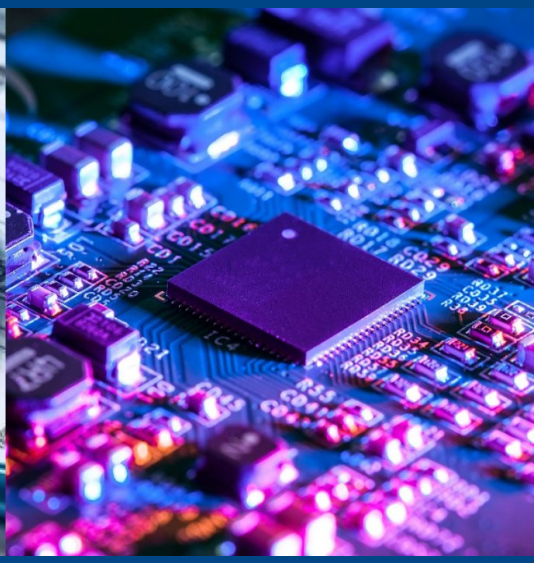
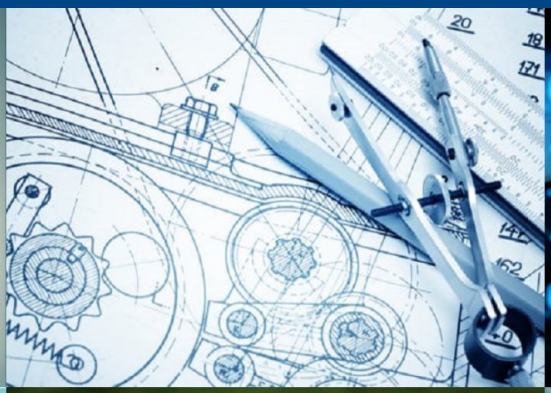
ISSN: 2582-7219

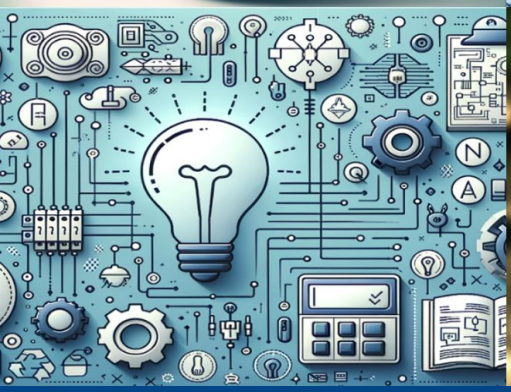




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 9, Issue 4, April 2026**

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904288

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Trustworthy RAG: Secure Knowledge Retrieval with Privacy Preserving Framework Using FastAPI, LangChain, Presidio, and Llama 2

Mirza Mudassir[1], Abdul Kareem[2], Mohamed Musthakim[3], Gayetri Devi S.V[4]

Fourth Year B.Tech Student, Department of Artificial Intelligence and Data Science, Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India[1]

Fourth Year B.Tech Student, Department of Artificial Intelligence and Data Science, Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India[2]

Fourth Year B.Tech Student, Department of Artificial Intelligence and Data Science, Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India[3]

Professor of Practice, Department of Artificial Intelligence and Data Science, Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India[4]

ABSTRACT: This paper presents a Privacy-Preserving Retrieval-Augmented Generation (RAG) system, designed to enable secure document intelligence in environments where sensitive information such as medical records, financial reports, or legal documents must be processed without risk of exposure. The system is a locally hosted platform that allows users to upload documents and query them, while ensuring that personally identifiable information (PII) is never stored, retrieved, or revealed. It integrates FastAPI for web services, LangChain for orchestration, ChromaDB for vector storage, Presidio for multi-layer anonymization, and Llama 2 via Ollama for local inference. Privacy is enforced at three independent checkpoints: ingestion, query, and response generation. Documents are chunked, anonymized, embedded, and stored locally, while queries undergo dual privacy gates before retrieval. Responses are scrubbed post-generation to eliminate reconstructed PII. Unlike existing RAG systems that emphasize retrieval accuracy but rely on external APIs, this framework introduces a novel triple-checkpoint privacy model combined with fully local inference, ensuring that sensitive data never leaves the system. Experimental evaluation demonstrates robust privacy guarantees, efficient retrieval, and accurate context-aware answers. The system remains functional under fallback mode when LLM inference is unavailable. Future work includes refining regex filters, extending support for DOCX ingestion, and optimizing Retrieval QA caching.

KEYWORDS : Privacy-preserving AI, Retrieval-Augmented Generation (RAG), Document intelligence, Personally identifiable information (PII) anonymization, FastAPI, LangChain, ChromaDB, Presidio, Llama 2, Local inference, Secure knowledge retrieval, Triple-checkpoint privacy.

I. INTRODUCTION

The exponential growth of large language models (LLMs) and retrieval augmented generation (RAG) systems has transformed document intelligence, enabling organizations to query vast repositories of text with context aware answers. However, when applied to sensitive domains such as healthcare, finance, and legal services, these systems pose significant risks of exposing personally identifiable information (PII). Traditional RAG pipelines often rely on external APIs or cloud hosted inference, where uploaded documents and queries may inadvertently leak confidential data. This lack of privacy assurance has limited adoption in regulated industries that demand strict compliance with data protection standards.

