

GOVERNED TEMPLATE RENDERING REFERENCE: EMBEDDING & VECTOR STRATEGY DEFAULT V2 (DRAFT)

definition b6e86011...d61e

EMBEDDING & VECTOR STRATEGY · PROJECT ARTIFACT

SPECIRA

Embedding & Vector Strategy: Dispatch Modernization

Meridian Field Services: instructional example, not project evidence

DRAFT · WATERMARK POLICY: DRAFT_ONLY

template embedding_vector_strategy v2 · pack: specira_default_delivery

embeddings of PII are PII; labels propagate; vectors rot silently

§1 Vectorized Content Inventory

mandatory 1 decision 1 evidence rule

validators: every_embedded_content_class_cites_a_classification_label · every_vector_source_field_cites_its_data_model_field · every_embedding_model_choice_documents_migration_cost

One index at pilot: override-note search for the exception review (the US-3 surface: "find past overrides with similar reasons"). The label propagates: the vectors and the index are as sensitive as their sources.

Field	Value
Source fields (cited)	<u>Assignment.reason_code · Assignment.note^[1]</u>
Class · propagated	CONFIDENTIAL worker-conduct records (<u>labels^[2]</u>); inversion reconstructs note text, so store access mirrors the source's: operations managers only (<u>permission model^[3]</u>)
Chunking	One note, one vector; notes run 20 to 200 tokens; splitting buys nothing; rationale recorded
Metadata	assignment id · hub · reason_code · created_at (the filter fields)

Model & store decisions

Model: the compact 384-dimension embedding model over the 1,536-dimension option. Quality on the held-out set differs by two points of Recall@10; storage and query cost by four times. **Migration cost, stated:** re-embedding the pilot corpus (60 thousand notes) is an afternoon and single-digit dollars, so the cheap model is reversible ([embedding model^{\[4\]}](#)).

Store: beside the relational data. The corpus sits far under the dedicated-store threshold and the filters are relational joins; the dedicated-store trigger is a ten-million-vector corpus or a hybrid-search need ([vector store^{\[4\]}](#)).

§2 Retrieval Design & Baselines

mandatory

1 decision

1 evidence rule

validators: every_retrieval_config_declares_top_k_and_threshold · hybrid_and_rerank_stages_carry_triggers · every_retrieval_path_has_a_recall_or_mrr_baseline

Declared, not defaulted, and staged additions earn their place with triggers, so the future conversation has criteria. Anecdote is not a baseline.

Config (exception-review search)	Value
Top-k · threshold	20 · similarity 0.72
Metadata filters · FIRST	hub · date range · reason_code: relational, cheap, and they cut the ca
Hybrid (dense + lexical)	DEFERRED : reason codes are already a filter column, so the exact-identifier case is free-text queries missing coded concepts past 10% of review sessions
Reranking	DEFERRED : first-stage precision meets the baseline; the trigger is Recall@10 below

Baseline: Recall@10 0.91, MRR 0.78 on the golden query set (40 real review questions from M. Chen's team plus 60 synthetic, human-reviewed, [register^{\[5\]}](#)); release gate: Recall@10 at 0.85. Retrieval-only, distinct from any generation eval, and none exists; the surface is retrieval-only at pilot ([patterns^{\[6\]}](#)).

§3 Freshness & Invalidation

mandatory

1 decision

1 evidence rule

validators: re_embed_sla_is_numeric · every_deleted_source_has_a_tombstone_rule · rot_check_wired_to_observability

A stale note scores as high as a fresh one; similarity cannot tell. This section exists because of that.

Rule	Value
Freshness · event-driven	A new override note embeds on write (the queue the worker already consumes, building blocks ^[7]); re-embed SLA: 5 minutes; the nightly re-index is the backstop only
Tombstones	The seven-year assignment archive transition (retention rules ^[2]) soft-deletes the vector (flagged, skipped at query time); asynchronous cleanup weekly; the retention rule drives the index, never the other way
Rot check	Nightly comparison of source write timestamps against vector timestamps; a gap past the SLA raises a ticket (alerting ^[8])

§4 Open Questions

mandatory 1 decision

Question	Owner	Answer by	Blocks
Do synthetic golden queries need union-visible review? (The corpus is worker-conduct text; shared thread with security open questions ^[9] .)	N. Duval	Sep 30, 2026	The golden set's refresh rule only

Refs References & Package Contents

In the Specira workspace

specira	[1] Data model: Assignment.reason_code, Assignment.note app.specira.ai/projects/dispatch-modernization/artifacts/data-model
specira	[2] Data classification: labels, retention rules app.specira.ai/projects/dispatch-modernization/artifacts/data-classification
specira	[3] Auth & authz policy: the permission model the store mirrors app.specira.ai/projects/dispatch-modernization/artifacts/auth-authz-policy#permission
specira	[4] Decision log: embedding model, vector store decisions app.specira.ai/projects/dispatch-modernization/artifacts/decision-log

- specira

[5] Test cases: the golden query set as a registered asset [app.specira.ai/projects/dispatch-modernization/artifacts/test-cases](#)
-
- specira

[6] AI orchestration blueprint: the retrieval-only pattern row [app.specira.ai/projects/dispatch-modernization/artifacts/ai-orchestration-blueprint](#)
-
- specira

[7] Architecture: building blocks (the queue the embed hook rides) [app.specira.ai/projects/dispatch-modernization/artifacts/architecture](#)
-
- specira

[8] Observability: the rot-check ticket wiring [app.specira.ai/projects/dispatch-modernization/artifacts/observability-monitoring#alerting](#)
-
- specira

[9] Security requirements: the shared union-review thread [app.specira.ai/projects/dispatch-modernization/artifacts/security-requirements](#)
-

Generated by Specira · template embedding_vector_strategy v2 (draft) · pack specira_default_delivery lineage b6e86011...d61e · page 1 of 4

SAMPLE
seeded demo data · specira.ai