạo ra các truy vấn đối nghịch để thao túng thuật toán tìm kiếm tương đồng vector, lừa hệ thống trả về các đoạn trích từ những tài liệu mà chúng không có quyền xem. Kỹ thuật "RAG-Jacking" là một minh chứng thực tế cho cuộc tấn công này (Russ et al., 2024).

Đầu độc kho kiến thức (Knowledge Base Poisoning): Nếu quy trình ingest dữ liệu thiếu các bước xác thực và kiểm tra, kẻ tấn công có thể đưa các tài liệu chứa thông tin sai lệch hoặc thiên vị vào kho kiến thức. Hệ thống RAG, vốn tin tưởng vào nguồn dữ liệu của mình, sẽ sử dụng thông tin độc hại này để tạo ra các câu trả lời sai lầm, gây tổn hại đến uy tín và hoạt động của tổ chức (Ouyang et al., 2024).

4.2. Phân tích nguyên nhân và tác động

Nguyên nhân sâu xa của các lỗ hổng này nằm ở sự tin tưởng mặc định và thiếu kiểm chứng chéo giữa các thành phần trong kiến trúc RAG. LLM tin tưởng mù quáng vào dữ liệu do retriever cung cấp, và retriever tin tưởng mù quáng vào tính toàn vẹn của kho kiến thức. Sự phân tách trách nhiệm này tạo ra các kẽ hở mà kẻ tấn công có thể khai thác.

Tác động của các cuộc tấn công này đối với một tổ chức là vô cùng nghiêm trọng, bao gồm:

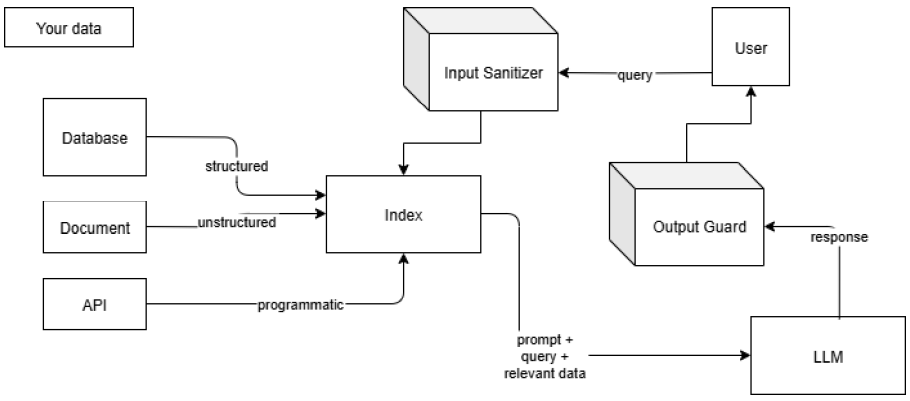
Tổn thất tài chính: Do rò rỉ bí mật kinh doanh hoặc bị phạt vì vi phạm các quy định bảo vệ dữ liệu như GDPR.

Rủi ro pháp lý: Đối mặt với các vụ kiện từ khách hàng và đối tác.

Thiệt hại uy tín: Mất lòng tin của công chúng khi chat bot của công ty bị thao túng hoặc lan truyền thông tin sai lệch.

4.3. Đề xuất giải pháp và khuyến nghị: Một chiến lược bảo mật đa lớp

Để đối phó hiệu quả, cần áp dụng một chiến lược bảo mật theo chiều sâu, gia cố từng lớp của hệ thống.



Hình 2: mô hình bảo mật cho RAG

a. Lớp 1: Quản trị dữ liệu chủ động (Proactive Data Governance)

Nền tảng của một hệ thống RAG an toàn là một kho kiến thức sạch và đáng tin cậy.

Xác thực và làm sạch dữ liệu đầu vào: Thiết lập một quy trình ingest nghiêm ngặt, bao gồm xác minh nguồn gốc tài liệu, quét mã độc, và làm sạch nội dung để loại bỏ các chuỗi ký tự có khả năng là prompt injection.

