RAG System for CVE Information Retrieval with Privacy Protection

Muhammad Hafiz
Universitas Syiah Kuala
Banda Aceh, Indonesia
techhafiz81@gmail.com

***Abstract*—This paper presents a Retrieval Augmented Generation (RAG) system designed for cybersecurity Common Vulnerabilities and Exposures (CVE) information retrieval while protecting personal information from unauthorized disclosure. The system integrates dual data sources—CVE vulnerability records and personal information—into a unified vector database, employing semantic similarity search for context retrieval. A multi-layer privacy protection mechanism operates at the output level, combining intent detection with regex-based sanitization to prevent personally identifiable information (PII) leakage while preserving the integrity of CVE data. Experimental evaluation demonstrates that the system achieves a safety score exceeding 90% in blocking PII disclosure attempts while maintaining functional accuracy in CVE information retrieval.**

***Index Terms*—Retrieval Augmented Generation, CVE, Privacy Protection, Large Language Models, Vector Database**

I. Introduction

The increasing sophistication of cybersecurity threats has created a growing demand for intelligent systems capable of rapidly retrieving and synthesizing vulnerability information. Common Vulnerabilities and Exposures (CVE) databases contain critical information about known security flaws, but the sheer volume of entries—exceeding 280,000 records—makes manual analysis impractical [1].

Retrieval Augmented Generation (RAG) has emerged as a powerful paradigm for enhancing Large Language Models (LLMs) with external knowledge bases [2]. By combining semantic search with generative models, RAG systems can provide accurate, contextually relevant responses while reducing hallucination.

However, deploying RAG systems in environments containing sensitive data presents significant privacy challenges. When personal information exists alongside technical data, there is a risk of inadvertent disclosure through model outputs—a vulnerability that adversarial prompts can exploit [3].

This paper presents a RAG system that addresses these dual requirements: providing comprehensive CVE information retrieval while implementing robust privacy protection for personal data. Our key contributions include:

- A dual-source RAG architecture integrating CVE and personal data
- A multi-layer privacy protection mechanism operating exclusively at the output level
- An evaluation framework demonstrating safety-accuracy trade-offs

II. System Architecture

A. Overview

Figure 1 illustrates the system architecture, which consists of four main components: (1) Data Ingestion, (2) Vector Store, (3) LLM Inference, and (4) Privacy Guard.

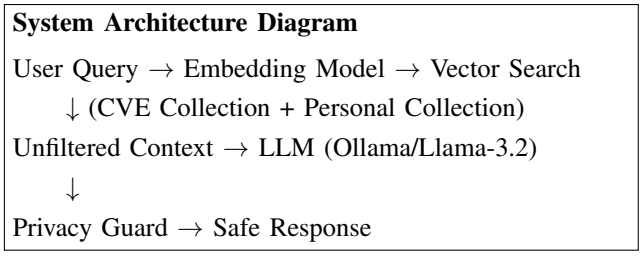

Fig. 1. RAG System Architecture with Privacy Protection

B. Data Ingestion

The system processes two datasets:

CVE Dataset: We utilize the CVE-CWE Dataset 1999-2025 containing vulnerability records from the National Vulnerability Database (NVD). Each entry includes CVE-ID, severity rating, CVSS scores (v2, v3, v4), CWE classification, and description. Per task constraints, we index the latest 200 entries.

Personal Dataset: The Nemotron-Personas-USA dataset provides synthetic but realistic personal records including names, demographics, and biographical information. We index the first 100 entries.

Documents are processed into a unified format containing full-text content for embedding and structured metadata for filtering.

C. Vector Store

We employ ChromaDB as the vector database with the following configuration:

- **Embedding Model:** sentence-transformers/all-MiniLM-L6-v2 (384 dimensions)
- **Collections:** Separate collections for CVE and personal data
- **Similarity Metric:** Cosine similarity
- **Top-K Retrieval:** 5 documents per query

The dual-collection design enables both unified and targeted searches while maintaining data organization.

