RAG Document Ingestion & Retrieval — Phase-Level Verification Note

1. Purpose

Define **minimal, practical verification checks** for each POC phase to ensure that the system behaves as expected without implementing full formal testing.

This aligns with industry practice: POC validates functionality and trust, not full correctness.

---

2. Verification Strategy Principle

- Each phase outputs **inspectable artifacts**
- Stop-and-verify approach: do not move to the next phase until current phase is verified
- Verification focuses on **observability and expected behavior**
- Manual inspection is acceptable at POC stage

---

3. Phase-Level Verification Checklist

3.1 Phase 1 — Document Ingestion / Registry

Goal: Verify the document is registered correctly

Verification Signals: - Document exists in `documents` table with UUID - Metadata matches original file (page count, file size, source) - Status is `received`

Artifacts for Inspection: - SQL record - Stored PDF location (local/S3)

Manual Checks: - Open the PDF from storage, compare with metadata

---

3.2 Phase 2 — Probe Engine / Structural Analysis

Goal: Verify that the document is correctly probed and structural parameters are detected

Verification Signals: - Pages sampled correctly (start, middle, end) - Page type classifications (text-heavy, table-heavy, scanned) are reasonable - Complexity score is computed - Number of images, tables detected roughly matches expectations

Artifacts for Inspection: - Probe output JSON or log - Sampled page info

Manual Checks: - Inspect 3–5 pages to confirm detected type matches actual content

---

3.3 Phase 3 — Deterministic Parsing

Goal: Verify deterministic extraction for text and tables

Verification Signals: - Extracted blocks are created for each page - Extraction method is recorded (`pipeline`) - Confidence score is assigned (heuristic) - Page classification matches actual content

Artifacts for Inspection: - Extracted text files or table CSVs - SQL `blocks` table entries

Manual Checks: - Open 2–3 blocks per type and confirm content matches page

---

3.4 Phase 4 — SQL Storage Verification

Goal: Verify all extracted data is correctly persisted

Verification Signals: - Document, pages, blocks tables populated - Lineage preserved (page_number, block_id, extraction method) - No missing blocks or duplicate entries

Artifacts for Inspection: - SQL queries checking counts and IDs - Sample block content

Manual Checks: - Cross-check block content with original PDF page

---

3.5 Phase 5 — Optional Vector / Retrieval Hooks (POC Light)

Goal: Verify readiness for vector-based retrieval (if included in POC)

Verification Signals: - Number of vectors == number of valid blocks - Basic similarity query returns expected chunk(s) - No missing embeddings

Artifacts for Inspection: - Vector DB entries or local embedding files - Sample queries and top results

Manual Checks: - Pick 1–2 keywords, verify the top chunk makes sense

---

4. Notes on Verification Philosophy

- **Manual inspection is expected at POC stage**
- **No full automation or test suite required**
- Focus on **detecting failures early** rather than exhaustive correctness
- Each phase **must emit artifacts** (logs, JSON, CSV, SQL) to allow easy inspection

- Allows the POC to scale later into full automated testing

---

5. Recommended Workflow

1. Complete phase 1 → verify → continue
2. Complete phase 2 → verify → continue
3. Complete phase 3 → verify → continue
4. Complete phase 4 → verify → continue
5. Optional phase 5 → verify → continue

   This is exactly how industry builds trustworthy POCs — stop, inspect, confirm, proceed.

End of Phase-Level Verification Note.