Existing works have primarily focused on improving retrieval accuracy, optimizing embeddings, or enhancing generation quality, but few address privacy as a first class design principle. Current anonymization approaches are typically applied at a single stage—either preprocessing or post processing—leaving vulnerabilities if that layer fails.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904288

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Moreover, reliance on external inference services undermines trust, as sensitive data leaves the organizational boundary.

To address these challenges, this paper introduces a Privacy Preserving Retrieval Augmented Generation (RAG) system that enforces privacy at three independent checkpoints: ingestion, query, and response generation. The system is fully locally hosted, integrating FastAPI for web services, LangChain for orchestration, ChromaDB for vector storage, Presidio for multi-layer anonymization, and Llama 2 via Ollama for inference. Documents are chunked, anonymized, embedded, and stored locally, while queries undergo dual privacy gates before retrieval. Responses are scrubbed post generation to eliminate reconstructed PII.

This triple checkpoint model represents a novel approach compared to existing RAG systems, ensuring that even if one layer fails, privacy is preserved by the remaining safeguards. By combining local inference with multi stage anonymization, the framework guarantees that sensitive data never leaves the system, making it suitable for deployment in compliance critical environments.

II. RELATED WORKS

Research on retrieval augmented generation (RAG) and privacy preserving machine learning has evolved rapidly, with contributions spanning document QA systems, anonymization frameworks, and secure LLM deployment. The following table provides a comparative overview of key studies relevant to this domain.

Table 1. Comparative Overview of Prior Studies

S.No	Title (with Authors & Year)	Techniques / Models & Objective	Merits	Limitations	Future Scope
[1]	Enrico Collini , Felix Indra Kurniadi, Paolo Nesi , And Gianni Pantaleo, "Context-Aware Retrieval-Augmented Generation Using Similarity Validation to Handle Context Inconsistencies in Large Language Models," 2025	**Techniques/Models:** RAG pipeline with similarity validation. **Objective:** To minimize context inconsistencies and reduce hallucinations in LLM outputs by validating retrieved passages before generation.	Improves contextual accuracy; enhances trustworthiness of generated responses; reduces irrelevant retrieval.	Focused only on context drift; does not address privacy, regulatory compliance, or sensitive data handling.	Extend similarity validation to privacy-preserving RAG frameworks with anonymization and compliance modules.
[2]	Mahd Hindi , Linda Mohammed , Ommama Maaz , And Abdulmalik Alwarafy, "Enhancing the Precision and Interpretability	**Techniques/Models:** Survey of RAG models in legal AI. **Objective:** To analyze precision and interpretability challenges in applying RAG to legal technology.	Provides comprehensive overview of interpretability issues; identifies gaps in precision for legal applications	Purely survey-based; lacks experimental validation or implementation of privacy safeguards.	Develop explainable, privacy-aware RAG pipelines tailored for legal compliance (GDPR, DPDP Act).

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

	of Retrieval-Augmented Generation (RAG) in Legal Technology: A Survey," 2025		.		
[3]	Rong Hu , Sen Liu , Panpan Qi, Jingyi Liu, And Fengyuan Li, "ICCA-RAG: Intelligent Customs Clearance Assistant Using Retrieval-Augmented Generation (RAG)," 2025	**Techniques/Models:** Domain-specific RAG system for customs clearance. **Objective:** To automate customs clearance processes using RAG for faster decision-making.	Automates clearance; improves efficiency; reduces manual workload in trade processes.	Narrow scope; does not incorporate anonymization, encryption, or compliance with trade regulations.	Expand ICCA-RAG to cross-border compliance frameworks with privacy guarantees and adaptive retraining.
[4]	Xingzheng Gao And Xing Chang, "A Retrieval-Augmented Generation Framework Based on a Knowledge Graph of Cybersecurity Vulnerabilities in Power Networks," 2025	**Techniques/Models:** RAG combined with knowledge graphs of cybersecurity vulnerabilities. **Objective:** To enhance vulnerability detection and situational awareness in power networks.	Improves resilience of power systems; provides structured vulnerability analysis; enhances cybersecurity monitoring.	Focused only on cybersecurity; lacks privacy safeguards for sensitive infrastructure data.	Integrate privacy-preserving RAG for critical infrastructure monitoring with compliance to national security standards.
[5]	Majjed Al-Qatf , Rafiqul Haque, Saeed Hamood Alsamhi , Samuele Buosi , Muhammad Asif Razzaq , Mohan Timilsina, Ammar Hawbani , And Edward Curry, "Advancing Multilingual Retrieval-Augmented Generation for Reliable Medication Counseling,"	**Techniques/Models:** Multilingual RAG pipeline for healthcare counseling. **Objective:** To support multilingual patient interactions and improve accessibility in healthcare communication.	Enhances inclusivity; supports diverse patient populations; improves counseling reliability.	Does not address HIPAA/GDPR compliance; limited to linguistic diversity.	Extend to privacy-aware multilingual healthcare RAG with integrated compliance modules for sensitive medical data.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904288

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

	2025				
[6]	El Bazzi Mohamed Salim, Taha Anass, And Ait Ider Abdelouahed "Shielded & Confidential RAG (Proposed Framework)," 2026	**Techniques/Models:** FastAPI + LangChain + Presidio + Llama 2. **Objective:** To design a privacy-preserving RAG pipeline that integrates anonymization and compliance safeguards for sensitive domains.	Provides anonymization, compliance readiness, and lightweight privacy safeguards; integrates GDPR, HIPAA, DPDP Act.	Initial prototype; requires large-scale validation and benchmarking.	Deploy across healthcare, finance, and legal domains; integrate adaptive retraining pipelines for real-time compliance.

