



# PRIVATE RAG AI

**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

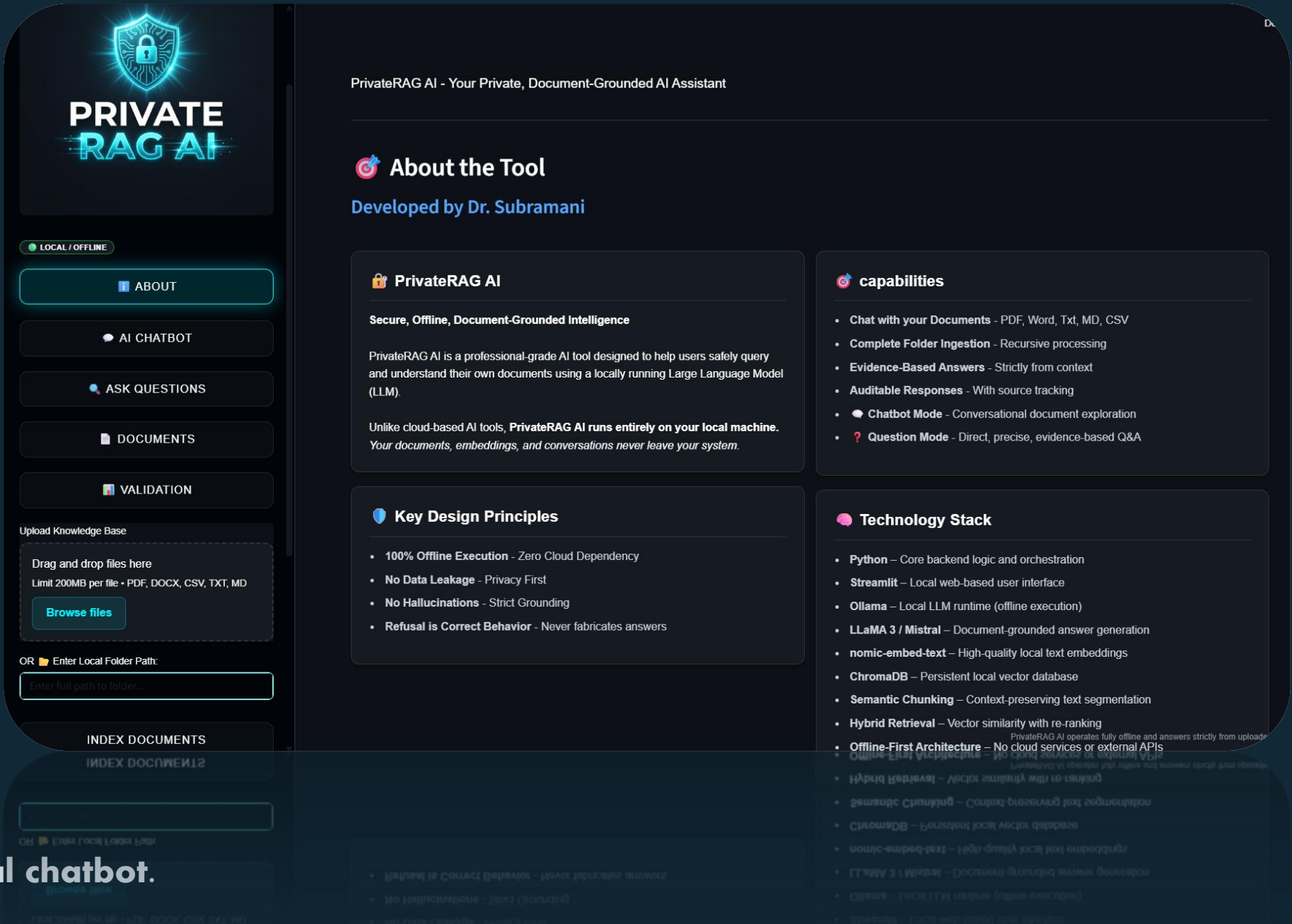
## What it is

- 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.



