# RAG Document Ingestion & Retrieval — Minimal POC Scope Note

## 1. Purpose of the POC

The goal of this POC is to **prove safe, auditable document ingestion** for large, mixed-content PDFs.

This POC validates **trust, traceability, and decision-driven parsing** — not advanced intelligence.

> Core question answered by this POC: **Can we ingest a complex PDF and produce reliable, reviewable extracted data without hallucination?**

---

## 2. What This POC Explicitly Proves

- Documents are uniquely registered and traceable
- Parsing decisions are driven by document structure, not guesswork
- Deterministic extraction is preferred over ML
- All extracted content has lineage and confidence
- SQL acts as the authoritative source of truth

---

## 3. In-Scope Components (Minimal & Mandatory)

### 3.1 Document Registry

**Scope** - Accept PDF input - Generate immutable `document_id` - Store document metadata and ingestion status

**Why Included** - Enables reprocessing, auditing, and lifecycle tracking

---

### 3.2 Probe Engine

**Scope** - Read PDF metadata - Sample a small set of pages - Detect: - Presence of text layer - Image density - Table likelihood

**Outputs** - Structural signals - Simple complexity score

**Why Included** - Prevents blind parsing - Demonstrates decision-first architecture

---

### 3.3 Deterministic Parsing Engine

**Scope** - Text extraction for text-heavy pages - Table extraction for table-heavy pages - Page-level classification

**Explicit Limitation** - No OCR - No ML-based content generation

**Why Included** - Ensures trust-first extraction - Minimizes hallucination risk

---

### 3.4 SQL as Authoritative Store

**Scope** - Persist documents, pages, and extracted blocks - Store extraction method and confidence score

**Core Tables** - `documents` - `pages` - `blocks`

**Why Included** - Enables full lineage - Acts as audit log and replay source

---

## 4. Out-of-Scope Components (Intentionally Excluded)

The following are **deliberately excluded** from the POC:

- OCR pipelines
- Knowledge Graph construction
- Vector embeddings
- LLM-based answering
- User-facing chat interface

**Rationale**

> Intelligence is added only after correctness is proven.

---

## 5. End-to-End POC Flow

```
PDF
  ↓
Document Registry
  ↓
Probe Engine
  ↓
Deterministic Parsing
```

```
  ↓
SQL (Source of Truth)
```

No data bypasses SQL.

## 6. Quality Signals Produced by the POC

For each extracted block: - Extraction method (pipeline) - Confidence score (heuristic) - Page number - Content type

These signals will later drive promotion to vector or graph systems.

## 7. Success Criteria (Definition of Done)

The POC is considered successful if:

- A large PDF (500+ pages) is ingested end-to-end
- Structural probing completes in seconds
- Majority of pages are parsed deterministically
- All extracted data is queryable from SQL
- Any skipped or failed pages are clearly recorded

## 8. How This POC Extends Forward

This POC directly enables: - OCR integration for low-confidence pages - Vector database ingestion for trusted blocks - Knowledge Graph construction for structured documents

No redesign is required — only extension.

## 9. One-Sentence Summary

**“This POC proves that we can ingest complex PDFs safely, make informed parsing decisions, and store auditable results — before adding ML or retrieval layers.”**

*End of Minimal POC Scope Note.*