Hexagon Smart Completions Intelligent Agent – Specification & Architecture (PDF-Only Scope)

(Scope narrowed to static PDF manuals only; live-data API and Excel ingestion deferred to a later phase.)

---

1. Vision & Use-Case

Deliver a **conversational assistant** that allows field engineers, supervisors, and administrators to query six key Smart Completions PDF manuals—including all embedded screenshots—and receive traceable, role-aware answers within seconds.

Role	Abilities
Admin	Upload / replace PDF revisions, manage users, view all content
Standard User	Chat with the assistant, view PDFs mapped to their role

---

2. Front-End Experience (React)

- Single-page app (React + Tailwind + shadcn/ui) with **futuristic side menu**: ① *Chat* ② *Document Library* (six base PDFs, plus any admin-uploaded revisions).
- **JWT status indicator** shows session validity.

---

3. Authentication & Authorisation

- **Clerk (OAuth 2 / JWT)** issues signed tokens containing `role` and permitted-document list.
- FastAPI middleware validates JWT on every call.

---

4. Document Ingestion & Storage (PDF-Only)

1. **Revision Handling**\ – Admin selects an existing manual or uploads a newer PDF revision. Old revision archived; new one processed.
2. **Screenshot & Visual Parsing**\ – **PaddleOCR** extracts raw text from images.\ – **LayoutLM** recognises bounding boxes, arrows, and highlighted regions, outputting enriched JSON (`{text, bbox, page, annotation_type}`).
3. **Chunking & Embedding**\ – Combine standard PDF text with OCR output → semantic chunks (\~300 tokens).\ – Generate embeddings with `all-MiniLM-L12` (or OpenAI ADA) and store in **Weaviate**.

4. **Databases**
5. **Weaviate** — vector storage for embeddings + minimal IDs.
6. **PostgreSQL** — raw PDF blob, revision metadata, LayoutLM JSON, upload history, chat logs.

   No separate DB is required for OCR / LayoutLM; enriched JSON is stored in Postgres as `jsonb`.

---

5. Agentic-RAG Pipeline (LangChain)

```
User ➜ FastAPI ➜ Aggregator Agent (LangChain)
          ├──► Retriever  (Weaviate)
          ├──► OCR/LayoutLM tool (on-demand)
          └──► Planning & Memory  (ReACT / CoT)
```

- **Agentic behaviour**: decides whether OCR tool is needed (e.g., user asks "what does the circled note on page 14 say?").
- Conversation memory: short-term in Redis, long-term in Postgres.
- Role context injected via prompt templates.
- Final prompt is sent to a **local open-source LLM served via Ollama** (e.g., Mixtral-8×7B or Llama-3-8B).

---

6. Non-Functional Requirements

Category	Target
Latency	≤ 3 s (95th pct)
Uptime	99 % (Mon-Fri 08:00-18:00)
OCR accuracy	≥ 90 % on clear screenshots (fallback: provide original image link)
Security	JWT exp = 60 min, refresh = 14 days; OWASP top-10 tested
Load	100 concurrent chat sessions

OCR / LayoutLM limitations & mitigation

- Low-resolution or noisy images can lower OCR accuracy ⇒ pre-processing (contrast, binarise).
- Hand-written notes may be missed ⇒ admins can attach manual text annotations.

---

7. Tech Stack

Layer	Technology
Front-end	React, Tailwind, shadcn/ui
Auth	Clerk (OAuth 2 / JWT)
API & Logic	FastAPI, LangChain 0.1+
Vector store	Weaviate
Metadata DB	PostgreSQL 15
OCR	PaddleOCR 2.6
Layout Understanding	LayoutLM-v3 (Hugging Face)
LLM	**Ollama-served open-source model** (Mixtral-8×7B / Llama-3-8B / Gemma-7B)
DevOps	**Docker** (build + run) · Docker Compose (local orchestration)

---

8. Development Phases (PDF-Only Scope)

1. **Foundations & Brain-storming** (Week 1)\ Architecture, spec, repo, Clerk skeleton.
2. **Ingestion & Storage Layer** (Week 2)\ PDF parsing POC, Postgres schema, Weaviate up.
3. **RAG & Visual Reasoning** (Week 3)\ Embedding service, LangChain skeleton, PaddleOCR + LayoutLM integration.
4. **API & Front-End Integration** (Week 4)\ FastAPI chat endpoints, JWT middleware, React chat UI.
5. **Testing & CI/CD** (Week 5)\ Integration tests, security hardening, **Docker image build & container-hardening pipeline**.
6. **Monitoring, Polish & Delivery** (Week 6)\ Prometheus metrics, UI polish, demo & acceptance.

---

9. Acceptance Criteria (PDF Scope)

- Query cycle ≤ 3 s; answer cites PDF revision + page.
- OCR answers reference highlighted image at least 8/10 times on test set.
- Admin can upload new PDF revision and see it reflected in search within 5 minutes.
- Role restrictions verified with 10 test users.
- CI pipeline passes > 90 % tests.

---

10. Glossary (excerpt)

Term	Meaning
Agentic RAG	RAG where an agent plans tool calls (retriever, OCR) instead of static chain.
Chunk	200–300 token text segment stored as an embedding vector.
LayoutLM	Transformer that encodes text and 2-D layout from document images.
Weaviate	Open-source vector DB with hybrid search.

---

11. Open Issues (TBD)

ID	Topic
TBD-1	Final open-source model selection (Mixtral vs. Llama-3 vs. Gemma)
TBD-2	AWS vs. Azure for hosting
TBD-3	Image pre-processing parameters for noisy screenshots