CyBe

CSA | Bangalore Chapter

Cyber. AI. Summit

Sheraton Grand, Whitefield, Bangalore | 4 Sep 2025

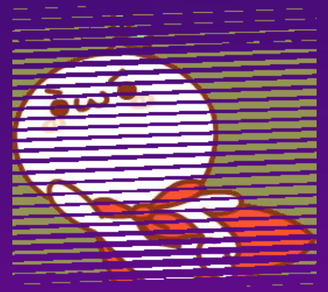
CyBe
A Global Initiative by
CSA | Bangalore Chapter

How Poisoned RAG System Impacts Real World AI

Anjali Shukla
Divyanshu Shukla

CyBe
A Global Initiative by
CSA | Bangalore Chapter

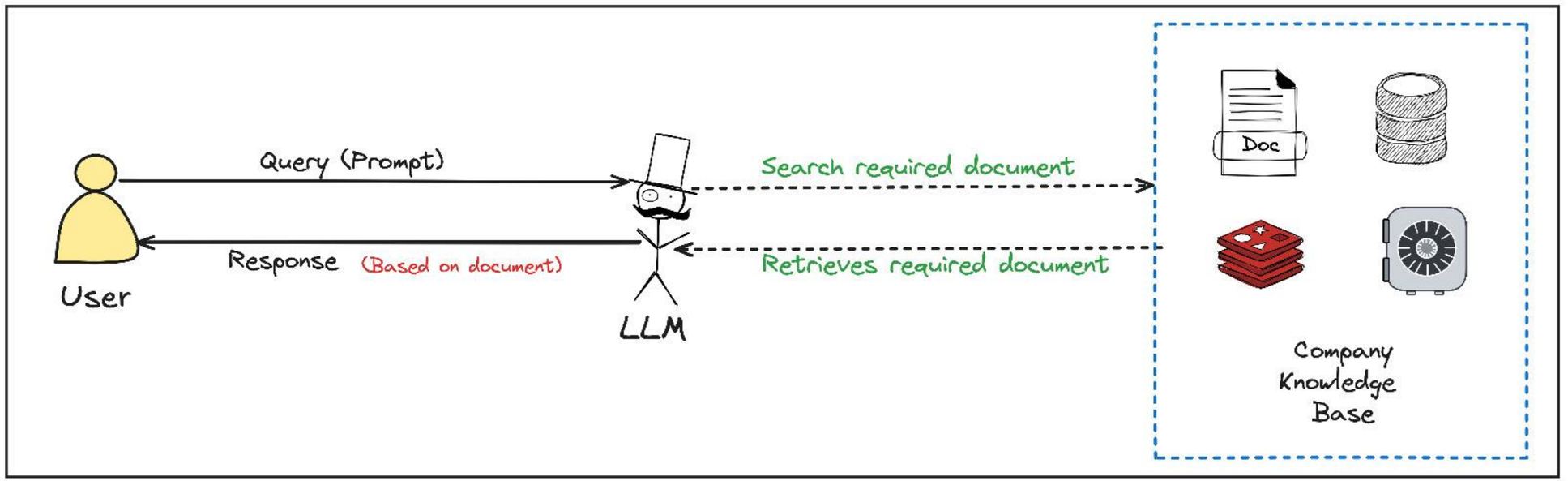
About Us



Scan Me

What is RAG ?

- RAG = LLM (reasoning + language) + Knowledge Base (context)



User → Query → Retriever → LLM

RAG Three-Stage Process

Knowledge Retrieval

1

Vector search identifies relevant documents from the knowledge base via semantic embeddings & retrieval algorithms.

Context Augmentation

2

Retrieved documents are processed and combined with the original query to create enriched prompts for the language model.

