

GOVERNED TEMPLATE RENDERING REFERENCE: AI ORCHESTRATION BLUEPRINT DEFAULT V2 (DRAFT)

definition 2d63546b...e56b

AI ORCHESTRATION BLUEPRINT · PROJECT ARTIFACT

SPECIRA

AI Orchestration Blueprint: Dispatch Modernization

Meridian Field Services: instructional example, not project evidence

DRAFT · WATERMARK POLICY: DRAFT_ONLY

template ai_orchestration_blueprint v2 · pack: specira_default_delivery

workflows beat agents until proven; the justification cuts both ways

§1 Agent Architecture

mandatory

1 decision

1 evidence rule

validators: every_feature_has_a_named_orchestration_pattern · every_agent_or_multiagent_choice_carries_a_complexity_justification · every_guardrail_layer_is_mapped_to_input_or_output_stage

Cheapest pattern first: single call, chain, router, parallelization, orchestrator-workers, evaluator-optimizer, agent, multi-agent. Anything agentic carries a written justification; so does the rejection.

Feature	Pattern	Model	Rationale	Fallback
Suggestion ranking	single constrained call	ranking role ^[1]	One bounded task (rank the eligible set), no tool use, no iteration	Full-list degradation (suggestions_unavail
Exception-review search	single retrieval call	none (retrieval-only at pilot)	The US-3 surface over the override-	Keyword filter fallback

Feature	Pattern	Model	Rationale	Fallback
			note index (retrieval ^[3])	

Complexity justification, cutting both ways

An agent framework was **REJECTED** for suggestion ranking: the task has fixed inputs, a closed output space, and a two-second budget; agent tiers trade latency and cost for autonomy the task cannot use; multi-agent burns 4 to 220x tokens on work this shape ([orchestration pattern](#)^[4]).

Guardrails, two layers. Input: the intake validation rules ([validation](#)^[5]). Output: schema-constrained ranked ids from the eligible set only; constrained decoding per [ADR-5](#)^[4], the strongest output guardrail. **Observability:** every call a span with model, input and output tokens, operation, and provider, landing in the LLM trace stream ([instrumentation](#)^[6]).

§2 LLM Provider Strategy

mandatory

1 decision

1 evidence rule

validators: every_provider_dependency_has_a_fallback_or_documented_spof ·
provider_egress_cites_security_and_classification

Decision: single provider behind the platform gateway

Why considered: one AI feature at pilot; a second provider's resilience buys less than its gateway-maintenance and eval-duplication costs; direct integration hard-wires the vendor shape into application code. **Findings:** single-provider-with-gateway wins; the gateway is the egress enforcement point; anonymized codes only cross ([SR-014](#)^[7], [INT-4](#)^[8]); switching exposure stated: prompts and evals are provider-portable by design, the gateway isolates the API shape ([provider strategy](#)^[4]).

SPOF, documented and owned: provider outage degrades suggestions to the full list (never blocks dispatch); accepted by A. Reyes; revisit trigger is a second AI feature entering the roadmap.

§3 Prompt Design Patterns

mandatory

1 decision

1 evidence rule

validators: no_prompt_content_in_blueprint · constrained_outputs_cite_their_decision

Orchestration-level shape only; the prompt CONTENT lives in the prompt spec, cited never restated.

The ranking call carries the system framing at the gateway layer; job attributes and candidate codes enter as typed variables; the instruction hierarchy is architectural: the system framing is not overridable by job-note content ([injection defenses](#)^[9]). **Structured output:** the ranking feature feeds a machine (the board), so schema-constrained decoding is enforced at the gateway, citing [ADR-5](#)^[4]; no free-text output path exists at pilot. **Prompt content:** one entry, [PR-01](#)^[9].

§4 State Management

mandatory

1 decision

1 evidence rule

validators: every_checkpointed_workflow_declares_its_persistence_backend · stateless_features_carry_rationale

Statelessness is a claim with rationale; the absence of a checkpointer is a decision, not an omission.

