3. System Requirements

3.1 Functional Requirements

FR-1: SOP Upload & Indexing

- Admin users shall upload SOP documents in PDF format.
- System shall chunk documents (1000 chars, 100 overlap).
- System shall generate embeddings and store them in MongoDB.

FR-2: Semantic Retrieval (RAG)

- System shall retrieve the top relevant SOP chunks using vector similarity.
- Only retrieved content shall be passed to the LLM.

FR-3: Answer Generation (Week 3 – Ollama + Phi-3)

- The system shall use **Phi-3** running locally via **Ollama**.
- The model shall generate answers strictly from retrieved SOP context.
- If context is insufficient, the model must respond:

"The provided SOP documents do not contain this information."

FR-4: Source Citation

- Every answer shall include:
    - SOP name
    - Page number
    - Section (if available)

FR-5: Streaming Response

- Responses shall be streamed token-by-token using **Server-Sent Events (SSE)**.

3.2 Non-Functional Requirements

NFR-1: Accuracy

- The system must not hallucinate beyond retrieved SOP content.

NFR-2: Performance

- Vector retrieval ≤ 500 ms
- First token response ≤ 2 seconds (local Phi-3)

NFR-3: Security

- SOP documents accessible only to authorized users.
- No external LLM API calls in Week 3 (fully local).

NFR-4: Scalability

- Must support at least 500+ pages of SOP documents.