NexusMind

Unified Offline Multimodal RAG System

BY:
Abhishek Kumar Vishwakarma.
Bhumika kumari
Anish Kumar kannaujiya

AGENDA

1. Motivation	2. Problem Statement	3. Objective
4. System Architecture	5. Data Flow	6. Technical Approach
7. Key Features	8. Limitations & Future Enhancement	9. Conclusion

Motivation

- Existing AI systems (ChatGPT, Gemini, etc.) are **cloud-dependent** and **privacy-limited**.
- Enterprises and research environments require **offline, secure AI solutions**.
- **Goal:** Enable ChatGPT-level intelligence **without internet**, ensuring transparency and autonomy.
- *Bridging the gap between cloud AI and secure local computing.*

Problem Statement

“Design a GPU-optimized, offline multimodal RAG system capable of understanding and reasoning over text, image, and audio inputs — without relying on any cloud API.”

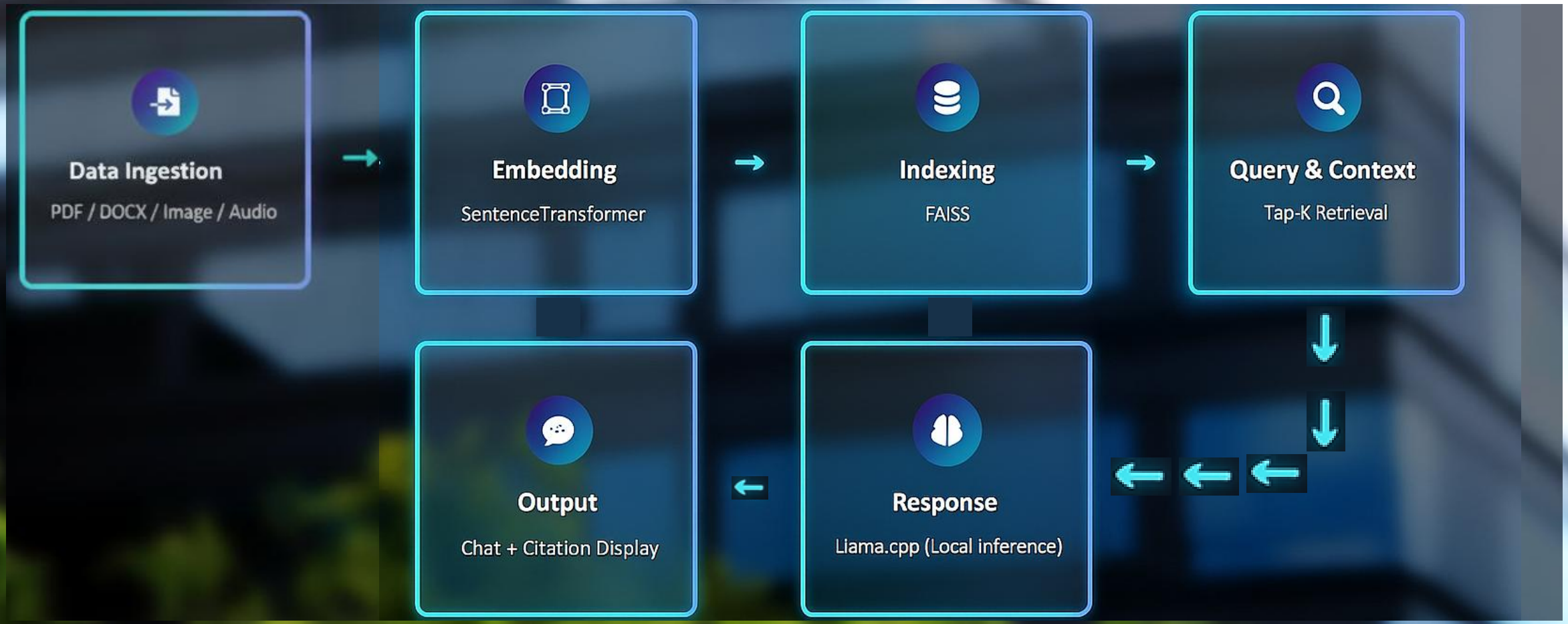
Challenges:

- Data privacy and dependency on external servers
- Multimodal data integration (PDF, image, audio)
- Transparent, source-verified responses

Objective

- Develop a **100% offline AI assistant** for multimodal understanding.
- Implement **RAG pipeline (FAISS + Llama)** for context-grounded reasoning.
- Provide **citation-based transparency** for all responses.
- Deliver an **ultra-simple single-page UI** — *no hidden abstraction; everything visible upfront.*
- Build a foundation for **future video input support** (multiframe + audio-text fusion).

