AI SECURITY ENGINEERING FIELD GUIDE · 2026

Chapter 04 · RAG Security

Standalone learning module for LMS delivery and required reading.

FORMAT	USE	SCOPE	AUDIENCE
Standalone PDF	LMS reading module	Single chapter	Learners

CHAPTER 04

RAG Security

> Retrieval is a data access decision. Treat it as one.

> Retrieval is a data access decision. Treat it as one.
>
> FIELD GUIDE

FIELD GUIDE: RAG BOUNDARY REVIEW

Inspect corpus sources, ACLs, chunk metadata, embedding pipelines, retrieval logs, citation behavior, and tenant boundaries before trusting an answer.

This domain covers retrieval-augmented generation systems: corpus intake, source authorization, chunking, embeddings, metadata, vector indexes, retrieval filters, context assembly, citation integrity, deletion propagation, and retrieval telemetry. It matters when an AI system answers from private documents, customer records, knowledge bases, tickets, code, policies, contracts, or tenant data.

Do not start by asking whether the final answer leaked a string. Start by asking which chunks entered the context window and whether the user was allowed to retrieve them. If unauthorized context reached the model, the boundary failed even when the answer was vague.

FIELD USE

Run RAG review as a data-access review. Inspect corpus authority, chunk metadata, retrieval filters, deletion propagation, and logs before judging answer quality.