## 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

## Problem Statement:

Challenges with Cloud-Based AI  
Data Privacy & Control at Risk



Challenges with Cloud AI  
Data Privacy & Control at Risk



### CLOUD-BASED AI RISKS



Sensitive documents sent to external servers



Privacy & compliance risks servers



No transparency risks (GDPR, HIPA)



Hallucinated or fabricated answers

Loss a data owenship

### THE CRITICAL 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**



# 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



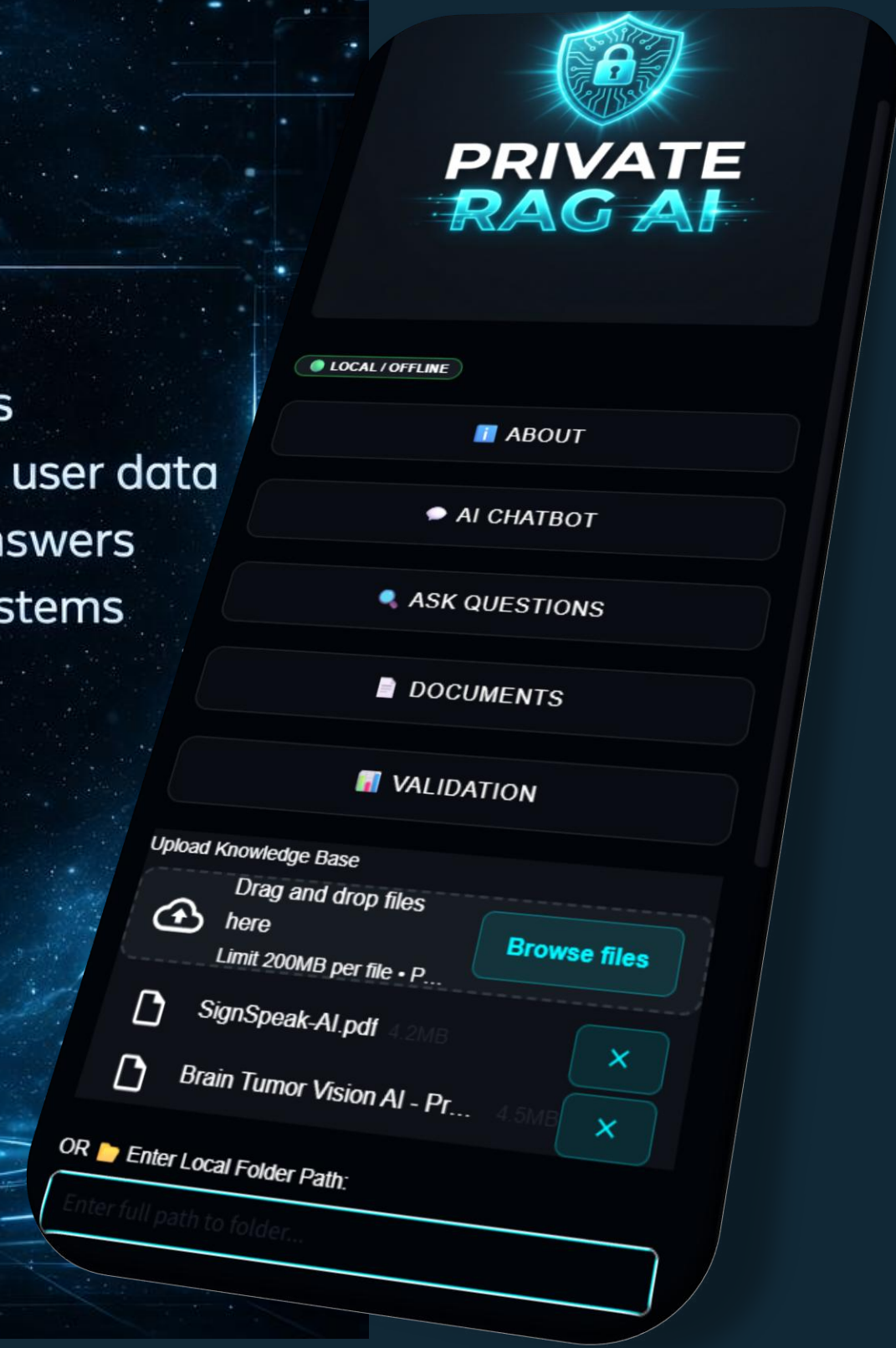
## In Scope

- 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



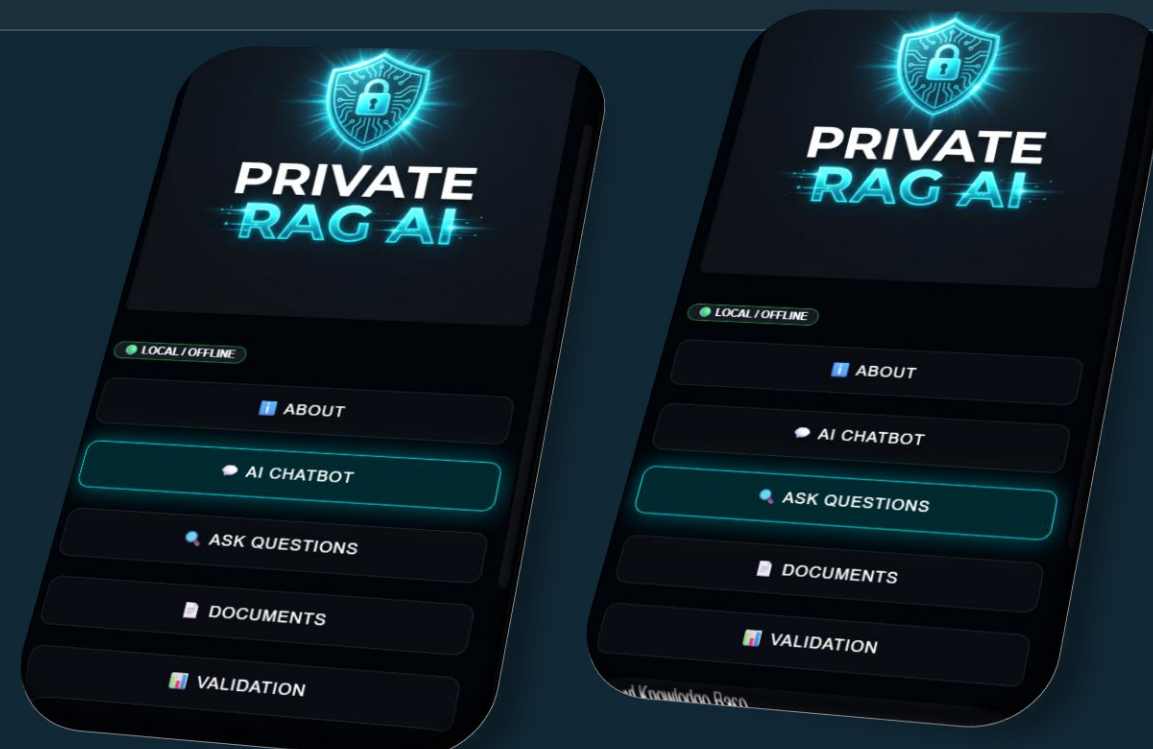
## Dual Interaction Design

# 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





# Chatbot Mode

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



PRIVATE  
RAG AI

LOCAL / OFFLINE

ABOUT

AI CHATBOT

ASK QUESTIONS

DOCUMENTS

VALIDATION

Upload Knowledge Base

Drag and drop files here  
Limit 200MB per file • PDF, ...

Browse files

SignSpe...  
4.2MB

Brain Tu...  
4.5MB

OR Enter Local Folder Path:

Enter full path to folder...

Chatbot

Deploy

DOCUMENTS INDEXED

2

TOTAL CHUNKS

44

KNOWLEDGE BASE STATUS

ACTIVE

LAST SYNC

17:13:45

PrivateRAG AI Chatbot