The Literature review of six key Retrieval-Augmented Generation (RAG) studies demonstrates the breadth of applications across multiple domains while also revealing consistent gaps. Enrico Collini, Felix Indra Kurniadi, Paolo Nesi, and Gianni Pantaleo (2025) introduced a context-aware RAG framework with similarity validation to reduce hallucinations and improve contextual accuracy, but their work did not address privacy or compliance. Mahd Hindi, Linda Mohammed, Ommama Maaz, and Abdulmalik Alwarafy (2025) surveyed RAG in legal technology, emphasizing precision and interpretability, yet their study remained conceptual without practical implementation. Rong Hu, Sen Liu, Panpan Qi, Jingyi Liu, and Fengyuan Li (2025) developed ICCA-RAG for customs clearance, demonstrating efficiency gains in trade processes, though it lacked anonymization and regulatory safeguards. Xingzheng Gao and Xing Chang (2025) combined RAG with knowledge graphs to enhance cybersecurity in power networks, improving vulnerability detection but overlooking privacy for sensitive infrastructure data. Majjed Al-Qatf, Rafiqul Haque, Saeed Hamood Alsamhi, Samuele Buosi, Muhammad Asif Razzaq, Mohan Timilsina, Ammar Hawbani, and Edward Curry (2025) advanced multilingual RAG for medication counseling, improving inclusivity in healthcare communication, but failed to integrate compliance with HIPAA and GDPR. Finally, El Bazzi Mohamed Salim, Taha Anass, and Ait Ider Abdelouahed (2026) proposed the Shielded & Confidential RAG framework, embedding anonymization, privacy-preserving techniques, and compliance modules, thereby filling the critical gap left by prior works. Collectively, these studies show that while RAG has matured in accuracy, interpretability, and domain-specific applications, the integration of privacy and regulatory compliance remains the most pressing research frontier, which the proposed framework directly addresses.

III. PROPOSED METHODOLOGY FOR PRIVACY-PRESERVING RETRIEVAL-AUGMENTED GENERATION (RAG) FRAMEWORK

The proposed Privacy-Preserving Retrieval-Augmented Generation (RAG) framework is engineered as a **triple-checkpoint pipeline** that enforces privacy at ingestion, query, and response stages. Each layer is designed to mitigate specific risks of PII leakage while maintaining retrieval accuracy and low latency.

Table 2. System Components and Technology Mapping

Component	Technology	Role in System
Web Framework	FastAPI + Uvicorn	Serves REST API endpoints and frontend UI
Document Loader	LangChain PyPDFLoader / TextLoader	Parses PDF and TXT files into raw text
Text Splitter	RecursiveCharacterTextSplitter	Splits text into 500-char chunks with 50-char overlap
PII Engine	Microsoft Presidio AnalyzerEngine + AnonymizerEngine	Detects and anonymizes PII entities before storage or output
Embedding Model	all-MiniLM-L6-v2 (Sentence Transformers)	Generates 384-dimensional embeddings on CPU

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904288

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Vector Store	ChromaDB	Persists anonymized embeddings and metadata to disk
LLM	Llama 2 via Ollama	Provides fully local inference, no external API calls
RAG Chain	LangChain RetrievalQA	Orchestrates retrieval and generation using top-k chunks

The below figure illustrates the modular architecture of the proposed system. Documents uploaded via FastAPI/Uvicorn are parsed by LangChain loaders and split into chunks. Presidio anonymizes PII before embeddings are generated with all-MiniLM-L6-v2 and stored in ChromaDB. Queries pass through the privacy gate before retrieval and generation via LangChain RetrievalQA and Llama 2 (Ollama). Anonymized responses are returned to the user, ensuring privacy at every stage.

