FJB

JOSEFF BETANCOURT

AI PLATFORM ARCHITECT | JOSEFFB.COM

AGENT SYSTEM RESEARCH SERIES | VOLUME 1

Provider-Agnostic AI Orchestration

Principles and Design

Roles, capability contracts, evidence, governance, and cost-aware routing.
Written for AI enthusiasts and professionals.

Prepared 13 Aug. 2026
Version 1.3.0
Author and publisher: Joseff Betancourt
Research support: Fernain Research

Publication field	Information
Version	1.3.0
Status	Publication version
Prepared	13 Aug. 2026
Author and publisher	Joseff Betancourt
Research practice	Fernain Research
Document type	Structured research note and candidate architecture
Review method	Bounded, non-systematic literature synthesis with adversarial source and citation audit
Empirical status	No original performance measurement or benchmark; cited empirical findings have not been reproduced by this series
DOI	Not assigned
Canonical URL	Pending publication
License	Pending author decision; no license assigned
ORCID	Pending author confirmation; not provided for this publication

Authorship disclosure: Research direction and methodology: Joseff Betancourt. Evidence discovery, analysis, adversarial review, and drafting: AI-assisted. Final evaluation, editorial judgment, and publication responsibility: Joseff Betancourt.

Suggested citation: Betancourt, Joseff. "Provider-Agnostic AI Orchestration: Principles and Design." *Agent System Research Series*, vol. 1, ver. 1.3.0, 13 Aug. 2026. Research note.

Abstract

This research note proposes a provider-agnostic architecture for allocating software-engineering work across an accountable human authority, an AI control plane, economical engineering workers, and bounded local or deterministic execution. It argues that stable role, authority, evidence, and acceptance contracts should precede model or provider selection. The economic hypothesis is to minimize total cost per accepted engineering outcome by selecting the cheapest admissible tier rather than the cheapest available model. A bounded literature synthesis establishes that model routing and cascades can improve cost-quality tradeoffs in specified benchmark settings, while software-agent and productivity studies expose transfer and external-validity limits. Volume 1 contributes no original benchmark and makes no provider ranking. It defines the architecture, evidence taxonomy, claim register, research questions, falsifiable hypotheses, validation plan, limitations, and protocol commitments for future empirical work. All external claims remain reported rather than locally reproduced, and provider attributes are treated as time-indexed variables by method rather than as an empirically demonstrated drift claim.

Keywords: AI orchestration; provider-agnostic architecture; large language model routing; software-engineering agents; multi-agent systems; evidence governance; human authority; cost-quality tradeoff.

Executive Summary

Software-engineering agent systems should be designed around stable roles, capability contracts, evidence, and acceptance boundaries rather than around a single model vendor. This volume does not present longitudinal evidence that provider names, prices, interfaces, or capabilities drift. Instead, it adopts an internal method commitment to record those attributes as time-indexed variables and to avoid assuming that an earlier observation remains current **[OPINION | V1-C017]**.

This research note proposes a provider-agnostic architecture with four distinct layers:

Layer	Current working candidate	Primary responsibility
Accountable Human Authority	Human operator	Scope, delegated authority, risk acceptance, override, and release
AI control plane - Seat 0	Sol	High-judgment decomposition, routing, coordination, evidence review, escalation, and validation ownership within delegated authority
Engineering worker tier	DeepSeek-class models	Economical implementation, testing, investigation, and documentation under bounded contracts
Bounded execution tier	Local models and deterministic tools	Low-cost, private, repeatable work with objective acceptance gates

The role topology is `Accountable Human Authority -> AI Seat 0 -> bounded worker tiers -> independent acceptance gates`. Sol, DeepSeek-class models, and local models are current candidate implementations of those roles, not a measured provider ranking **[FACT | V1-C011]**. Without authoritative runtime evidence, the candidate mapping also does not prove that any named runtime honored a requested route **[INFERENCE | V1-C040]**.

The dispatch hypothesis is to use the cheapest **admissible** tier: the least expensive tier that can satisfy the task's quality, evidence, authority, data, latency, and validation requirements. The cheapest available model is not necessarily the cheapest path to an accepted engineering outcome because orchestration, retries, review, failed runs, and human correction also consume resources **[HYPOTHESIS | V1-C010]**.

Peer-reviewed and preprint research reports that model cascades and learned routers can reduce inference cost while preserving selected benchmark quality in bounded settings (Chen et al.; Ding et al.; Hu et al.; Ong et al.) **[REPORTED | V1-C001]**. The same literature reports results that depend on the evaluated task distribution, quality target, model pool, budget, and routing method (Ding et al.; Hu et al.; Ong et al.) **[REPORTED | V1-C002]**. Those results do not establish the economics of multi-turn, tool-using software-engineering orchestration on mature repositories **[INFERENCE | V1-C047]**.