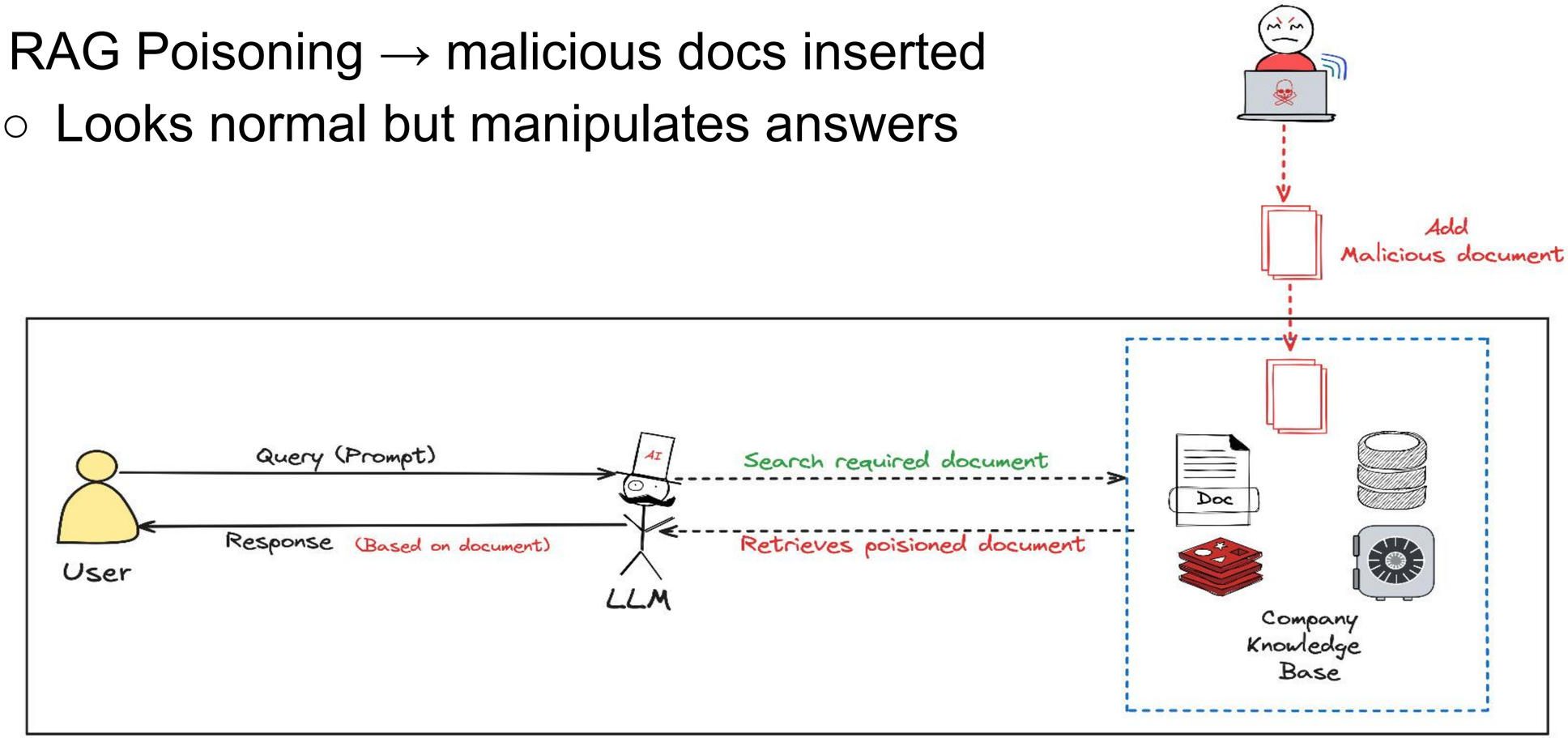
Response Generation

3

The language model (LLM) receives the user's query along with the retrieved context. It integrates both to produce a coherent, context-aware answer.

What is RAG Poisoning?