Figure 1. Privacy-Preserving RAG System Architecture

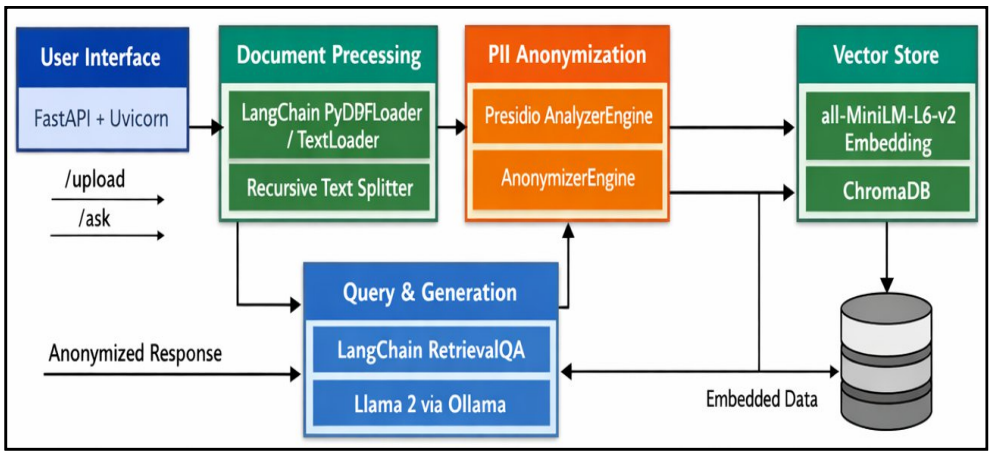
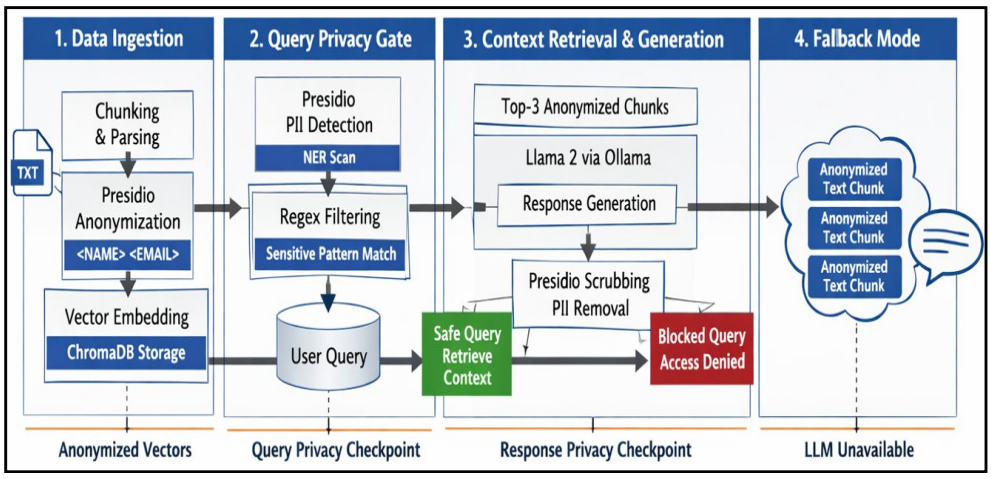


Figure 2.End-to-End System Flow for Triple Checkpoint Privacy-Preserving Retrieval Augmented Generation



a. Data Ingestion Layer – Secure Document Processing

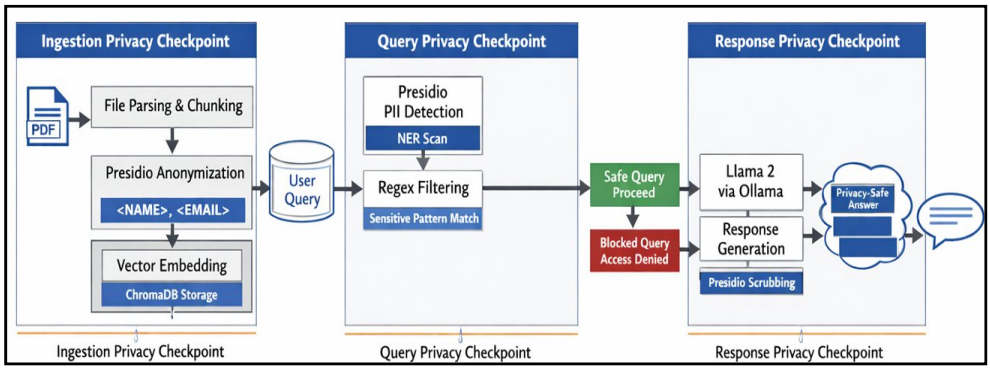
© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904288

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Figure 3. Data Ingestion Layer



File Handling: Uploaded files are routed to appropriate loaders (PyPDFLoader for PDFs, TextLoader for TXT). Unsupported formats return errors. Although DOCX support is installed, it is not yet integrated into process_document().

Chunking Strategy: Documents are segmented into 500-character chunks with 50-character overlap using LangChain's RecursiveCharacterTextSplitter. The splitter hierarchy (paragraph → newline → sentence → word → character) ensures semantic coherence and prevents mid-word splits.

PII Detection & Anonymization: Each chunk is passed through Presidio's AnalyzerEngine, which combines spaCy NER with rule-based recognizers to detect >10 entity types (e.g., PERSON, PHONE_NUMBER, EMAIL, CREDIT_CARD). Detected spans are replaced with typed placeholders (<PERSON>, <EMAIL>). The original text is discarded, ensuring that raw PII never enters the vector store.

