



PRIVATE RAGAI

Your Private, Document-Grounded AI Assistant

PrivateRAG AI
**Secure, Offline, Document-
Grounded Intelligence System**

Developed by:

Dr. Subramani

Mentored By:
Ramisha Rani K
Ramya Dinesh

System Overview

PrivateRAG AI at a Glance

What it is

- Fully offline, document-grounded AI system
- Built for privacy, safety, and auditability

Key Highlights

- 100% local execution
- Zero hallucination tolerance
- Mandatory refusal logic
- Evidence-backed answers only
- Dual interaction modes:
 - Chatbot Mode
 - Question Mode

A controlled intelligence system — not a general chatbot.

The screenshot displays the PrivateRAG AI web application. At the top, there's a navigation bar with a shield icon and the text "PRIVATE RAG AI". Below the navigation, a sidebar on the left lists "LOCAL / OFFLINE" and four main buttons: "ABOUT", "AI CHATBOT", "ASK QUESTIONS", and "DOCUMENTS". A "VALIDATION" button is also present. To the right of the sidebar, a central area shows a "Upload Knowledge Base" section with a "Drag and drop files here" input field and a "Browse files" button. Below this is a "Enter Local Folder Path:" input field with a placeholder "Enter full path to folder...". Further down, there are sections for "INDEX DOCUMENTS" and "STORAGE". On the far right, the page content includes:

- About the Tool:** Developed by Dr. Subramani.
- PrivateRAG AI:** Secure, Offline, Document-Grounded Intelligence. Describes the tool as a professional-grade AI designed for safe querying of local documents using a locally running Large Language Model (LLM). It emphasizes that unlike cloud-based tools, PrivateRAG AI runs entirely on the user's local machine, ensuring no documents, embeddings, or conversations leave the system.
- capabilities:** A list of features including Chat with your Documents (PDF, Word, Txt, MD, CSV), Complete Folder Ingestion (Recursive processing), Evidence-Based Answers (Strictly from context), Auditable Responses (With source tracking), Chatbot Mode (Conversational document exploration), and Question Mode (Direct, precise, evidence-based Q&A).
- Key Design Principles:** A list of principles: 100% Offline Execution - Zero Cloud Dependency, No Data Leakage - Privacy First, No Hallucinations - Strict Grounding, and Refusal is Correct Behavior - Never fabricates answers.
- Technology Stack:** A list of technologies used: Python (Core backend logic and orchestration), Streamlit (Local web-based user interface), Ollama (Local LLM runtime (offline execution)), LLaMA 3 / Mistral (Document-grounded answer generation), nomic-embed-text (High-quality local text embeddings), ChromaDB (Persistent local vector database), Semantic Chunking (Context-preserving text segmentation), Hybrid Retrieval (Vector similarity with re-ranking), and Offline-First Architecture (No cloud services or external APIs).

Problem Statement

Challenges with Cloud-Based AI

- Sensitive documents sent to external servers
- Privacy & compliance risks (GDPR, HIPAA)
- No transparency or audit trail
- Hallucinated or fabricated answers
- Loss of data ownership

Need

A secure AI system that works locally on private documents



Solution Overview



Introducing PrivateRAG AI

Fully offline Retrieval-Augmented Generation (RAG)

