Technical Requirements: Local LLM+RAG AI Platform Project

1. Core Technology Requirements

Component	Technology	Version	Purpose
Language	Python	3.10+	App development
Framework	Llama Index	0.10.0+	RAG orchestration
Vector DB	Chroma DB	0.4.0+	Embedding storage
Metadata DB	SQLite	3.0+	Document tracking
LLM Server	Ollama	Latest (0.12.8+)	Model hosting
PDF Parser	PyPDF2	3.0.0+	Document parsing

2. Python Package Requirements

Llama Index Packages:

- llama-index-core>=0.10.0 - Core framework functionality

- llama-index-embeddings-ollama>=0.1.0 - Ollama embedding integration

- llama-index-llms-ollama>=0.1.0 - Ollama LLM integration

- llama-index-vector-stores-chroma>=0.1.0 - ChromaDB integration

Supporting Libraries:

- chromadb>=0.4.0 - Vector database

- PyPDF2>=3.0.0 - PDF processing

3. Model Requirements

Embedding Model:

- Name: nomic-embed-text

- Parameters: 137M

- Dimensions: 768

- Size: 274 MB

- Purpose: Convert text to semantic vectors.

Primary LLM:

- Name: llama3:latest (llama3.1:8b)
- Parameters: 8B
- Context Window: 8K tokens
- Size: 4.7 GB
- Purpose: Response generation
- Model Configurability: Default primary LLM is “llama3:latest” unless overridden at runtime.
- Primary and/or fallback models can be changed via environment variables.
    - Example: change to gemma model with “LLM_MODEL=gemma3:4b”

Fallback LLM (Optional LLM Model 1):

- Name: deepseek-r1:latest (deepseek-r1:8b)
- Parameters: 8B
- Context Window: 8K tokens
- Size: 5.1 GB
- Purpose: Backup if primary fails.
- Note: more RAM (20 GB+) required

Optional LLM Model 2:

- Name: gemma3:4b
- Parameters: 4B
- Context Window: 128K tokens
- Size: 2.1 GB
- Purpose: Optional model if user prefers it or primary and fallback models fail.
- Note: model from Google

4. Infrastructure Requirements

Hardware Specifications:

- Device: Laptop
- Processor: 11th Gen Intel(R) Core(TM) i5-11400H @ 2.70GHz (2.69 GHz)
- RAM: 16 GB
- System type: 64-bit operating system, x64-based processor
- Graphics card: 4 GB (multiple GPUs installed)
- Storage: 40 GB

Recommended Hardware Specifications:

- Device: Advanced computer or server system.
- CPU: Multi-core processor (8+ cores recommended for faster processing; 4+ cores minimum)
- RAM: 32 GB for comfortable operation, but 16 GB minimum (8 GB for OS/apps, 8 GB for models)
- Storage: 100 GB free space (ideally up to 200 GB+ SSD for better I/O performance)
  - 50 GB for Ollama models
  - 10 GB for vector database (scales with document count)
  - 40 GB for workspace and dependencies
- OS: Windows 10/11 (primary target)

Software:

Operating System and Tools:

- Windows 10/11 (64-bit)
- PowerShell 5.1 or later

Runtime Environment:

- Python 3.10 or later
- Pip package manager
- Python venv module (for virtual environment; .venv)

Other Notes:

- Ollama server running on localhost:11434.
- No internet connection is required after the initial setup.

Network:

Initial Setup:

- Internet connection for:
  - Python package downloads
  - Ollama software installation
  - Local model downloads (~20-30 GB total)

Operation:

- No internet connection required (100% local operation).
- Model inference and communication takes place entirely offline (via localhost).