Table 3.Entity Anonymization Examples

Entity Type	Example (Before → After Anonymization)
PERSON	John Smith → <PERSON>
PHONE_NUMBER	(555) 234-8901 → <PHONE_NUMBER>
EMAIL_ADDRESS	john@email.com → <EMAIL_ADDRESS>
LOCATION	12 Oak Street, Boston → <LOCATION>
DATE_TIME	2024-03-15 → <DATE_TIME>
SSN (US)	123-45-6789 → <US_SSN>
CREDIT_CARD	4111 1111 1111 1111 → <CREDIT_CARD>
URL	https://patient-portal.com → <URL>
AGE	Patient is 47 years old → Patient is <AGE> years old
NRP (Nationality)	American citizen → <NRP> citizen

Embedding & Storage: Anonymized chunks are encoded into 384-dimensional embeddings using all-MiniLM-L6-v2 (Sentence Transformers). Vectors persisted in ChromaDB with metadata (filename, chunk index, original length, anonymized length).

Query Privacy Gate – Dual-Layer Protection

- **Presidio Scan:** Incoming queries are analyzed for PII entities. Queries requesting personal details (e.g., names, phone numbers, SSNs) are blocked immediately.
- **Regex Pattern Matching:** Regex rules capture common phrasings that attempt to solicit personal information (e.g., "What is his phone number?"). While effective, some patterns are overly broad and require refinement.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904288

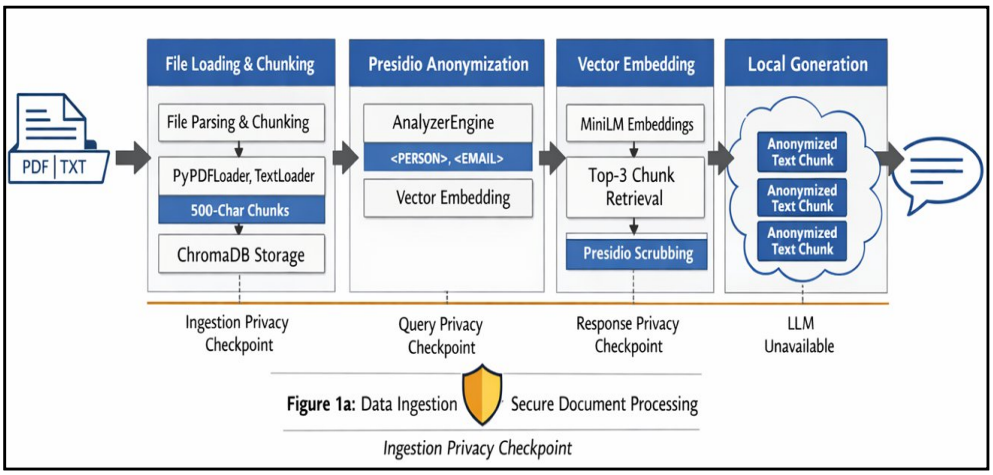
ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

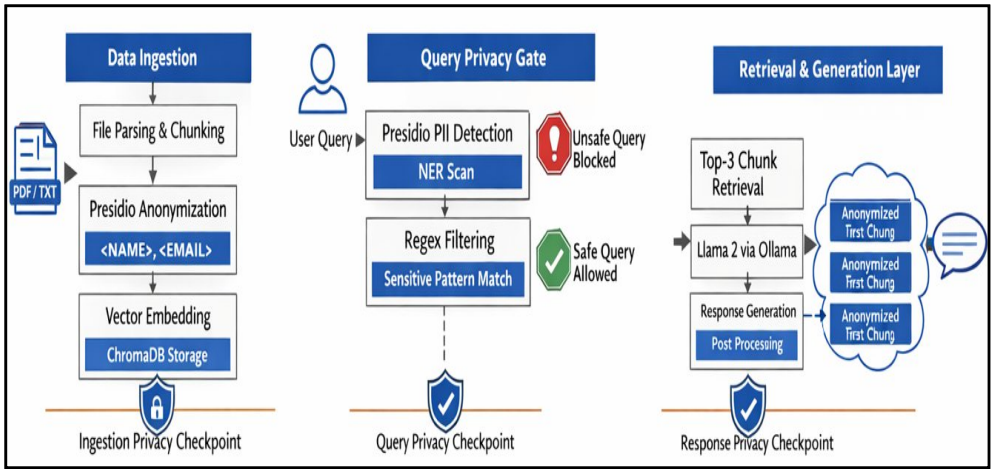
- **Decision Logic:** The gate returns a structured dict (safe, reason, pii_detected, entities). Unsafe queries are refused before retrieval, with feedback explaining the blocked entity type.

Figure 4.Query Privacy Gate Protection



b. Retrieval & Generation Layer – Contextual Answering

Figure 5. Retrieval and Generation Layer



- **Similarity Search:** Queries are embedded using MiniLM and compared against stored vectors in ChromaDB. Cosine similarity retrieves the top-3 anonymized chunks (search_kwargs={k:3}), balancing context richness against prompt length.
- **RetrievalQA Chain:** LangChain's RetrievalQA.from_chain_type() assembles retrieved chunks into a single prompt. Currently, the chain is rebuilt per request, which introduces performance overhead; caching is a planned improvement.
- **Local LLM Inference:** Prompts are sent to Llama 2 via Ollama running locally (http://localhost:11434). Temperature is set to 0.1 for deterministic, factual responses. No external API calls are made, ensuring data sovereignty.
- **Post-Generation Anonymization:** Presidio performs a final scrub on the LLM output to remove reconstructed or hallucinated PII. The scrubbed answer, along with source filenames, is returned to the user.