Answers strictly from uploaded documents

Refuses when evidence is insufficient

Transparent confidence & source display

Not a general-purpose chatbot

A document intelligence system

Project Objectives



Core Objectives

1

Secure interaction with private documents

2

Accurate, evidence-based answers

3

Complete elimination of hallucinations

4

Mandatory refusal when unsupported

5

100% offline execution

6

Full data sovereignty

7

User-friendly interface

8

Auditability & confidence transparency

Project Scope



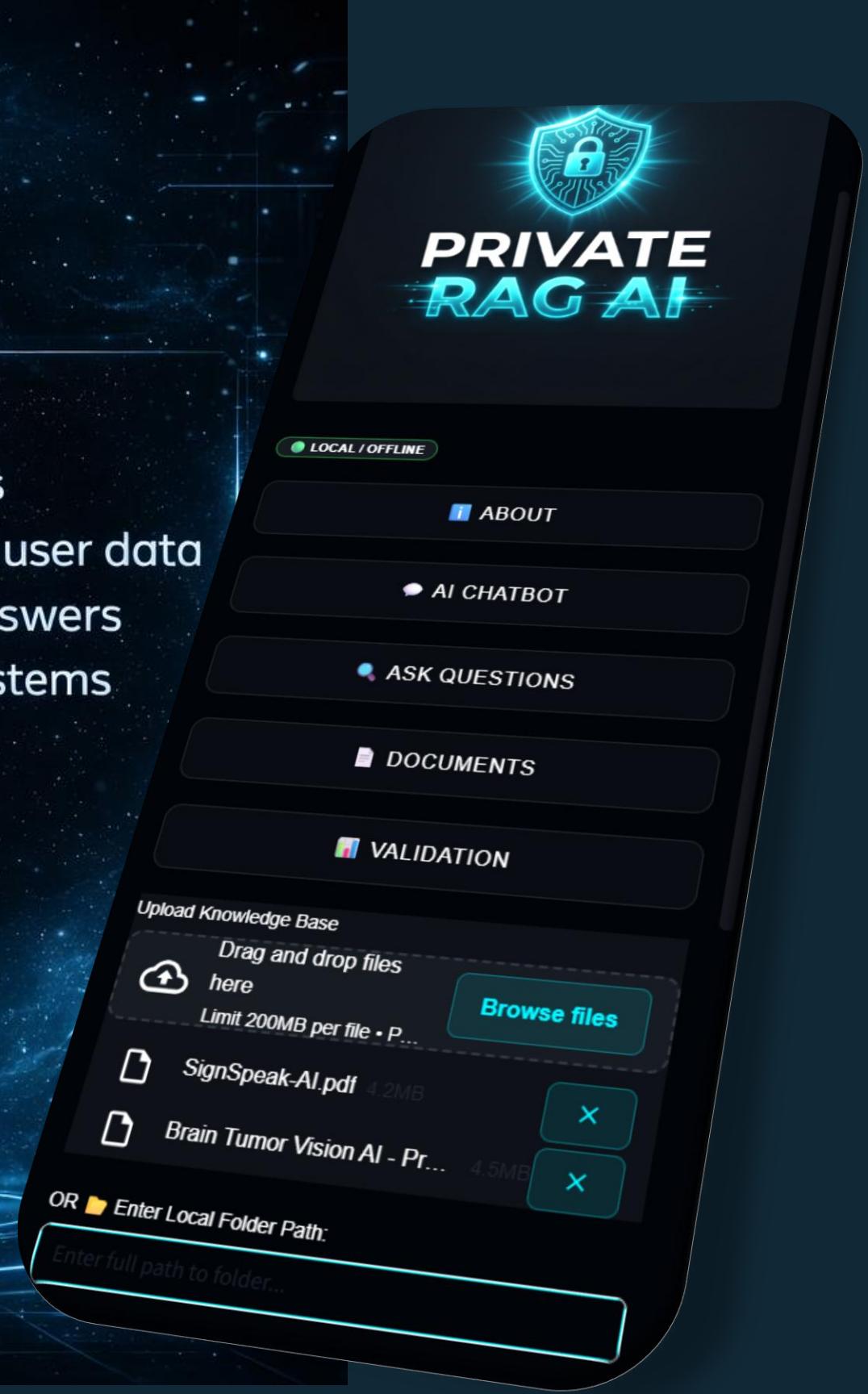
In Scope

- Local document ingestion & indexing
- Semantic vector search
- Document-grounded Q&A
- Chatbot & question interfaces
- Guardrails & refusal logic
- Confidence scoring & evaluation
- Persistent local storage



Out of Scope

- Cloud APIs or online LLMs
- Training or fine-tuning on user data
- Creative or speculative answers
- Autonomous reasoning systems



Technology Stack Used

Core Technologies

Programming & UI

- Python (Backend logic)
- Streamlit (Web-based UI)

AI & Models

- Ollama (Local LLM runtime)
- nomic-embed-text (Embeddings)
- LLaMA 3 / Mistral (Answer generation)

Data & Retrieval

- ChromaDB (Local vector database)
- Semantic chunking with overlap
- Hybrid retrieval & re-ranking

System Design

- Offline-first architecture
- Deterministic execution
- Guardrail-driven safety model