Stateless, as a claim: suggestion ranking is stateless per-request; every call carries its full context (job attributes plus the eligible set), nothing resumes, nothing checkpoints. Rationale: the task completes inside one budget window and the board re-requests on staleness ([board read model](#)^[10]). The exception-review search is likewise stateless. No checkpointer, no persistence backend, no conversation state, and the first conversational feature on the roadmap re-opens this section.

§5 Cost Architecture

mandatory

1 decision

1 evidence rule

validators: every_feature_has_a_cost_per_task_budget · budget_breach_rule_wired_to_observability

Cost per task is designed, not discovered on the invoice.

Feature	Budget / task	Caching	Measured actual
Suggestion ranking	1,800 tokens (1,500 in / 300 out)	Prompt-prefix caching on the stable system framing (the eligible-set payload	1,410 tokens mean over pi

Feature	Budget / task	Caching	Measured actual
		varies, the framing does not)	

Semantic caching: rejected at pilot: dispatch contexts rarely repeat within the cache window; *recorded so the rejection is a decision*. Cost per thousand suggestions is budgeted and tracked on the value-case dashboard ([dashboards^{\[6\]}](#)). **Breach rule:** a day above 120% of budget raises a ticket to A. Reyes; the alert is wired ([alert catalog^{\[6\]}](#)).

§6 Error Recovery & Fallbacks

mandatory

1 decision

1 evidence rule

validators: every_external_call_has_a_timeout_and_retry_budget ·
degrade_paths_cite_api_degradation_contracts

Rung	Suggestion ranking
Timeout	2 s, the boundary the process flow owns (FLW-01^[11])
Retry budget	One retry inside the window, none after; a late suggestion is discarded, never late-applie
Fallback chain	DEGRADED, declared : the full technician list with the degradation notice (suggestions
Circuit breaker	5 consecutive provider failures → open 5 min; suggestions degrade without
Posture	Degrade before error: dispatch never blocks on the model

The failure paths themselves are owned by the process flows and edge-case analysis; cited, not remodeled.

§7 Open Questions

mandatory

1 decision

Question	Owner	Answer by	Blocks
Does the exception-review search add a generation step (summarize similar overrides) or stay retrieval-only through pilot?	A. Reyes	Sep 26, 2026	The pattern table's second row only

Refs References & Package Contents

In the Specira workspace

specira	[1] Model selection: the ranking role, eval runs app.specira.ai/projects/dispatch-modernization/artifacts/model-selection
specira	[2] API contracts: suggestions_unavailable degradation contract app.specira.ai/projects/dispatch-modernization/artifacts/api-contracts#errors
specira	[3] Embedding & vector strategy: the override-note retrieval app.specira.ai/projects/dispatch-modernization/artifacts/embedding-vector-strategy
specira	[4] Decision log: orchestration pattern, provider strategy, ADR-5 app.specira.ai/projects/dispatch-modernization/artifacts/decision-log
specira	[5] Business rules: validation family app.specira.ai/projects/dispatch-modernization/artifacts/business-rules#validation
specira	[6] Observability: LLM instrumentation, dashboards, alert catalog app.specira.ai/projects/dispatch-modernization/artifacts/observability-monitoring
specira	[7] Security requirements: SR-014 gateway-only egress app.specira.ai/projects/dispatch-modernization/artifacts/security-requirements#sr-014
specira	[8] Integration inventory: INT-4 model provider app.specira.ai/projects/dispatch-modernization/artifacts/integration-inventory#int-4
specira	[9] Prompt engineering spec: PR-01, injection defenses app.specira.ai/projects/dispatch-modernization/artifacts/prompt-engineering-spec
specira	[10] Architecture: the board read model app.specira.ai/projects/dispatch-modernization/artifacts/architecture
specira	[11] Process flows: FLW-01 boundary timer app.specira.ai/projects/dispatch-modernization/artifacts/process-flows#flw-01
specira	[12] Edge case analysis: the AI families app.specira.ai/projects/dispatch-modernization/artifacts/edge-case-analysis#ai

Generated by Specira · template ai_orchestration_blueprint v2 (draft) · pack specira_default_delivery lineage 2d63546b...e56b · page 1 of 5

SAMPLE
seeded demo data · specira.ai