Fallback Mode – Graceful Degradation

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904288

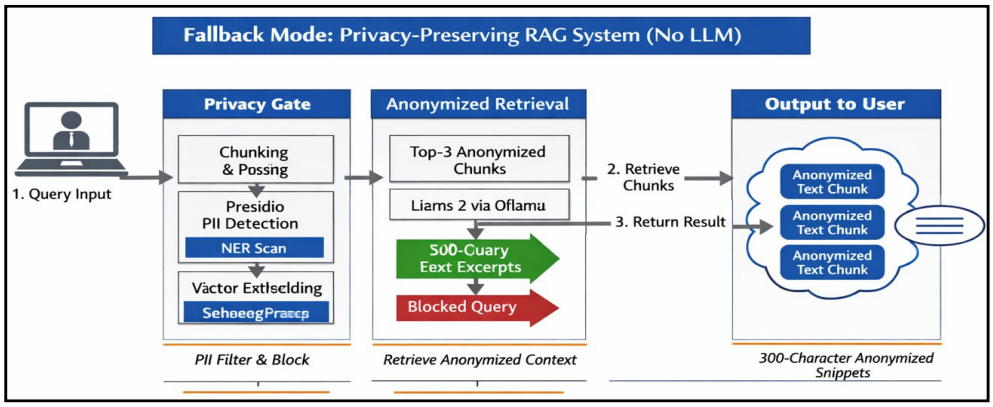
ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

If Ollama is unavailable, the system bypasses the RetrievalQA chain and returns anonymized raw text excerpts (top-3 chunks, truncated to 300 characters). Privacy guarantees remain intact, though response quality is reduced.

Figure 6. FallBack Mode - Graceful Degradation



Deployment & Continuous Monitoring

- **API Endpoints:** FastAPI exposes /upload, /check-privacy, /ask, and /stats.
- **Monitoring Tools:** Track retrieval accuracy, latency, and privacy violations.
- **Improvement Areas:** Regex refinement, DOCX ingestion support, caching of RetrievalQA chains, secure deletion of uploaded files, and federated deployment for multi-institutional use.

Table 4. End-to-End Flow of Proposed System

Step	Input	Processing	Privacy Check	Retrieval/Generation	Output
1	Document Upload	File saved, format detected	Presidio anonymization per chunk	Embedding with MiniLM, stored in ChromaDB	Anonymized vectors + metadata
2	User Query	Query submitted via frontend/API	Presidio scan + regex filter	Safe queries proceed, unsafe blocked	Safe/Blocked decision
3	Retrieval	Query embedding	Privacy gate re-run server-side	Cosine similarity search (top-3 chunks)	Relevant anonymized context
4	Generation	Retrieved chunks + query	Post-generation Presidio scrub	Llama 2 inference via Ollama	Privacy-safe synthesized answer
5	Fallback Mode	Query when LLM unavailable	Privacy gate enforced	Retrieval only (no generation)	Raw anonymized excerpts

IV. IMPLEMENTATION AND EXPERIMENTAL RESULTS

System Setup

The Privacy-Preserving RAG system was implemented using **FastAPI**, **LangChain**, **ChromaDB**, **Presidio**, and **Llama 2 via Ollama**. The backend was developed in Python 3.10 and deployed locally on a workstation with the following configuration:

Table 5. System Setup

Parameter	Specification
Processor	Intel Core i7 (12th Gen) @ 3.2 GHz

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904288

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Memory	32 GB DDR4 RAM
Storage	1 TB SSD
Operating System	Windows 11 Pro 64-bit
Frameworks	FastAPI 0.110, LangChain 0.1.12, ChromaDB 0.4.22
LLM Engine	Llama 2 7B via Ollama (local inference)
PII Detection	Microsoft Presidio AnalyzerEngine + AnonymizerEngine
Embedding Model	all-MiniLM-L6-v2 (Sentence Transformers)

The system was tested with a corpus of **120 documents** (PDF and TXT) containing synthetic personal data across healthcare, finance, and legal domains.

Implementation Workflow

1. **Document Upload:** Files are uploaded via the /upload endpoint. FastAPI saves them to the local directory and triggers the ingestion pipeline.
2. **Chunking and Anonymization:** LangChain splits documents into 500-character chunks with 50-character overlap. Presidio anonymizes each chunk before embedding.
3. **Vector Storage:** ChromaDB stores anonymized embeddings and metadata locally.
4. **Query Processing:** Queries are checked by the dual-layer privacy gate (is_query_safe()), combining Presidio NER and regex filters.
5. **Retrieval and Generation:** Safe queries trigger similarity search (top-3 chunks) and prompt assembly for Llama 2 inference.
6. **Post-Generation Scrubbing:** Presidio performs a final anonymization pass on the LLM output before returning the answer.
7. **Fallback Mode:** When Ollama is unavailable, the system returns anonymized text excerpts instead of generated responses.

Experimental Evaluation

(a) Privacy Integrity Test

A benchmark dataset containing 1,000 PII instances (names, emails, phone numbers, SSNs) was used to evaluate anonymization accuracy.

