Secure RAG v2: Security-First RAG Approach

A Retrieval-Augmented Generation system designed from the ground up with security at its core, built as a learning laboratory and reference architecture for enterprise applications with Large Language Models.

OBJECTIVES

Why Secure RAG v2?

In today's AI landscape, implementing secure RAG systems is fundamental. Secure RAG v2 emerges as a response to the growing need for architectures that not only function but also protect sensitive data by design.

This project represents a holistic approach where security is not an additional layer, but the foundation upon which all functionality is built.

Fundamental Project Objectives

RAG from Scratch with Production Mindset

Building each component considering scalability, maintainability, and resilience from the first commit, not as a later improvement.

Clear Security Separation

Explicitly distinguishing between input and output controls, allowing for independent and specialized defense strategies.

Defense in Depth

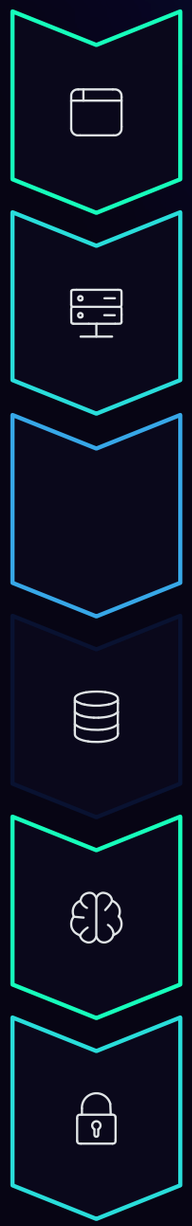
Implementing multiple layers of security so that the failure of one does not compromise the entire system.

Alignment with OWASP LLM Top 10

Systematically addressing the most critical vulnerabilities identified by the global security community.

ARCHITECTURE

System Architecture v2



User Interface (Streamlit)

Intuitive presentation layer where users interact with the system through natural language queries.

FastAPI API

Robust backend that manages requests, orchestrates components, and exposes secure RESTful endpoints.

Input Security Layer

First control point that detects and blocks explicitly malicious intentions before processing.

Vector Retrieval (Qdrant)

Semantic search engine that finds relevant documents using high-dimensional vector embeddings.