Continuous dialogue with your Private AI. Results are grounded strictly in your indexed data.

what models used in my document?

Three leading CNN architectures - VGG16, DenseNet121, and MobileNetV3-Large - were evaluated under identical training conditions.

Figure 4 information?

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.

Deployment information?

The key deployment details and performance metrics for a deep learning-based tumor detection system using MRI scans of brain regions, particularly glioma and meningioma, are as follows:


Ask me anything...

Training time: Not explicitly mentioned in the provided context.

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



LOCAL / OFFLINE

ABOUT

AI CHATBOT

ASK QUESTIONS

DOCUMENTS

VALIDATION


Upload Knowledge Base

Drag and drop files here  
Limit 200MB per file • PDF, ...

Browse files

Hr Interv...  
9.7MB

×

OR  Enter Local Folder Path:

INDEX DOCUMENTS

Ask Questions

Your Private, Document-Grounded AI Assistant

DOCUMENTS INDEXED  
1

TOTAL CHUNKS  
17

KNOWLEDGE BASE STATUS  
ACTIVE

LAST SYNC  
17:37:28

Intelligence Workspace

Enter Question:

What is the most recent project you have worked on?

SEARCH

CLEAR

Intelligence Response

The most recent project I worked on involved developing a mobile app for streamlining task management and collaboration within a team environment.