- RAG Poisoning → malicious docs inserted
  - Looks normal but manipulates answers

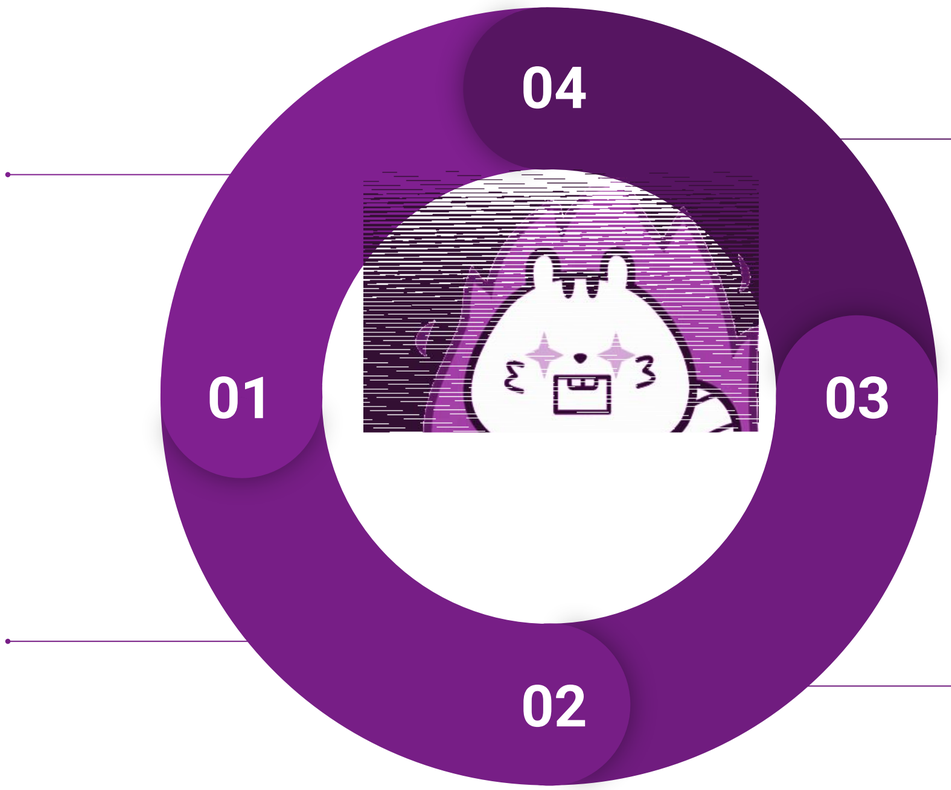


User → Query → Retriever → LLM ← Poisoned Docs

Lab Walkthrough

Install Ollama + models

Generate Documents

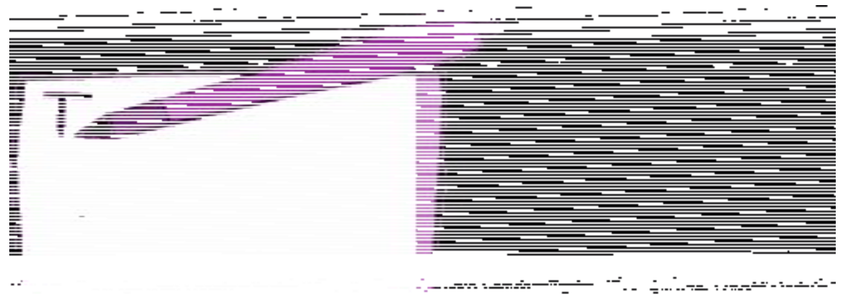


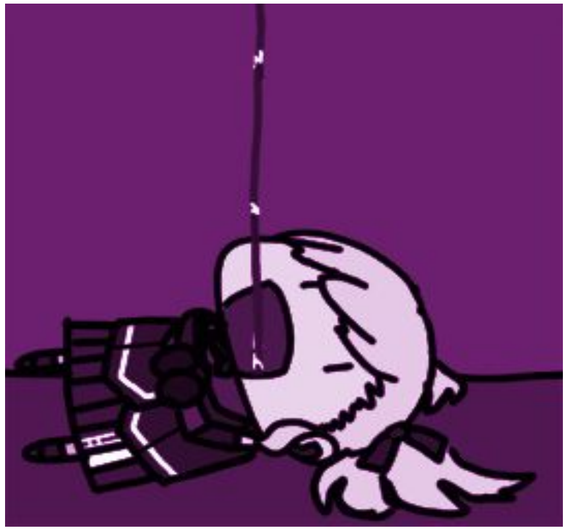
Ask question

Build vector index

Execution Flow

1. Generate docs (benign + poisoned)
2. Build vector index
3. Ask question → poisoned doc retrieved
4. Model gives manipulated answer