This volume therefore does not publish a benchmark leaderboard or conclude that Sol, DeepSeek, or a local model is universally superior. It defines the architecture, evidence language, research questions, experimental obligations, limitations, and falsifiable hypotheses for future empirical work.

Research position: A durable orchestration system should minimize total cost per accepted engineering outcome subject to quality and governance constraints. Provider selection is a replaceable implementation detail; accountable authority and evidence are not.

1. Problem

The practical problem is not "which model is smartest?" It is how to allocate real engineering work across systems with different capabilities, prices, latencies, privacy properties, and failure modes while preserving an accepted quality threshold.

A fixed strongest-model policy can pay premium cost for work that a cheaper tier could complete. A fixed cheapest-model policy can create rework, false completion, and review burden. An informal routing policy can hide both effects because it does not state the quality constraint, record rejected attempts, or count the control plane and operator **[HYPOTHESIS | V1-C013]**.

The routing studies reviewed here report results for particular model pools, datasets, budget settings, training signals, and evaluator assumptions. This publication does not generalize those results to arbitrary changes in prompts, scaffolds, harnesses, repositories, or acceptance tests (Ding et al.; Hu et al.; Ong et al.) **[INFERENCE | V1-C044]**. A model that performs well on a public issue-resolution benchmark may therefore still be a poor fit for a private legacy repository, a security-sensitive migration, or a ten-hour multi-agent delivery sequence **[INFERENCE | V1-C006]**.

The research problem is:

> How can an AI control plane operating under accountable human authority route engineering work across a high-judgment AI control tier, an economical engineering tier, and a bounded local tier so that total cost falls without reducing accepted quality?

2. Research Questions

This document organizes its research around the following questions:

1. Which orchestration responsibilities are stable roles, and which are replaceable provider implementations?
2. What evidence makes a model or provider admissible for a role?
3. Which task properties justify local execution, economical cloud execution, or high-judgment escalation?
4. How should total orchestration cost be measured?
5. What is the correct unit of engineering quality?
6. When does routing reduce cost, and when does it merely move cost into retries, supervision, validation, or delay?
7. Which authority must remain human?
8. How should long-horizon experiments be recorded so that successes, failures, and inconclusive results form a usable research corpus?
9. How should time-varying provider attributes, benchmark contamination, and runtime-attestation gaps limit conclusions?

3. Observations from Prior Research

3.1 Routing can improve a cost-quality tradeoff

FrugalGPT and Hybrid LLM, both peer-reviewed, and the RouteLLM and RouterBench preprints report that query-level model selection or cascades can reduce cost while preserving a selected measure of quality on their evaluated datasets (Chen et al.; Ding et al.; Hu et al.; Ong et al.) **[REPORTED | V1-C001]**.

These studies do not establish one universal router. Their reported performance varies with the model pool, task distribution, budget, training signal, and quality target (Ding et al.; Hu et al.; Ong et al.) **[REPORTED | V1-C002]**. The supported conclusion is that routing is measurable and conditional, not that routing is automatically economical.

A 2026 preprint reports a failure mode in its RouterBench and MMR-Bench experiments in which evaluated routers increasingly select the strongest model as the permitted budget rises. The authors call this "routing collapse"; the result remains preprint evidence and has not been reproduced by this series (Lai and Ye) **[REPORTED | V1-C018]**.

3.2 Software-engineering evaluation is more than code generation

The peer-reviewed SWE-bench paper frames evaluation as resolving real repository issues by editing a codebase and evaluating the resulting patch (Jimenez et al.) **[REPORTED | V1-C003]**. The current SWE-bench FAQ lists SWE-bench Verified as 500 instances verified by engineers as solvable (SWE-bench) **[REPORTED | V1-C028]**.

The implication for this series is that a provider comparison is not transferable unless the compared candidates receive the same task draw, base repository state, harness, tools, time policy, and acceptance gates **[INFERENCE | V1-C006]**.

3.3 Long-horizon capability is setting-dependent

The peer-reviewed NeurIPS 2025 paper *Measuring AI Ability to Complete Long Software Tasks* reports a relationship between model success and the time skilled humans require for its benchmark tasks, while explicitly discussing external-validity limits (Kwa et al.) **[REPORTED | V1-C004]**. A separate randomized study, available as a preprint, reports that early-2025 AI tools increased completion time for the experienced open-source developers and mature repositories studied, despite participants expecting acceleration (Becker et al.) **[REPORTED | V1-C046]**.

