

RAG from First Principles

Days 1–6: A Complete, Safe RAG Pipeline

This document provides a complete, end-to-end explanation of a Retrieval-Augmented Generation (RAG) system built from first principles. Each day adds one irreversible safety or correctness layer. By Day 6, the system guarantees that every factual claim in an answer is semantically supported by evidence.

Day 1 — RAG Foundations

Goal

Understand why naïve LLM prompting fails and why retrieval must be explicit and controlled.

Core Idea

LLMs are probabilistic text engines, not databases.

RAG separates knowledge access from language generation.

Flow

Query → Document Corpus → Retrieval → Context → Answer

Example

User asks: “How long do refunds take?”

Without RAG: the model guesses.

With RAG: the answer must come from retrieved policy text.

Invariant Introduced

Answers must be grounded in external evidence.

Day 2 — Document → Chunks

Goal

Transform raw documents into clean, retrievable units.

Chunking Rules

Chunks must be semantically coherent.

Chunks must fit within model context limits.

Flow

Raw Document → Paragraph Split → Merge/Trim → Chunk objects

Invariant Introduced

Retrieval operates on chunks, not entire documents.

Day 3 — Retrieval

Goal

Select candidate chunks relevant to a user query.

Core Mechanism

Embedding-based similarity search.

Recall is prioritized over precision.

Flow

Query → Embedding → Vector Search → Candidate Chunks

Invariant Introduced

Retrieval narrows the evidence space but does not validate truth.

Day 4 — Retrieval → Context

Goal

Turn raw retrieval output into a controlled ContextPack.

ContextPack

Approved chunks (usable evidence).

Dropped chunks (explicitly excluded).

Flow

Candidate Chunks → Filtering → Ordering → ContextPack

Invariant Introduced

Only approved chunks may influence answers.

Day 5 — Context → Answer

Goal

Generate an answer strictly grounded in the ContextPack.

Rules

No context → refusal.

Citations required for factual statements.

Flow

ContextPack → Grounded Prompt → Answer

Invariant Introduced

The model may only speak from provided evidence.

Day 6 — Semantic Validation (Claim \leftrightarrow Context Entailment)

Goal

Guarantee that every factual claim in the answer is supported by context.

Claim Extraction

Answers are decomposed into atomic factual claims.

Entailment

Each claim is checked against approved context.

The LLM acts only as a semantic verifier.

Fail ■ Closed Rules

No context \rightarrow NOT_ENTAILED.

UNKNOWN \rightarrow failure.

Any single failure invalidates the answer.

Final Invariant

ENTAILED is the only success state.