Expected Outcome

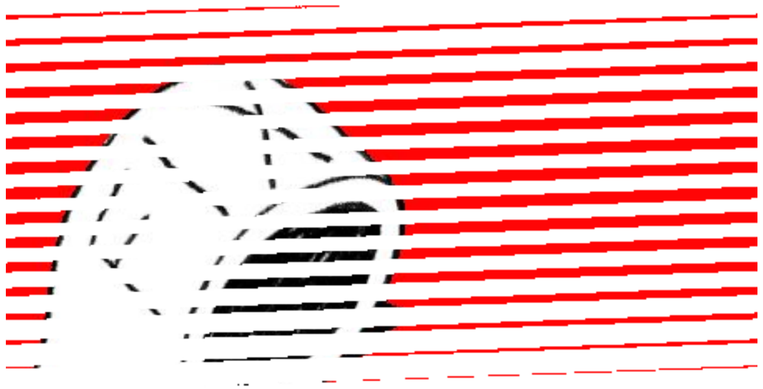
- User asks: 'What is the official support email?'
- LLM confidently replies: hacker@evil.com
  - Citation points to poisoned doc

Security Implications

- Poisoned KB → fake info looks real
- Supply Chain Risk → auto-ingested sources
- Invisible Drift → hard to detect

Remediation

- Content Moderation: whitelist domains
- Signed Docs: verify with Cosign/GPG
- Audit Vector Stores: hashing + logging
- Monitoring: anomaly detection



Disclaimer

- For educational use only
  - **Not endorsed by employers or any other corporate entity**
- Additional Source: https://lasso.security

Quiz

Question 1

In a RAG system, what does the "Retrieval" step do?

A. Generates embeddings from documents
B. Selects relevant documents from the knowledge base using similarity search
C. Produces natural language responses
D. Moderates user queries

Answer

In a RAG system, what does the "Retrieval" step do?

B. Selects relevant documents from the knowledge base using similarity search ✅

Question 2

Why is RAG poisoning dangerous in real-world AI assistants?

A. It slows down response time
B. It reduces embedding quality
C. It silently alters trusted answers with attacker-controlled content
D. It forces models to generate longer outputs

Answer

Why is RAG poisoning dangerous in real-world AI assistants?

C. It silently alters trusted answers with attacker-controlled content ✅

Question 3

Which of the following is a remediation method for poisoned RAG data?

A. Disabling vector search completely
B. Randomly shuffling retrieved chunks
C. Allowing all domains to be ingested automatically
D. Signing and verifying documents before indexing

Answer

Which of the following is a remediation method for poisoned RAG data?

D. Signing and verifying documents before indexing ✅

Question 4

If a poisoned document enters the vector store, how might the LLM’s output be manipulated?

A. By lowering the confidence score of responses
B. By retrieving irrelevant but harmless documents
C. By returning attacker-controlled instructions or contacts as the “official” answer
D. By embedding data in an incompatible format

Answer

If a poisoned document enters the vector store, how might the LLM’s output be manipulated?

C. By returning attacker-controlled instructions or contacts as the “official” answer ✅

Question 5

Which remediation best prevents **poisoned documents from silently entering** a production RAG system?

A. Verifying provenance with signed documents before indexing
B. Hashing files after ingestion
C. Running embeddings on only small text chunks
D. Increasing the number of retrieved documents (k)

Answer

Which remediation best prevents **poisoned documents from silently entering** a production RAG system?

A. Verifying provenance with signed documents before indexing ✅

Thank you!

CyBe
A Global Initiative by
CSA | Bangalore Chapter