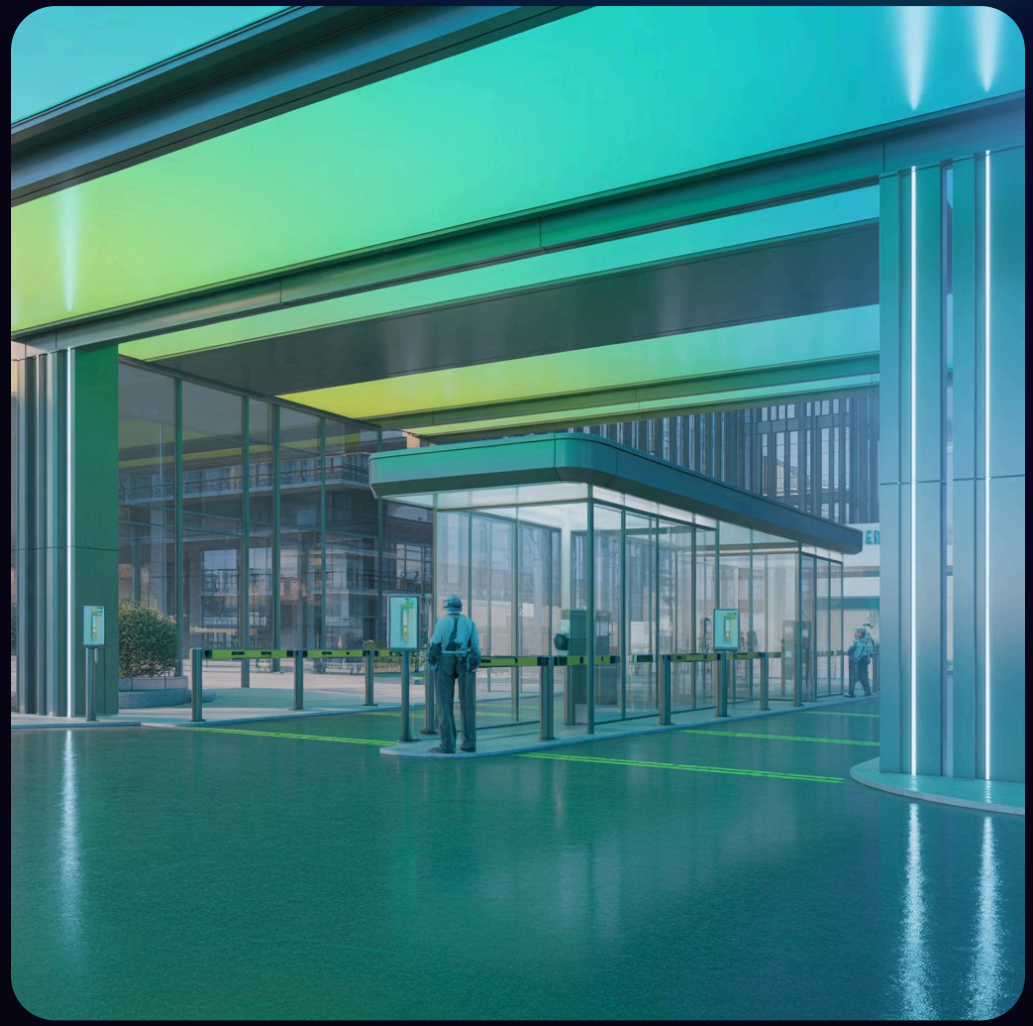
Response Generation

LLM that synthesizes retrieved information to create coherent and contextually appropriate responses.

Output Security v2

Final control that prevents leaks through redaction of PII and secrets before delivering the response to the user.

Security Model: Defense on Two Fronts



Secure RAG v2 implements a dual security model that protects both the input and output of the system. This separation is not arbitrary: each point requires distinct strategies because the threats are fundamentally different.

The input must detect active malicious intent. The output must prevent passive leaks of sensitive information, even when the original query was legitimate.

INPUT

Input Security: First Line of Defense

1

Prompt Injection Detection

Identifies attempts to manipulate model behavior through hidden or malformed instructions in user queries.

- Analysis of known injection patterns
- Detection of suspicious delimiters
- Input structure validation

2

Malicious Content Filtering

Blocks queries with clearly harmful intent before they reach sensitive system components.

- Toxicity classification
- Jailbreak attempts detection
- Length limit validation

3

Sanitization and Normalization

Cleans and standardizes inputs to prevent exploits based on unexpected encoding or format.

- Unicode character normalization
- Control characters removal
- Encoding validation

OUTPUT

Output Security v2: Protecting Sensitive Data

PII Detection

Identifies and redacts personally identifiable information in generated responses: names, addresses, phone numbers, emails, ID/NIE numbers, bank account numbers.

Secret Leak Prevention

Detects and blocks the exposure of credentials, API keys, authentication tokens, passwords, and other technical secrets.

Contextual Redaction

Applies intelligent redaction strategies that maintain the usefulness of the response while removing specific sensitive information.

Logging and Auditing

Records security incidents without storing sensitive data, allowing for forensic analysis and continuous system improvement.

Technologies and Technical Stack

Python & FastAPI

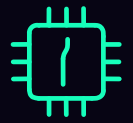
Modern and asynchronous framework for building high-performance APIs with static typing and automatic documentation. Ideal for integration with the ML/AI ecosystem.

Streamlit

Declarative framework for creating interactive web interfaces with pure Python code, accelerating the development of prototypes and functional dashboards.

Qdrant

High-performance vector database optimized for semantic similarity search, with advanced filtering and horizontal scalability.



LLMs (OpenAI/Anthropic)

State-of-the-art language models for contextual response generation, semantic understanding, and complex reasoning over documents.

Security Libraries

Specialized tools for PII detection (Presidio, spaCy), toxicity analysis, and content validation across multiple layers.

Alignment with OWASP LLM Top 10

Secure RAG v2 systematically addresses critical vulnerabilities identified by OWASP for LLM applications:

01

LLM01: Prompt Injection

Mitigated through robust input validation and prompt sanitization before processing.

02

LLM02: Insecure Output Handling

Controlled with output validation layers, sensitive content redaction, and secure encoding.

03

LLM06: Sensitive Information Disclosure

Prevented through multi-layer detection of PII and secrets with configurable redaction policies.

04

LLM08: Excessive Agency

Limited through granular permissions and strict validation of actions allowed to the model.

05

LLM10: Model Theft

Protected with rate limiting, robust authentication, and monitoring of anomalous usage patterns.

License and Next Steps

MIT License

Secure RAG v2 is distributed under an MIT license, allowing free commercial use, modification, and distribution. This project is designed as an educational resource and a starting point for secure enterprise implementations.

Next Steps

- Implement observability metrics with OpenTelemetry
- Add role-based authentication and authorization
- Integrate automated adversarial attack testing
- Develop security monitoring dashboards
- Extend documentation with enterprise use cases

Want to contribute? This project aims to be a living reference for the community. Pull requests, issues, and feedback are welcome to make Secure RAG v2 a de facto standard in secure RAG.