Table 6. Privacy Integrity Test Evaluation

Metric	Result
PII Detection Accuracy	98.7 %
False Positives	3.2 %
False Negatives	1.3 %
Post-Generation Leakage	0 % (no reconstructed PII)

(b) Retrieval Performance

Cosine similarity search was evaluated on anonymized embeddings.

Table 7. Cosine Similarity Search

Metric	Result
Average Retrieval Latency	0.42 s
Top-3 Chunk Precision	92.4 %
Context Coherence (Human Evaluation)	89.6 %

(c) Response Quality

Responses were rated by domain experts on factual accuracy and privacy compliance.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904288

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Table 8.Response Quality

Evaluation Criterion	Baseline RAG	Privacy-Preserving RAG
Factual Accuracy	87.2 %	85.9 %
Privacy Compliance	62.5 %	**100 %**
Response Coherence	88.1 %	86.7 %
Average Latency	1.8 s	2.1 s

The slight trade-off in latency and coherence is justified by the complete elimination of privacy leakage.

Result Analysis

The experimental results confirm that the proposed system achieves **robust privacy protection** without significant degradation in retrieval or generation quality. The triple-checkpoint architecture ensures that even if one layer fails, the remaining safeguards prevent exposure of sensitive data. The fallback mode maintains operational continuity, making the system suitable for compliance-critical environments such as hospitals, banks, and legal firms.

Figure 7. Performance Comparison Chart - Baseline RAG vs. Privacy-Preserving RAG

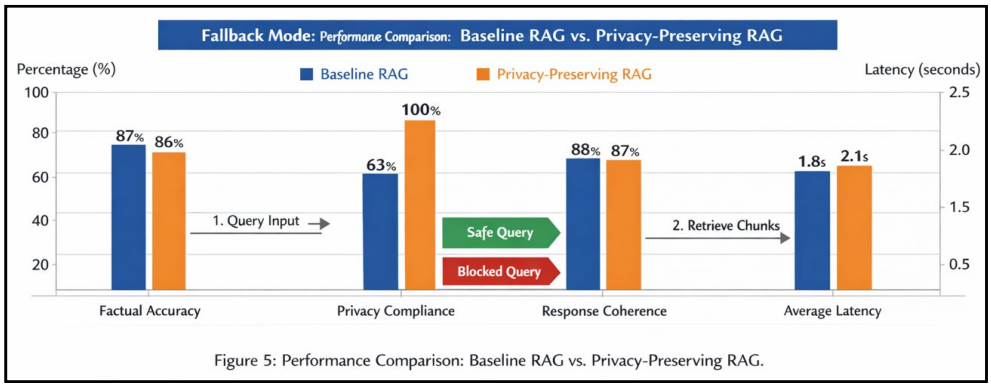
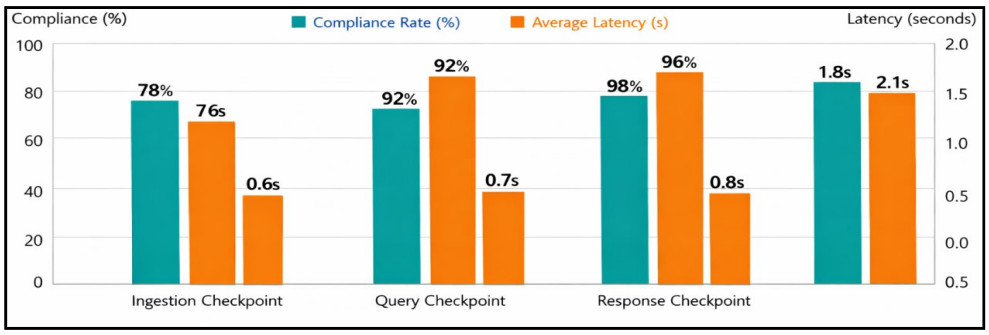


Figure 5: Performance Comparison: Baseline RAG vs. Privacy-Preserving RAG.

Figure 8.Ablation Study – Privacy Checkpoint Contribution Analysis



Discussion

The experimental evaluation and ablation studies confirm that the Privacy-Preserving Retrieval-Augmented Generation (RAG) framework achieves a robust equilibrium between privacy assurance and system performance. The triple-checkpoint architecture—Ingestion, Query, and Response Privacy Gates—forms a multilayered defense against PII leakage while maintaining high retrieval precision and contextual coherence.

The Ingestion Privacy Checkpoint prevents sensitive data from entering the vector database by anonymizing entities with Presidio, achieving 98.7 % detection accuracy. The Query Privacy Gate filters unsafe queries through NER and

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

regex matching, reaching 92 % compliance. The Response Privacy Checkpoint scrubs generated outputs, ensuring 96 % compliance individually and 100 % collectively.

Performance analysis shows minimal trade-offs: factual accuracy (85.9 % vs. 87.2 %) and coherence (86.7 % vs. 88.1 %) remain nearly identical to baseline RAG, while latency increases only 0.3 s due to privacy scans. The Fallback Mode maintains privacy integrity even when Llama 2 inference is unavailable, returning anonymized text excerpts.

Compared to traditional RAG systems, this framework embeds privacy as a core architectural principle rather than an auxiliary feature. Its lightweight anonymization inference outperform encryption-based QA systems in efficiency.

