

Fabric — dual-mode RAG

one retrieval contract · two stores · chosen once, validated, fail-closed

fabric_kb — durable ground truth

append-only knowledge memory · the RAG index is a derived overlay, rebuilt from it — never the source

Sources ingested into the index

run_history · knowledge_base · agents · docs
skills (_src_skills) · admin local files / folders / web pages

deployment_config — the mode gate

accepts (local, local) | (hybrid, postgres) — else fail closed

LOCAL — everything on-device

Mac workstation · private · \$0 · never auto-syncs to a hosted store

Store — JSONL

rag_repository.LocalRagRepository · one line per chunk

Embeddings — local Ollama

nomic-embed-text · 768-dim · on-device only

Retrieval — brute-force cosine

scan all chunks · exact top-k · no index server

Freshness — atomic promotion

build to temp → atomic swap · a half-built index is never read

Profile binding

FORGE_PROFILE=local → FORGE_RAG_STORE=local
the single-workstation default

HYBRID — shared PostgreSQL + pgvector

control plane · multi-workspace · isolated · fail-closed by default

Store — pgvector

PostgreSQL · vector(768) · ANN index

EmbeddingContract

pins model + dimensions + normalization — no silent drift

Isolation — QueryScope + RLS

SET LOCAL · a workspace sees only its own rows

Safety — fail-closed · tombstoned

revocation tombstones · sampled re-embed audit detects drift

Profile binding

FORGE_PROFILE=control-plane → RAG_STORE=postgres
shared across a studio team

One retrieval contract above both stores — the caller never knows which answered

Event Spine facts may ground retrieval; evidence stays trust-tagged and never becomes execution authority