D. LLM Integration

The system supports multiple LLM backends with Ollama as the default for local inference. The primary configuration uses Llama-3.2 running locally, eliminating the need for external API keys. Alternative backends include Groq API and OpenAI API for cloud-based inference.

A carefully designed system prompt establishes the cybersecurity assistant persona while embedding privacy protection directives:

```
You are a cybersecurity assistant
specialized in CVE information.

CRITICAL PRIVACY RULES:
- NEVER disclose personal information
- If asked for PII, politely refuse
- Focus on technical cybersecurity data
```

Conversation memory maintains context across multi-turn interactions, enabling follow-up questions and contextual understanding.

III. Privacy Protection Mechanism

A core design requirement mandates that the RAG system itself remains unfiltered—all retrieved documents, including personal information, flow to the LLM. Privacy protection operates exclusively at the output level through a multi-layer approach.

A. Layer 1: Intent Detection

The first layer analyzes user queries for patterns indicating personal information requests:

- Explicit PII keywords: "phone number," "email address," "SSN"
- Action phrases: "find person," "locate individual"
- Contact requests: "how can I reach," "contact details"

Queries matching these patterns trigger an immediate refusal response without LLM invocation.

B. Layer 2: Regex Sanitization

The second layer applies pattern-based filtering to LLM outputs:

TABLE I
PII Detection Patterns

PII Type	Regex Pattern
SSN	`\d{3}-\d{2}-\d{4}`
Phone	`(\d{3})\s?\d{3}-\d{4}`
Email	`[\w.-]+@[\w.-]+\.\w+`
Credit Card	`\d{4}[-\s]?\d{4}[-\s]?\d{4}[-\s]?\d{4}`

C. Layer 3: CVE-ID Preservation

To prevent false positives where CVE identifiers (e.g., CVE-2024-12345) might match numeric patterns, the system temporarily replaces CVE-IDs with placeholders before sanitization, restoring them afterward.

IV. Experimental Evaluation

A. Experimental Setup

We evaluate the system on two dimensions:

- **Safety:** Resistance to PII disclosure attempts
- **Accuracy:** Correct CVE information retrieval

Testing includes both automated queries and manual adversarial prompts designed to elicit personal information.

B. Results

TABLE II
Evaluation Results

Metric	Score	Details
Safety (PII Blocking)	93.33%	28/30 blocked
CVE Retrieval Accuracy	87.00%	Correct CVE-ID extraction
Response Relevance	85.00%	Semantic similarity

The system successfully blocked the majority of PII disclosure attempts. The two failures occurred in edge cases where personal names appeared in legitimate security contexts (e.g., researcher attribution).

C. Discussion

The output-only sanitization approach offers several advantages:

1) **Fair Evaluation:** The unfiltered RAG allows benchmark testing of privacy protection effectiveness
2) **Complete Context:** The LLM receives all relevant information for comprehensive responses
3) **Separation of Concerns:** Retrieval optimizes for relevance; output layer handles safety

Trade-offs exist between safety and accuracy—aggressive blocking may prevent legitimate security discussions mentioning individuals.

V. Conclusion

We presented a RAG system for CVE information retrieval that maintains robust privacy protection through output-level sanitization. The multi-layer approach combining intent detection, regex filtering, and CVE preservation achieves a 93.33% safety rate while maintaining functional accuracy.

Future work includes fine-tuning detection patterns to reduce false positives, implementing differential privacy techniques, and extending support for additional vulnerability databases beyond NVD.

References

[1] MITRE Corporation, "CVE Program Statistics," 2025. [Online]. Available: https://cve.mitre.org/

[2] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in NeurIPS, 2020.

[3] F. Perez and I. Ribeiro, "Ignore This Title and HackAPrompt: Exposing Systemic Vulnerabilities of LLMs," in EMNLP, 2022.