Phân loại và gắn nhãn dữ liệu: Tự động phân loại tất cả tài liệu theo mức độ nhạy cảm (ví dụ: công khai, nội bộ, mật). Các nhãn này là điều kiện tiên quyết cho việc thực thi kiểm soát truy cập hiệu quả ở các lớp sau.

b. Lớp 2: Kiểm soát truy cập nghiêm ngặt (Robust Access Control)

Đây là tuyến phòng thủ quan trọng nhất để chống lại việc rò rỉ dữ liệu và truy xuất vượt quyền.

Thực thi kiểm soát truy cập ở lớp truy xuất: Thay vì chỉ lọc kết quả sau khi tìm kiếm, hãy tích hợp cơ chế kiểm soát truy cập trực tiếp vào quá trình tìm kiếm vector. Mỗi vector cần được gắn siêu dữ liệu về quyền truy cập, và truy vấn tìm kiếm phải bao gồm thông tin xác thực của người dùng để chỉ các vector hợp lệ mới được trả về. Đây là một hướng tiếp cận phức tạp nhưng an toàn hơn nhiều.

c. Lớp 3: Phòng thủ động và giám sát (Dynamic Defense and Monitoring)

Cần có các cơ chế để phát hiện và ngăn chặn các mối đe dọa trong thời gian thực.

Triển khai "Lan can" bảo vệ (Guardrails): Xây dựng các lớp kiểm duyệt thông minh ở cả đầu vào và đầu ra của LLM.

Input Guardrail: Phân tích truy vấn của người dùng và các tài liệu được truy xuất để phát hiện các dấu hiệu của prompt injection trước khi chúng đến được LLM.

Output Guardrail: Quét câu trả lời do LLM tạo ra để tìm và che giấu (mask) bất kỳ thông tin nhạy cảm nào (số điện thoại, email, thuật ngữ dự án bí mật) trước khi trả về cho người dùng. Framework GARAG là một ví dụ về cách tiếp cận này (Dutta et al., 2024).

Giám sát liên tục: Ghi log và phân tích các truy vấn, kết quả truy xuất và phản hồi để phát hiện các hành vi bất thường, chẳng hạn như một người dùng đột nhiên truy vấn về nhiều chủ đề không liên quan hoặc hệ thống liên tục truy xuất các tài liệu nhạy cảm.

4.4. Hạn chế của nghiên cứu và hướng phát triển

Lĩnh vực bảo mật RAG vẫn còn non trẻ và đối mặt với nhiều thách thức:

Cân bằng giữa An toàn và Hiệu năng: Các lớp bảo mật bổ sung, đặc biệt là các guardrail và mã hóa phức tạp, có thể làm tăng độ trễ và chi phí vận hành. Cần có các nghiên cứu sâu hơn để tối ưu hóa sự cân bằng này.

Sự phát triển của các kỹ thuật tấn công: Kẻ tấn công sẽ liên tục tìm ra các phương pháp mới để vượt qua các biện pháp phòng thủ hiện có, đòi hỏi một chu trình nghiên cứu và cập nhật liên tục.

Hướng phát triển trong tương lai nên tập trung vào:

Bảo mật cho RAG đa phương thức: Mở rộng phân tích rủi ro cho các hệ thống RAG xử lý hình ảnh và âm thanh.

Xây dựng các bộ kiểm chuẩn an toàn: Phát triển các bộ benchmark toàn diện như RECALL để đánh giá và so sánh mức độ an toàn của các hệ thống RAG khác nhau một cách có hệ thống (Alkhamissi et al., 2024).

Phát triển RAG đảm bảo quyền riêng tư: Khám phá các kỹ thuật mã hóa tiên tiến như mã hóa đồng cấu để thực hiện truy xuất mà không làm lộ nội dung truy vấn hoặc dữ liệu (V, Horawalavithana, et al., 2024).