View Retrieved Evidence Chunks

Chunk 1 (Source: Hr Interview Question.pdf)  
The most recent project I worked on involved developing a mobile app for streamlining task management and collaboration within a team environment. 66

Chunk 2 (Source: Hr Interview Question.pdf)  
Honestly, I can't foresee any specific aspects of the job that I would dislike without experiencing them firsthand. However, I believe every role comes with its challenges, and I'm prepared to adapt and grow through any difficulties that may arise. 45 What skills would you bring to this job? I would bring strong

TRUST LEVEL  
99%

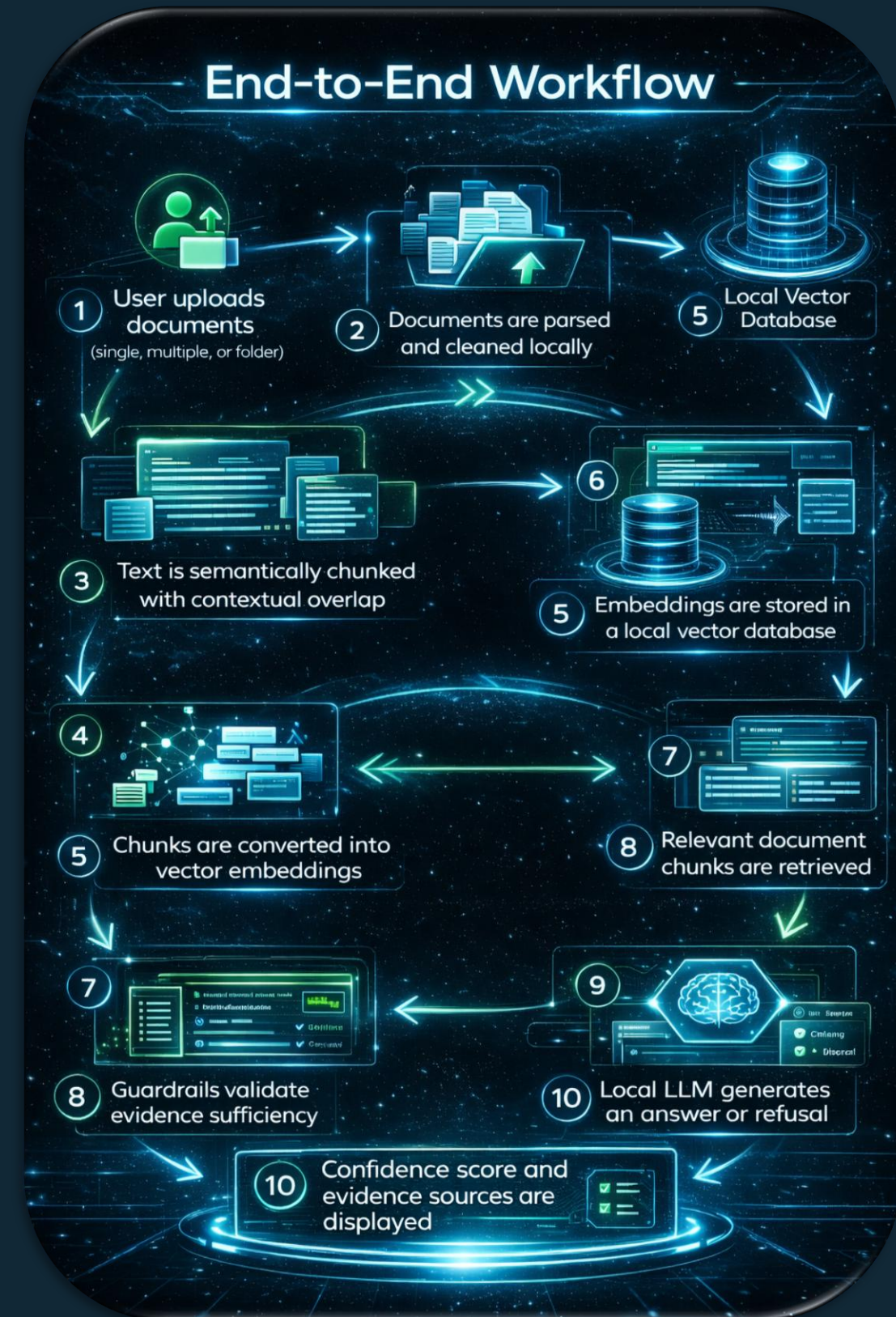
Dep

PrivateRAG AI operates fully offline and answers strictly from uploaded documents.

# 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

LinkedIn

## THANK YOU