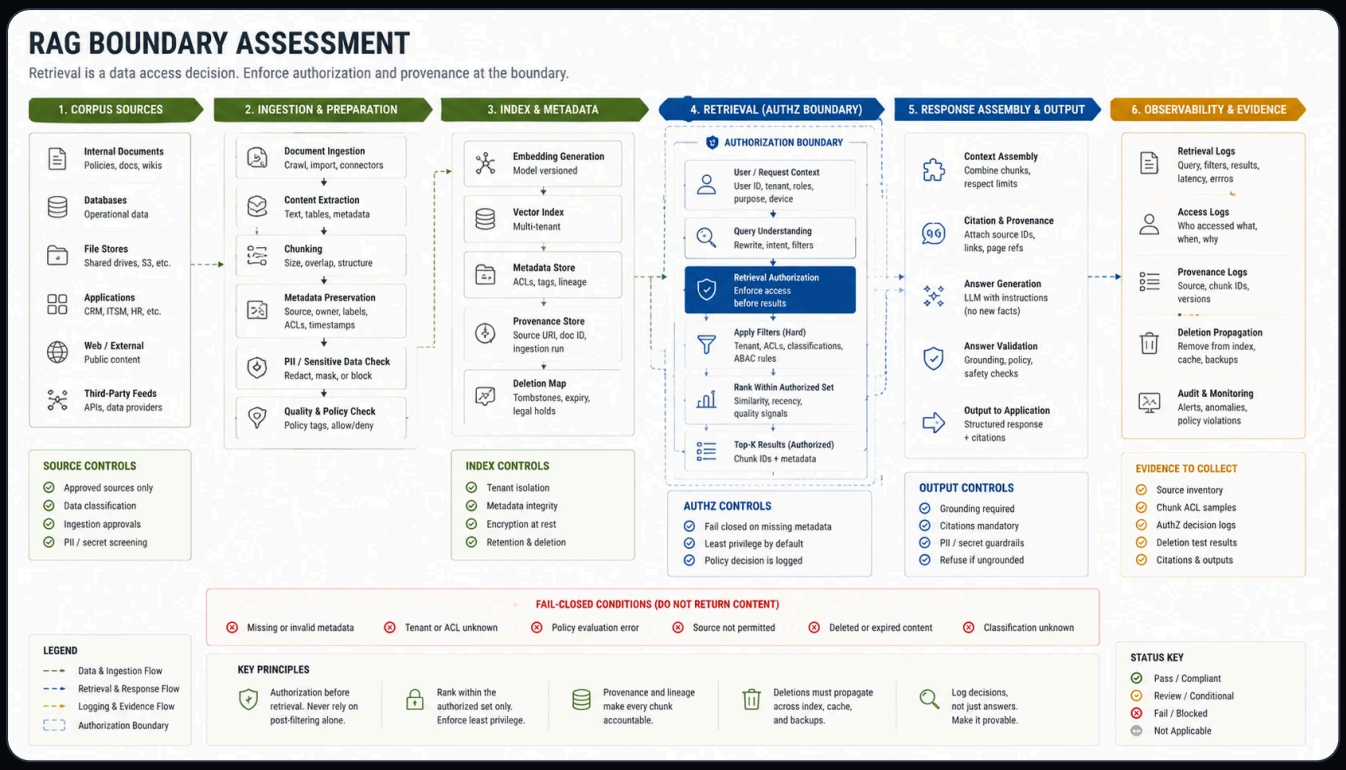

FIGURE 1: FIGURE 5: RAG BOUNDARY ASSESSMENT MAP COVERING CORPUS SOURCES, ACL LAYERS, CHUNKING PIPELINE, RETRIEVAL FILTERS, AND CONTEXT ASSEMBLY

WHAT BREAKS HERE?

- Retrieval leakage exposes private or high-privilege context.
- Stale permissions remain in chunks after source access changes.
- A poisoned corpus inserts adversarial instructions or false facts.
- Missing provenance hides which source influenced an answer.
- Weak tenant boundaries let one customer retrieve another customer's content.
- Citations imply support from sources the model did not actually use.

WHAT TO INSPECT

- Corpus sources, owners, write permissions, classification, and review gates.
- ACLs at source, ingestion, chunk, index, and retrieval layers.
- Chunk metadata: source ID, tenant, owner, classification, permission, version, freshness, and deletion state.
- Embedding pipeline, re-index jobs, deletion jobs, and permission-change propagation.
- Retrieval logs, citation behavior, source filtering, and context assembly.
- Query rewriting, hybrid search, reranking, caching, and post-retrieval filtering.

WHAT TO ASK

- Who can write to each corpus source?
- Which permission check runs before a chunk enters context?
- Does retrieval rank only inside the authorized result set?
- What happens when a source document is deleted or reclassified?
- Can the team reconstruct which chunks influenced a specific answer?
- How are poisoned documents detected, quarantined, and removed?
- Which retrieval failures block launch, and which create monitored backlog items?

WHAT TO TEST

- Unauthorized retrieval with low-privilege users.
- Cross-tenant queries against shared indexes.
- Poisoned document ingestion and retrieval.
- Stale permission scenarios after role, group, or document ACL changes.
- Deleted document retrieval after re-index delay.
- Misleading, missing, or hallucinated citation behavior.
- Query rewriting or reranking that drops authorization filters.

CONTROLS AND GUARDRAILS

- Retrieval-time authorization before context assembly.
- Hard tenant isolation or mandatory tenant filters before ranking.
- Source filtering, provenance checks, chunk metadata preservation, and corpus hygiene.
- Ingestion review for writable or low-trust sources.
- Deletion and permission-change propagation with verification.
- Retrieval logging with chunk IDs, authorization decisions, and citation support.
- Fail-closed behavior when metadata, source identity, or ACL state is missing.

EVIDENCE TO COLLECT

- Retrieval test results.
- Data-source map.
- Authorization notes.
- Chunk metadata schema.
- Corpus hygiene records.
- Deletion propagation checks.
- Citation validation results.
- Fail-closed test evidence.
- Remediation backlog.

OUTPUT ARTIFACTS

- RAG boundary assessment.
- Leakage findings.
- Evidence map.
- Corpus source inventory.
- Retrieval authorization test plan.

RELATED SERVICES AND WORKBENCH TOOLS

TYPE	RELATED PATHS
Workbench	RAG Test Harness, Runtime Proxy, Threat Canvas
Services	AI Product Security Assessment, AI Red Team Validation, Product Security Baseline
Handbook	RAG Authorization, Data Exposure and Privacy, Logging and Telemetry

ARTIFACT: RAG EVIDENCE PACK

Produce a source map, ACL review, retrieval tests, citation checks, deletion proof, and leakage backlog.