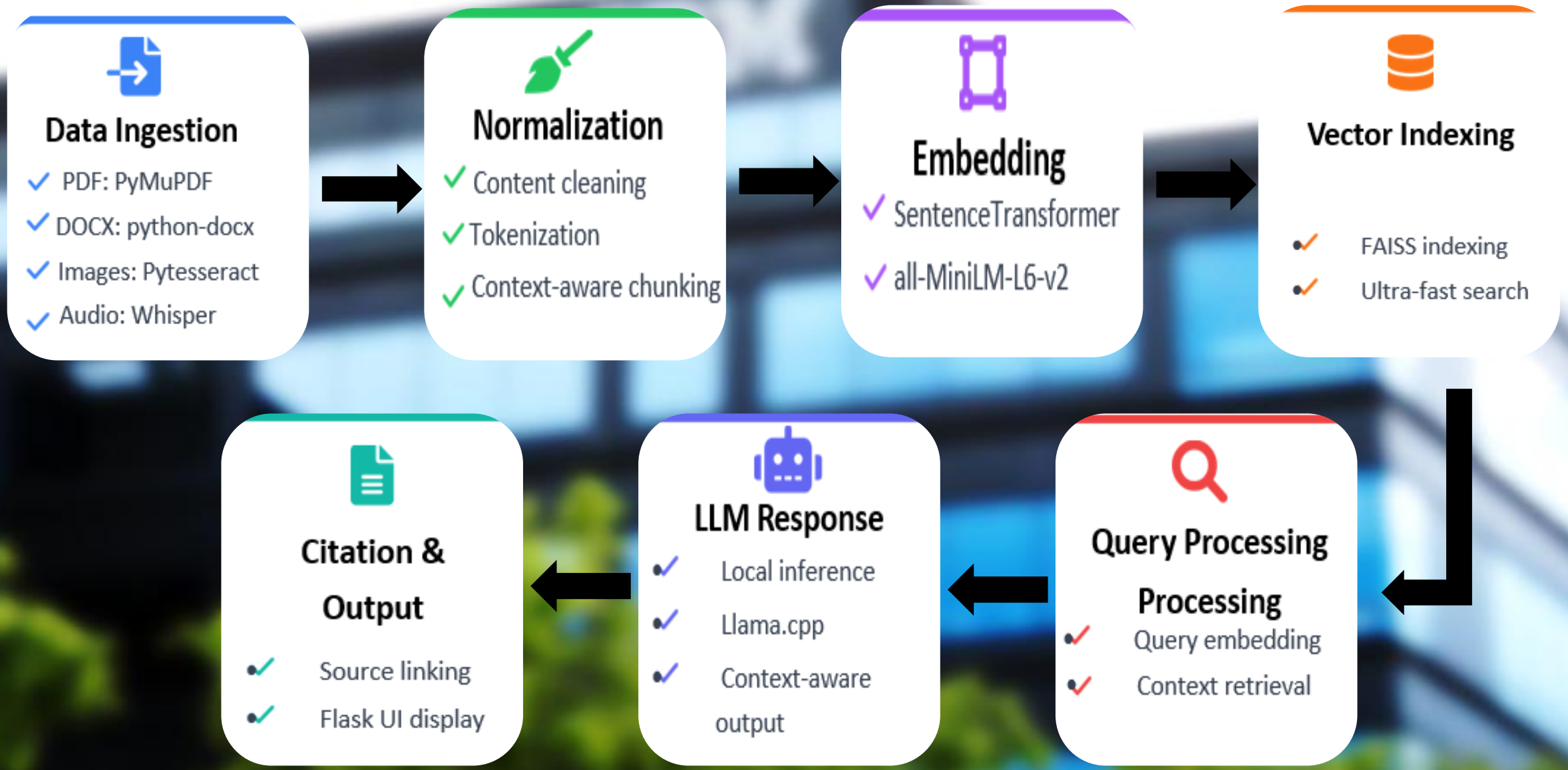
System Architecture



Data Flow


User Input
Upload files or ask query
Extraction
OCR / Whisper / DOCX Parser
Embedding
Generate embeddings
Indexing
FAISS vector indexing
Retrieval
Top-K context retrieval
Inference
Llama model response
Output
Display + citations

Technical Approach



Key Features

- 100% Offline Functionality.
- Multimodal Input (Text + Image + Audio).
- Transparent Citation System.
- Real-time CPU, GPU, RAM Monitoring.
- Adaptive Model Loading (1B/3B/8B).
- Automatic model selection by detecting hardware.
- **Single-Page, Zero-Abstraction UI — everything visible and interactive instantly.**
- Scalable for **future video-based reasoning.**

Limitations & Future Enhancement

- **Current Limitations:**
  - Requires GPU for optimal speed
  - Whisper & OCR accuracy depend on input quality
- **Future Enhancements:**
  - Integration of **video format ingestion** (frame + audio context)
  - Domain-specific fine-tuning (medical, legal, academic)
  - Multi-user collaborative mode
  - Database integration for automatic document ingestion
  - Hybrid Cloud Support (optional IBM Cloud offloading)

Conclusion

NexusMind proves that **AI intelligence doesn't need the internet.** It delivers **multimodal understanding**, **source transparency**, and **local autonomy** — all on a single workstation.

"NexusMind isn't just a project — it's the start of a secure AI revolution."