✓ No cloud services ✓ No paid APIs ✓ Fully open-source



Dual Interaction Design



Two Interaction Modes

Mode	Purpose
💬 Chatbot Mode	Conversational document exploration
❓ Question Mode	Direct, precise, evidence-based Q&A

Both modes:

- Share the same backend
- Follow identical guardrails
- Enforce zero hallucinations



PrivateRAG AI Chatbot Modes

Chatbot Mode

- Natural conversation flow
- Greeting handling
- Context-aware follow-ups
- Expandable evidence sections
- Confidence score per answer
- Clear refusal responses

The screenshot shows the PrivateRAG AI Chatbot interface. On the left, there's a sidebar with a shield logo and the text "PRIVATE RAG AI". It has buttons for "LOCAL / OFFLINE", "ABOUT", "AI CHATBOT" (which is highlighted in blue), "ASK QUESTIONS", "DOCUMENTS", and "VALIDATION". Below these are sections for "Upload Knowledge Base" with a "Drag and drop files here" area (showing "SignSpe..." and "Brain Tu...") and a "Browse files" button. There's also an "OR Enter Local Folder Path:" field with "Enter full path to folder...". At the top right, there's a "Deploy" button. The main area is titled "Chatbot" with the subtitle "Your Private, Document-Grounded AI Assistant". It shows stats: DOCUMENTS INDEXED (2), TOTAL CHUNKS (44), KNOWLEDGE BASE STATUS (ACTIVE), and LAST SYNC (17:13:45). Below this is a section titled "PrivateRAG AI Chatbot" with the subtitle "Continuous dialogue with your Private AI. Results are grounded strictly in your indexed data." It shows a conversation history:

- User: what models used in my document?
- AI: Three leading CNN architectures - VGG16, DenseNet121, and MobileNetV3-Large - were evaluated under identical training conditions.
- User: Figure 4 information?
- AI: Figure 4 shows a steady decrease in both losses and a consistent rise in accuracy, indicating stable learning, minimal overfitting, and strong generalization. The close alignment between training and validation accuracy demonstrates that dataset balancing significantly improved model performance.
- User: Deployment information?
- AI: The key deployment details and performance metrics for a deep learning-based tumor detection system using MRI scans of brain regions, particularly glioma and meningiota, are as follows:

At the bottom, there's a "Training time: Not explicitly mentioned in the provided context." note and a "Ask me anything..." input field with a right-pointing arrow. A small note at the bottom right says "PrivateRAG AI operates fully offline and answers strictly from uploaded docum".