These findings do not prove that AI slows engineering generally. They support the narrower inference that task context, repository familiarity, tooling, review behavior, and workflow design can reverse an expected productivity gain in some evaluated settings (Becker et al.; Kwa et al.) **[INFERENCE | V1-C042]**.

3.4 Organizational systems amplify tool effects

DORA's 2025 industry report describes AI as an amplifier of existing organizational strengths and weaknesses and argues that returns depend on the surrounding delivery system, not tool access alone (DORA) **[REPORTED | V1-C005]**.

This supports evaluating orchestration, validation, and governance together with the model rather than attributing an outcome to a model in isolation (DORA) **[INFERENCE | V1-C043]**.

3.5 Evaluation requires multiple metrics and transparent artifacts

HELM argues for scenario coverage, multi-metric evaluation, and transparent release of evaluation artifacts (Liang et al.). Model Cards, Datasheets for Datasets, and the NeurIPS reproducibility program provide peer-reviewed documentation patterns for intended use, evaluation conditions, dataset provenance, environments, code, and repeated runs (Gebru et al.; Mitchell et al.; Pineau et al.) **[REPORTED | V1-C036]**.

NIST's AI Risk Management Framework is a voluntary framework, not a technical standard. Its Core uses the functions Govern, Map, Measure, and Manage; the Generative AI Profile is a companion resource that applies those functions to generative-AI risks (Autio et al.; Tabassi) **[FACT | V1-C020]**.

3.6 The engineering-orchestration evidence gap remains bounded

Under the queries, sources, inclusion criteria, and stopping rule stated in Appendix A, this review did not locate a peer-reviewed study that jointly evaluates the complete proposed stack--accountable human authority, high-judgment AI Seat 0, economical cloud engineering workers, and bounded local workers--across long-running, tool-using delivery on mature private repositories while measuring total cost per accepted outcome and observed runtime routes **[UNKNOWN | V1-C007]**.

That bounded non-location is not proof that no such work exists. It defines the search limit and motivates the future empirical work described here.

4. Methodology

4.1 Study type

Volume 1 is a structured research note: a bounded literature synthesis combined with a candidate architecture derived from operating requirements. It contains no new performance measurement and no internally reproduced experiment.

External empirical results are tagged `REPORTED`, even when their authors describe controlled measurements. `MEASURED` and `EXPERIMENT` are reserved for this series' own evidence so readers can distinguish literature from local reproduction **[INFERENCE | V1-C021]**.

4.2 Source selection

The review prioritizes:

- peer-reviewed final versions of routing, evaluation, and reproducibility research when available;
- original benchmark papers and canonical project publications;

- official government frameworks and first-party technical reports when they define a relevant system or claim;
- current preprints only when clearly labeled as non-peer-reviewed evidence; and
- persistent DOI or canonical URLs instead of secondary summaries.

Provider marketing, leaderboard positions, and pricing pages may become inputs to time-bound provider evaluations, but they are insufficient evidence for the architecture by themselves **[OPINION | V1-C022]**.

4.3 Evidence taxonomy

Every material empirical or prescriptive claim receives one primary label:

Label	Meaning in this series
`FACT`	Directly verifiable document, artifact, specification, or stable observed state
`MEASURED`	A raw or directly derived result recorded by this project under a documented procedure
`REPORTED`	A claim from an external source or first party that this project has not reproduced
`EXPERIMENT`	A conclusion supported by one or more `MEASURED` results under a registered protocol, reproduced internally but not externally validated
`INFERENCE`	A conclusion derived from identified evidence and assumptions
`HYPOTHESIS`	A falsifiable claim proposed for testing
`OPINION`	A normative judgment or method commitment
`UNKNOWN`	Evidence is absent, inaccessible, contradictory, or inconclusive

Epistemic status is separate from claim type and source metadata. The claim register records each claim's function--such as empirical claim, definition, hypothesis, inference, or normative principle--plus source scope and maturity. A peer-reviewed result remains `REPORTED` until this series reproduces it. An internal `EXPERIMENT` does not become externally validated merely because it was repeated internally **[INFERENCE | V1-C049]**.

Claims use the form `[TAG | V1-C###]`. Headings, questions, definitions, citations, and connective text do not require tags. Repeated uses of the same claim retain the same claim identifier. The tag communicates epistemic status; the claim register carries the independent claim-type and source metadata.

