

GOVERNED TEMPLATE RENDERING REFERENCE: MODEL SELECTION DEFAULT V2 (DRAFT)

definition c294ea23...4b76

MODEL SELECTION & EVALUATION · PROJECT ARTIFACT

SPECIRA

Model Selection: Dispatch Modernization

Meridian Field Services: instructional example, not project evidence

DRAFT · WATERMARK POLICY: DRAFT_ONLY

template model_selection v2 · pack: specira_default_delivery

evidence, never vibes: pinned versions, event triggers, classed fallbacks

§1 Model Matrix

mandatory

1 decision

1 evidence rule

validators: every_model_choice_carries_task_eval_evidence · every_model_role_has_a_pinned_version_not_a_floating_alias · triggers_are_events_with_thresholds

One model role at pilot: RANKING (order the eligible technician set for a job). Blended cost per resolved task, latency at the committed percentile.

Candidate	Rank agreement	Blended cost / task	p95 @ pilot concurrency	Verdict
Small-fast model	0.87	0.4 ¢	620 ms: inside the 800 ms budget (NFR-P2 ^[1])	SE
Mid-tier model	0.89	2.1 ¢	1,340 ms: blows the budget	reje

Verdict rationale: two points of agreement do not buy a busted latency budget and five times the cost; logged (ranking-model decision^[2]), eval run attached (eval runs^[3]). **Pin:** the versioned model id held in platform config; the floating alias is banned from production.

Drift tripwire: the nightly eval runs the pinned version and the provider's newest candidate; a two-point agreement drop opens a ticket (alerting^[4]).

Re-evaluation trigger (event)	What re-runs
Provider deprecation notice on the pinned version	Full matrix
Any release in the small-fast class	Candidate row
Agreement drift below 0.84 on the nightly	Full matrix + golden-set review

§2 Evaluation Criteria

mandatory

1 decision

1 evidence rule

validators: golden_sets_have_provenance_and_refresh_rule · every_eval_metric_is_matched_to_task_type · no_judge_use_stated_as_claim_where_absent · every_fine_tune_decision_documents_cheaper_levers_tried_first

Golden set: 500 historical assignments with the dispatcher's actual choice and the eligible set at decision time, drawn from the legacy system's seven-year history ([migration inventory](#)^[5]); refreshed quarterly with pilot decisions; the set is a test asset ([register](#)^[6]) cited by id, not duplicated.

Metric fit: rank agreement (the model's top three contains the dispatcher's choice), a deterministic overlap measure; no rubric needed for a closed output space. **LLM-as-judge: NOT USED, as a claim:** the task has ground truth (what the dispatcher chose), so a judge would add bias without adding signal.

Lever ladder: prompt engineering carried the quality bar alone ([PR-01](#)^[7]); retrieval adds nothing (the eligible set is already the context); fine-tuning **rejected with the ladder documented:** 0.87 agreement meets the pilot bar ([success measures](#)^[8]) without tuning cost or drift risk.

§3 Fallback Routing

mandatory

1 decision

1 evidence rule

validators: every_fallback_classed_equivalent_or_degraded · degraded_fallbacks_carry_signoff · routing_rules_are_config_not_code

Role	Primary	Fallback	Class	Trig
ranking	Pinned small-fast model via the gateway	No model: the full technician list with the degradation notice (suggestions_unavailable ^[10])	DEGRADED, declared & signed ; quality delta: the dispatcher ranks manually, the pre-AI baseline; signed A. Reyes	2 s

Role	Primary	Fallback	Class	Trig
	(INT-4 ^[9])			

No capability-equivalent fallback exists at pilot (single provider, the documented SPOF, [provider strategy^{\[11\]}](#)), stated, not hidden. Routing lives in gateway config; a chain edit is a config change, no deploy.

§4 Open Questions

mandatory

1 decision

Question	Owner	Answer by	Blocks
Does the quarterly golden-set refresh reweight toward pilot decisions? (The legacy set encodes pre-suggestion habits, the thing the pilot is changing.)	A. Reyes with the QA owner	Oct 10, 2026	The eval-criteria refresh rule only

Refs References & Package Contents

In this export package

package [3] Eval runs: matrix scoring exports ./ai/eval-runs/

In the Specira workspace

- specira [1] NFR catalog: NFR-P2 (the committed latency budget) [app.specira.ai/projects/dispatch-modernization/artifacts/nfr-catalog](#)
- specira [2] Decision log: the ranking-model decision [app.specira.ai/projects/dispatch-modernization/artifacts/decision-log](#)
- specira [4] Observability: the drift-tripwire ticket wiring [app.specira.ai/projects/dispatch-modernization/artifacts/observability-monitoring#alerting](#)
- specira [5] Data migration: the assignment-history inventory the golden set draws from [app.specira.ai/projects/dispatch-modernization/artifacts/data-migration#inventory](#)
- specira [6] Test cases: the golden set as a registered test asset [app.specira.ai/projects/dispatch-modernization/artifacts/test-cases](#)
- specira [7] Prompt engineering spec: PR-01 [app.specira.ai/projects/dispatch-modernization/artifacts/prompt-engineering-spec#pr-01](#)

specira	[8] BRD: the pilot success measures app.specira.ai/projects/dispatch-modernization/artifacts/brd#objectives
specira	[9] Integration inventory: INT-4 model provider app.specira.ai/projects/dispatch-modernization/artifacts/integration-inventory#int-4
specira	[10] API contracts: the degradation contract app.specira.ai/projects/dispatch-modernization/artifacts/api-contracts#errors
specira	[11] AI orchestration blueprint: error recovery, provider strategy app.specira.ai/projects/dispatch-modernization/artifacts/ai-orchestration-blueprint

Generated by Specira · template model_selection v2 (draft) · pack specira_default_delivery lineage c294ea23...4b76 · page 1 of 4

SAMPLE
seeded demo data · specira.ai