## **Functional Scope**

### **1. Arabic-Native Semantic Search (Chat with Document)**

* Support **Arabic-language queries** directly (no translation pipeline).
* Enable **upload** of diverse document formats:
  + PDF
  + DOCX
  + PPT / PPTX
  + HTML
* **Parse and extract text**, perform chunking for long documents.
* **Generate embeddings** from Arabic text for semantic understanding.
* Implement **semantic search** to retrieve and match relevant chunks.
* **Chat interface** for querying uploaded content in Arabic.
* Ensure **accurate, context-aware** answers are returned in Arabic.
* No English-based intermediate reasoning (Arabic-native processing only).

### **2. Metadata Extraction & Tagging**

* Extract and retain metadata including:
  + **Document titles**
  + **Page numbers**
  + **Sections or headers**
  + **Keywords**
* Enable semantic chunk mapping with this metadata to enhance contextual relevance in answers.

### **3. Arabic LLM Integration**

* Replace the current multilingual LLaMA 3.1 with **Arabic-native LLMs**. Candidates include:
  + **Jais**
  + **Qwen 2.5**
* Tasks:
  + Model fine-tuning / prompt optimization.
  + Hyperparameter tuning for better instruction-following.
  + Add guardrails to **minimize hallucinations** and improve factuality.
  + Evaluate response **fluency, relevance, and accuracy** in Arabic.

### **4. Streamlit-Based Testing Interface (Temporary UI)**

* Build a minimal but functional frontend for internal testing using **Streamlit**:
  + File upload panel with basic **document stats** (size, pages, word count).
  + Interactive **chat interface**.
  + Language toggle (Arabic/English) only for debugging/testing.
  + Status indicators (loading, chunking, embedding, etc.).
* To be replaced in later phases by a production-ready web app.

## **Deliverables**

* Arabic-native semantic search system (query-to-chunk pipeline).
* Document upload & parsing module with metadata tagging.
* Integrated Arabic LLM with optimized inference settings.
* Temporary Streamlit UI for testing .
* Documentation (system architecture, deployment instructions, LLM setup).
* Evaluation benchmarks (qualitative and quantitative metrics).