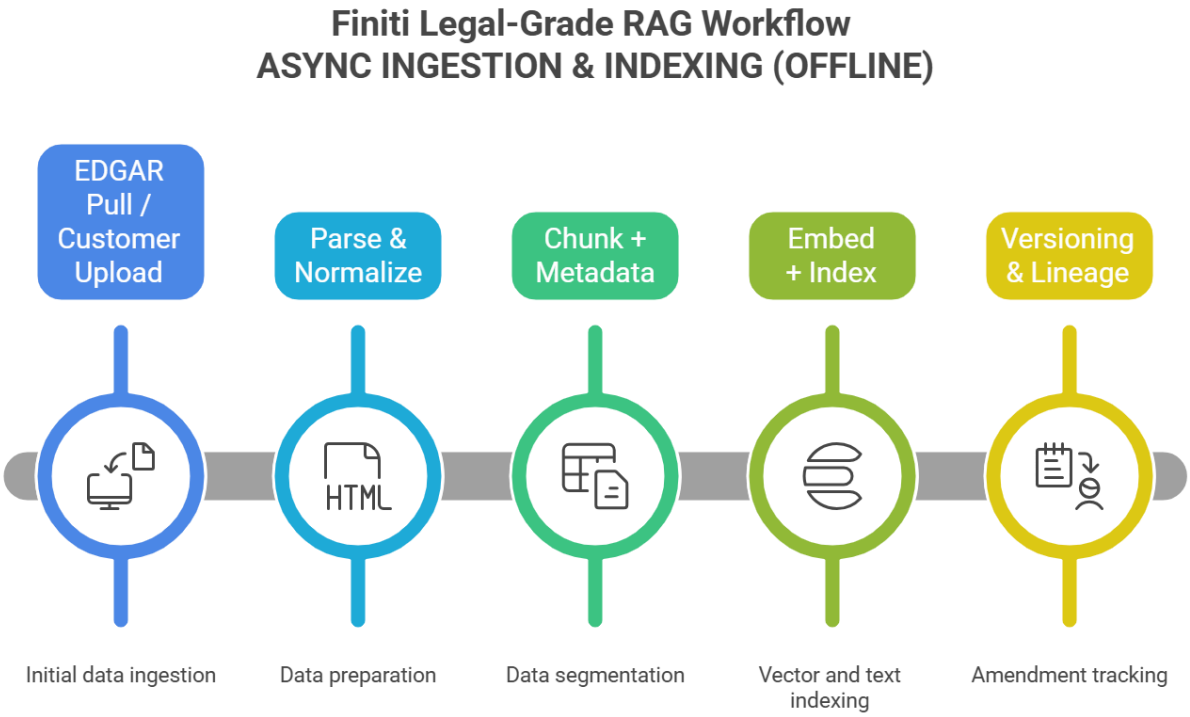
Architecture Diagram



LANE 1 — ASYNC INGESTION & INDEXING (OFFLINE)

1. EDGAR Pull / Customer Upload
    - idempotency_key = tenant_id + filing_version_id
2. Parse & Normalize
    - inputs: HTML / PDF / XBRL (table-aware extraction; OCR if needed)
3. Chunk + Metadata
    - paragraph/table chunks
    - metadata: section_key, filing_type, period_end, company_id, peer tags
    - span pointers: char_start/char_end (and page/bbox if PDF)
4. Embed + Index
    - embeddings (vector index) + BM25/full-text index
    - filters: section_key, filing_type, period_end, company_id, peer_set_id
5. Versioning & Lineage
    - amendment chain via superseded_by_id (e.g., 10-K → 10-K/A)
    - replay roots: stable references to source artifacts

Finiti Legal-Grade RAG Workflow
DRAFT / REVIEW / BENCHMARK (ONLINE)

API Gateway

Auth, rate limiting, observability

Request Parse & Classify

Intent, constraints

Peer Set Resolver

Chooses peer companies

Hybrid Retrieval

Structure-based and semantic search

Evidence Pack Builder

Selects top chunks and spans

Grounded Comparison Engine

Claim extraction and checks

Schema-Constrained Generation (JSON)

Generates draft/review output

Citation Validator & Guardrails

Span-level coverage validation

Confidence Scoring

Combines retrieval and verification signals

Review & Sign-off

Human approval gates

LANE 2 — ONLINE DRAFT / REVIEW / BENCHMARK (REQUEST-TIME)

1. API Gateway
    - auth, rate limiting, observability (tracing + metrics)
2. Request Parse & Classify
    - intent: draft / review / benchmark
    - constraints: section_key, period window, filing_type, tone/format rules
3. Peer Set Resolver
    - chooses peer companies using rules/manual selection
    - stores audit reason + configuration
4. Hybrid Retrieval
    - structure-based section matching (Item alignment)
    - BM25 keyword search + vector semantic search
    - pre-filters: section_key, filing_type, period_end, peer_set
5. Evidence Pack Builder
    - selects top chunks + spans + relevance scores
    - includes doc metadata and version references
6. Grounded Comparison Engine
    - claim extraction: split text into atomic factual claims
    - deterministic checks: numbers/dates/units/period alignment
    - contradiction detection: flag conflicting evidence
    - outputs structured findings + required citations per claim
7. Schema-Constrained Generation (JSON)
    - generates draft/review output using a strict JSON schema
    - rule: each claim must cite >= 1 span (chunk_id + start/end)
8. Citation Validator & Guardrails
    - span-level coverage validation
    - block/flag: UNSUPPORTED_CLAIM, CONTRADICTED, AMBIGUOUS, OUTDATED_SOURCE
    - prompt-injection hardening: retrieved text treated as untrusted data
9. Confidence Scoring
    - combines: retrieval score + rerank score + citation coverage + verifier signals
    - returns calibrated score or qualitative label (High/Med/Low)
10. Review & Sign-off

- human approval gates; export controls

11. Output

- JSON + HTML report (benchmarks + citations + issues)

CROSS-CUTTING

Audit & Reproducibility

- Append-only audit log with tamper-evident hash chain (tenant-scoped)
- Every step logs:
    - retrieval_trace (queries, filters, scores, chosen chunks)
    - model_config (model name, temperature, safety settings)
    - prompt_version (template + variables)
    - outputs + edits + exports
- workflow_runs enable full replay for any draft/review decision

Reliability & Operations

- Workflow engine / state machine for step transitions
- retries with exponential backoff
- DLQ for failed jobs after N retries (triage + replay)
- key SLOs: P95 latency, cost per draft, DLQ rate, citation coverage %, unsupported claim rate