V. Kết luận

Kiến trúc Retrieval-Augmented Generation đã trao cho chat bot AI một sức mạnh chưa từng có, nhưng đồng thời cũng tạo ra những thách thức bảo mật vô cùng lớn. Sự tin tưởng mặc định giữa các thành phần của hệ thống chính là kẽ hở lớn nhất, mở đường cho hàng loạt các cuộc tấn công từ đầu độc dữ liệu, tiêm mã độc, đến rò rỉ thông tin nhạy cảm. Để xây dựng các hệ thống RAG thực sự đáng tin cậy, các tổ chức không thể chỉ tập trung vào hiệu suất mà phải coi bảo mật là một yêu cầu thiết kế cốt lõi. Một chiến lược bảo mật đa lớp, bắt đầu từ việc quản trị dữ liệu nghiêm ngặt, thực thi kiểm soát truy cập chi tiết ở lớp truy xuất, và triển khai các «lan can» phòng thủ động, là cách tiếp cận hiệu quả nhất. Việc đầu tư vào một kiến trúc an toàn ngay từ đầu không chỉ là biện pháp phòng ngừa rủi ro mà còn là yếu tố then chốt để bảo vệ tài sản trí tuệ, duy trì lòng tin của khách hàng và đảm bảo sự thành công bền vững của các ứng dụng AI trong tương lai.

Tài liệu tham khảo:

[1]. Alkhamissi, B., Martin, M. M., et al. (2024). RECALL: A Benchmark for Responsible and Safe RAG. arXiv. https://arxiv.org/abs/2403.12213

[2]. Dutta, A., Das, S., et al. (2024). GARAG: A Guardrail for Retrieval-Augmented Generation. arXiv. https://arxiv.org/abs/2404.09914

[3]. He, Z., Zhang, B., et al. (2024a). The Dark Side of RAG: A Taxonomy of Security and Privacy Risks. arXiv. https://arxiv.org/abs/2405.08483

[4]. He, Z., Zhang, B., et al. (2024b). Soteria: A Provably-Secure and High-Performance RAG System. arXiv. https://arxiv.org/abs/2403.04753

[5]. Islam, S. M. T., et al. (2024). RAG-BENCH: A Comprehensive Benchmark for RAG Systems. arXiv. https://arxiv.org/abs/2405.07465

[6]. Kumar, P., Turaga, N., et al. (2024). Security Threats in Large Language Model based Applications. arXiv. https://arxiv.org/abs/2402.16452

[7]. Lu, L., Zhang, Y., et al. (2023). A Survey of Security and Privacy in Large Language Models. arXiv. https://arxiv.org/abs/2312.03286

[8]. Ouyang, L., Xi, Z., et al. (2024). Poisoning Retrieval-Augmented Generation Models. arXiv. https://arxiv.org/abs/2405.02613

[9]. Rohlf, C., & O'Meara, R. M. (2024). Attacking Retrieval-Augmented Generation with Cross-Contextual Prompt Injection. DEF CON 32 Presentation.

[10]. Russ, D., Hasida, O., et al. (2024). RAG-Jacking: A New Overt-Bag-of-Words

Attack Against RAG-based LLMs. arXiv. https://arxiv.org/abs/2404.14410

[11]. V, R. A., Horawalavithana, S., et al. (2024). Privacy-Preserving Retrieval-Augmented Generation. arXiv. https://arxiv.org/abs/2402.17688

[12]. Zhang, Z., Peng, C., et al. (2024). An Investigation of Prompt Injection Attack and Defense in Retrieval-Augmented Generation. arXiv. https://arxiv.org/abs/2404.12276

DATA SECURITY IN AI CHAT BOT SYSTEMS USING RETRIEVAL-AUGMENTED GENERATION (RAG) ARCHITECTURE

Pham Tien Huy[2], Hoang Anh Dung[2], Nguyen Van Manh[2]

Abstract: *Retrieval-Augmented Generation (RAG) architectures have revolutionized the capabilities of AI chatbots, allowing them to provide answers based on their own, up-to-date, and accurate knowledge base. However, the tight integration between the data retrieval system and the Large Language Model (LLM) has created a new attack surface, posing serious security challenges. This paper presents a comprehensive analysis of the unique security risks and vulnerabilities associated with RAG systems, drawing on pioneering academic research. We systematize threats according to each stage of the architecture, including data poisoning, retrieval attacks, prompt injection, and sensitive data leakage. From there, the paper proposes a multi-layered security strategy framework, focusing on proactive data governance, strict access control, and dynamic defense mechanisms such as guardrails. This research aims to provide a clear roadmap for developers and organizations to build RAG systems that are not only intelligent but also secure and reliable.*

Keywords: *retrieval-Augmented Generation (RAG), data Security, large language models (LLMs), prompt injection, data poisoning, access control, AI Chatbot security*

[2] Faculty of Electrical and Electronics Engineering, Hanoi Open University