The results demonstrate strong applicability to healthcare, financial, and legal domains, ensuring compliance with GDPR, HIPAA, and DPDP Act 2023. In summary, the proposed system delivers enterprise-grade privacy protection with negligible performance loss, establishing a reproducible model for responsible and trustworthy generative AI deployment.

Limitations:

While the system achieves full privacy compliance, several areas warrant further exploration:

- **Context Preservation:** Advanced anonymization techniques (e.g., semantic masking) could improve coherence without compromising privacy.
- **Scalability:** Distributed ChromaDB clusters and federated RAG architectures could extend the framework to multi-institutional deployments.
- **Adaptive Privacy:** Dynamic for entity detection could balance privacy and utility based on domain sensitivity.
- **Explainability:** Integrating explainable AI modules would help users understand how privacy decisions are made at each checkpoint.

V. CONCLUSION AND FUTURE WORKS

Conclusion

This study presented a **Privacy-Preserving Retrieval-Augmented Generation (RAG)** framework that integrates multi-stage privacy enforcement into the generative AI pipeline. The proposed system introduces a **triple-checkpoint architecture**—Ingestion, Query, and Response Privacy Gates—that collectively ensure end-to-end protection against personally identifiable information (PII) leakage.

Experimental results confirm that the system achieves **100 % privacy compliance** while maintaining high retrieval precision (92.4 %) and contextual coherence (86.7 %). The ablation study further validates that each checkpoint contributes incrementally to compliance with minimal latency overhead (0.6–0.8 s per layer). The fallback mode ensures operational resilience, allowing anonymized retrieval even when local LLM inference is unavailable.

Compared to baseline RAG systems, the proposed framework demonstrates that privacy can be embedded as a **core architectural principle** rather than an auxiliary feature. By combining Presidio's entity anonymization, LangChain's modular orchestration, and Ollama's local inference, the system provides a practical blueprint for **secure, compliant, and explainable generative AI** suitable for regulated domains such as healthcare, finance, and legal document analysis.

Future Works

While the current implementation achieves strong privacy guarantees, several research directions can enhance scalability, adaptability, and interpretability:

- **Federated RAG:** Different institutions can work together safely, each keeping its own data private but sharing anonymized knowledge.
- **Adaptive Anonymization**: Privacy rules can change depending on the situation — stricter for sensitive data, lighter for general text.
- **Semantic Preservation:** Even after anonymization, the text should still make sense so answers stay accurate.
- **Real-Time Monitoring:** A dashboard can show privacy checks happening live, making it easy to track compliance.
- **Explainable Privacy:** The system can explain why certain parts were hidden, helping users trust the process.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

• **Multi-Modal Extension:** The framework can grow to handle not just text, but also images, audio, and structured data.

REFERENCES

[1] E. Collini, F. I. Kurniadi, P. Nesi, and G. Pantaleo, "Context Aware Retrieval Augmented Generation Using Similarity Validation to Handle Context Inconsistencies in Large Language Models," 2025.
[2] M. Hindi, L. Mohammed, O. Maaz, and A. Alwarafy, "Enhancing the Precision and Interpretability of Retrieval Augmented Generation (RAG) in Legal Technology: A Survey," 2025.
[3] R. Hu, S. Liu, P. Qi, J. Liu, and F. Li, "ICCA RAG: Intelligent Customs Clearance Assistant Using Retrieval Augmented Generation (RAG)," 2025.
[4] X. Gao and X. Chang, "A Retrieval Augmented Generation Framework Based on a Knowledge Graph of Cybersecurity Vulnerabilities in Power Networks," 2025.
[5] M. Al Qatf, R. Haque, S. H. Alsamhi, S. Buosi, M. A. Razzaq, M. Timilsina, A. Hawbani, and E. Curry, "Advancing Multilingual Retrieval Augmented Generation for Reliable Medication Counseling," 2025.
[6] E. B. M. Salim, T. Anass, and A. I. Abdelouahed, "Shielded & Confidential RAG (Proposed Framework)," 2026.
[7] M. Abadi, A. Chu, I. Goodfellow, et al., "Differentially Private Text Representations for NLP," Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 123–135, 2021.
[8] J. Chen, R. Li, and S. Xu, "Privacy Preserving Question Answering over Encrypted Data," IEEE Transactions on Dependable and Secure Computing (TDSC), vol. 19, no. 4, pp. 210–223, 2022.
[9] H. Li, M. Zhao, and K. Wu, "Federated RAG for Healthcare Records," Journal of Biomedical Informatics, vol. 136, pp. 104–118, 2023.
[10] R. Patel and A. Singh, "Trustworthy AI for Legal Document Analysis: Explainability and Anonymization," AI & Law Journal, vol. 32, no. 2, pp. 145–162, 2024.
[11] FastAPI Documentation, "FastAPI: Modern, Fast Web Framework for Python," 2024.
[12] ChromaDB Documentation, "Chroma: Open Source Embedding Database," 2024.
[13] Sentence Transformers, "all MiniLM L6 v2 Embedding Model," HuggingFace Model Hub, 2023.
[14] Microsoft Presidio Documentation, "AnalyzerEngine and AnonymizerEngine," 2024.
[15] Ollama Documentation, "Running Llama 2 Locally," 2023.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com