❓ Question Mode

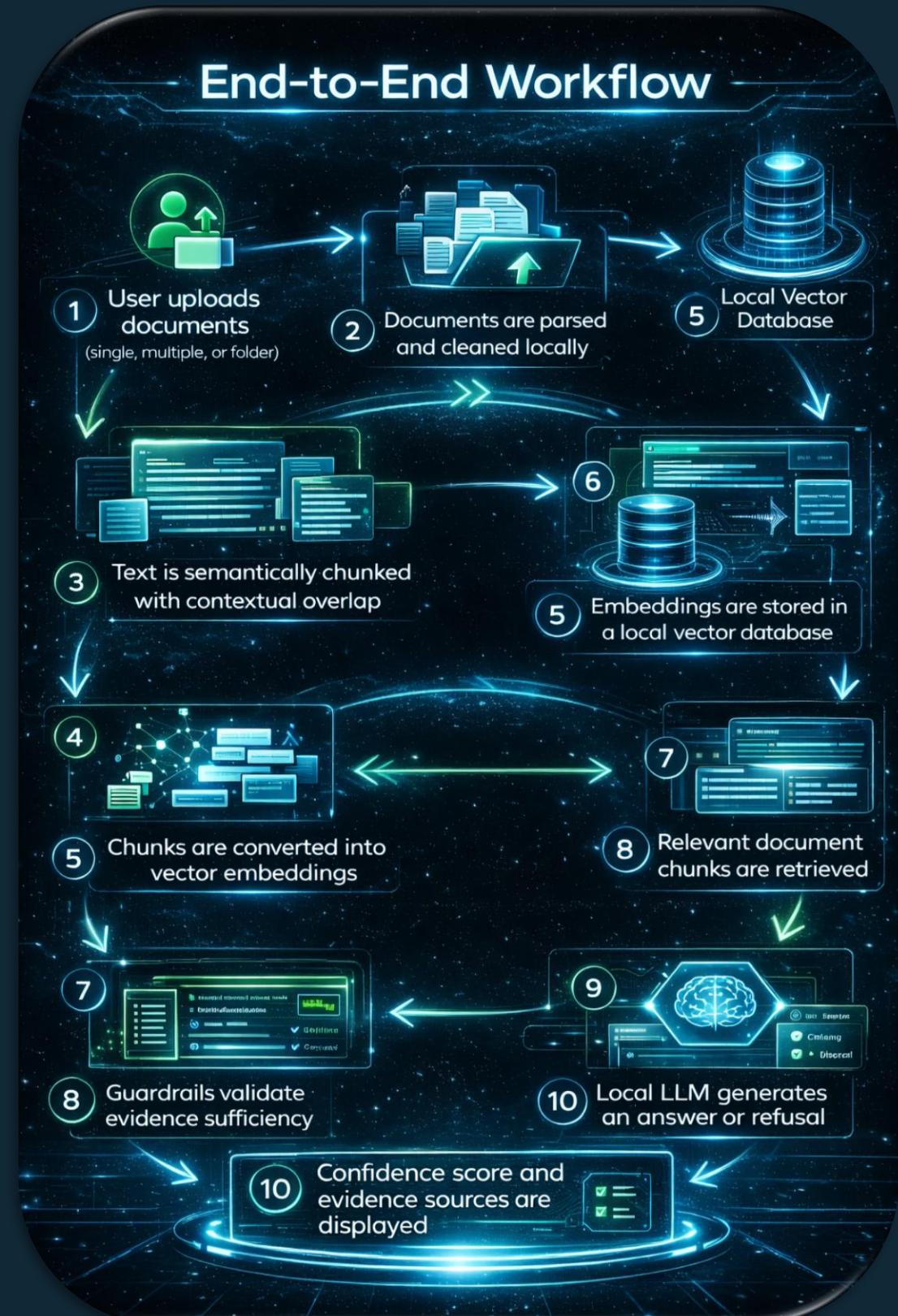
- Single-question input
- No conversational noise
- Deterministic retrieval
- Strict evidence enforcement
- Clear Answer / Refusal output
- Source-focused responses
- View Retrieved Evidence Chunks

The screenshot displays the PrivateRAG AI application interface in 'Ask Questions' mode. The top navigation bar features a shield icon and the text 'PRIVATE RAG AI'. Below the navigation are four buttons: 'ABOUT', 'AI CHATBOT', 'ASK QUESTIONS' (which is highlighted in blue), and 'DOCUMENTS'. A sidebar on the left contains sections for 'Upload Knowledge Base' (with a 'Drag and drop files here' area and a 'Browse files' button), 'INDEX DOCUMENTS', and 'VALIDATION'. The main workspace is titled 'Ask Questions' and includes a status bar with 'DOCUMENTS INDEXED 1', 'TOTAL CHUNKS 17', 'KNOWLEDGE BASE STATUS ACTIVE', and 'LAST SYNC 17:37:28'. The 'Intelligence Workspace' section contains a search bar with the query 'What is the most recent project you have worked on?' and a 'SEARCH' button. The 'Intelligence Response' section shows the AI's answer: 'The most recent project I worked on involved developing a mobile app for streamlining task management and collaboration within a team environment.' A 'View Retrieved Evidence Chunks' section lists two chunks: 'Chunk 1 (Source: Hr Interview Question.pdf)' and 'Chunk 2 (Source: Hr Interview Question.pdf)'. A circular 'TRUST LEVEL' indicator shows '99%'. At the bottom right, a note states: 'PrivateRAG AI operates fully offline and answers strictly from uploaded documents. If you need to update your knowledge base, simply upload new files or enter a local folder path.'

System Architecture

End-to-End Workflow

1. **User uploads documents (single, multiple, or folder)**
2. **Documents are parsed and cleaned locally**
3. **Text is semantically chunked with contextual overlap**
4. **Chunks are converted into vector embeddings**
5. **Embeddings are stored in a local vector database**
6. **User interacts via the chatbot interface**
7. **Relevant document chunks are retrieved**
8. **Guardrails validate evidence sufficiency**
9. **Local LLM generates an answer or refusal**
10. **Confidence score and evidence sources are displayed**



Dr. Subramani



THANK YOU

