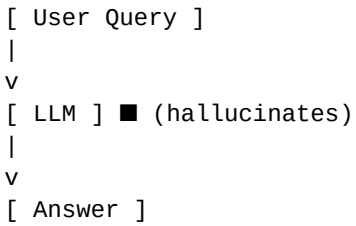
RAG from First Principles
A Practical, Safe RAG System (Days 1–6)

This book documents the step-by-step construction of a Retrieval-Augmented Generation (RAG) system from first principles. Each chapter introduces one irreversible invariant that improves correctness, safety, and auditability. By Day 6, the system enforces semantic entailment between claims and evidence.

Day 1 — RAG Foundations

Architecture Diagram



Detailed Explanation

Large Language Models generate text probabilistically and do not have guaranteed access to factual data.

Without external grounding, models hallucinate confidently.

Day 1 introduces the separation of knowledge retrieval from language generation.

Common Failure Modes

- Hallucinated facts

- Confident but incorrect answers

- No audit trail

Defenses Introduced

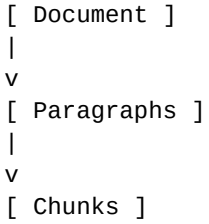
- External document corpus

- Explicit retrieval step

- Evidence-first mindset

Day 2 — Document → Chunks

Architecture Diagram



Detailed Explanation

Documents are split into semantically coherent chunks.

Chunks are the atomic units of retrieval.

Poor chunking directly degrades retrieval quality.

Common Failure Modes

- Overly large chunks dilute relevance

- Overly small chunks lose meaning

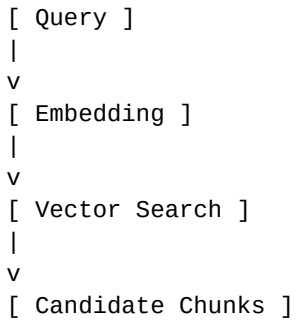
Defenses Introduced

- Paragraph-aware splitting

- Minimum and maximum chunk size rules

Day 3 — Retrieval

Architecture Diagram



Detailed Explanation

The system retrieves candidate chunks using semantic similarity.

Recall is prioritized to avoid missing evidence.

Noise is expected at this stage.

Common Failure Modes

- Missing relevant evidence

- Retrieving irrelevant chunks

Defenses Introduced

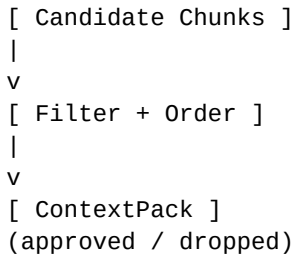
- Top-K retrieval

- Distance thresholds

- Deferred filtering

Day 4 — Retrieval → Context

Architecture Diagram



Detailed Explanation

Retrieved chunks are filtered into approved and dropped sets.

ContextPack becomes the official evidence bundle.

Reasons for dropping chunks are preserved.

Common Failure Modes

- Context pollution

- Hidden evidence removal

Defenses Introduced

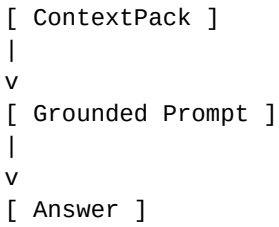
- Explicit approval rules

- Audit metadata

- Hard boundary before generation

Day 5 — Context → Answer

Architecture Diagram



Detailed Explanation

The model is prompted using only approved context.

If no context exists, the system must refuse.

Citations are required for factual claims.

Common Failure Modes

- Answering without evidence

- Citation hallucination

Defenses Introduced

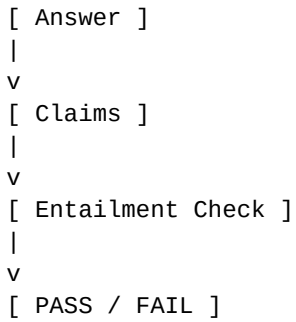
- Refusal behavior

- Citation enforcement tests

- Hallucination injection tests

Day 6 — Semantic Validation (Claim ↔ Context Entailment)

Architecture Diagram



Detailed Explanation

Answers are decomposed into atomic factual claims.

Each claim is checked against approved context.

Only ENTAILED claims are allowed.

Common Failure Modes

- Unsupported factual claims

- Partial hallucinations

- Misleading answers

Defenses Introduced

- Fail-closed entailment

- Claim-level verification

- Dropped-context isolation