4.4 Research protocol commitments

The following are normative protocol commitments for future empirical work **[OPINION | V1-C019]**:

1. Preregister hypotheses, task classes, task sampling, acceptance gates, baselines, and primary metrics before collecting result data.
2. Pin and report the repository base, model, provider, harness, scaffold, prompts, tools, dependency environment, hardware, and runtime settings.
3. Preserve every attempted run, including failed, interrupted, timed-out, and rejected runs.
4. Separate requested configuration from observed runtime evidence.
5. Use identical task draws and acceptance gates across compared routes.
6. Keep development tasks separate from held-out evaluation tasks.
7. Record contamination checks and known training-data uncertainty.
8. Report sample size, median, dispersion, confidence interval where applicable, and the complete denominator.
9. Publish negative and inconclusive findings with the same visibility as positive results.
10. Record funding, provider relationships, operator identity class, and other material conflicts or confounders.

5. Assumptions

The candidate architecture depends on explicit assumptions and method commitments:

ID	Assumption or commitment	Classification	Consequence if false or omitted
A1	Real engineering workloads contain sufficiently different task classes to make routing useful	HYPOTHESIS	A single qualified tier may be more efficient
A2	Objective acceptance gates can be defined for the studied task classes	HYPOTHESIS	Quality comparison becomes judge-dependent
A3	Provider attributes will be recorded as time-indexed observations, without assuming longitudinal stability	OPINION	Provider conclusions may be reused outside their observed conditions
A4	Local compute cost can be measured, including hardware, energy, maintenance, and operator overhead	HYPOTHESIS	"Local is cheaper" remains untestable
A5	The human operator can remain the accountable authority without becoming an unmeasured bottleneck	HYPOTHESIS	Operator capacity must enter the routing constraint
A6	Provider-neutral contracts can expose the runtime evidence required by the study	HYPOTHESIS	Actual route compliance remains `UNKNOWN`
A7	Private project work can be reported with enough redaction or synthetic substitution to preserve reproducibility	HYPOTHESIS	External validation will remain limited

6. Principles

6.1 Provider lock-in is the wrong architectural abstraction

Vendors are deployment choices. Decomposition, implementation, validation, evidence review, escalation, and acceptance are system roles. Binding those roles directly to one provider makes provider replacement an architectural change instead of an adapter change **[INFERENCE | V1-C008]**.

Provider-specific capabilities still matter. A provider is admissible only when it satisfies the role contract, data boundary, tool requirements, runtime evidence requirements, and validation threshold **[OPINION | V1-C039]**.

6.2 Roles precede models

The architecture first defines the responsibility, authority, inputs, outputs, evidence, and stop conditions of each role. A model is then evaluated as a candidate implementation of that role **[OPINION | V1-C038]**.

This prevents a model's marketing category or leaderboard position from silently expanding its authority.

6.3 Capability outranks vendor identity

Routing should use observed task-class capability, not brand prestige. Model identity is relevant only as evidence about the tested candidate and version **[OPINION | V1-C048]**.

The same provider may qualify for several roles, and different providers may qualify for one role. Provider agnosticism means replaceability under a stable contract, not pretending all models are equivalent.

6.4 Human authority remains accountable

The Accountable Human Authority owns scope, delegated authority, risk acceptance, override, and release. AI Seat 0 is the lead control-plane role, not the legal, organizational, or ethical owner of the outcome **[OPINION | V1-C009]**.

AI Seat 0 may classify, decompose, route, coordinate, synthesize, review evidence, and validate within the delegated boundary. It cannot expand that authority, accept human risk, or release work. This distinction keeps the research vocabulary aligned with Agent System's implementation semantics.

6.5 Route to the cheapest admissible tier

The economic objective is not minimum token price. It is minimum expected total cost per accepted engineering outcome subject to quality, evidence, authority, data, latency, and validation constraints **[HYPOTHESIS | V1-C010]**.

Detailed equations and calibrated thresholds require a dedicated economic-routing study. This document establishes only the constrained objective.

6.6 Evidence outranks confidence

Model confidence, fluent prose, a routing directive, and a successful tool request are not completion evidence. Accepted source changes, executable validation, runtime metadata, review decisions, and reproducible records are stronger evidence **[OPINION | V1-C015]**.

Requested model identity and actual runtime identity must remain separate. If authoritative runtime metadata is absent, the actual model or route is `UNKNOWN`, not inferred from style or behavior **[INFERENCE | V1-C040]**.

6.7 Governance is part of the architecture

Authority, privacy, destructive effects, release decisions, and evidence admission are constraints on routing rather than after-the-fact paperwork **[OPINION | V1-C041]**.

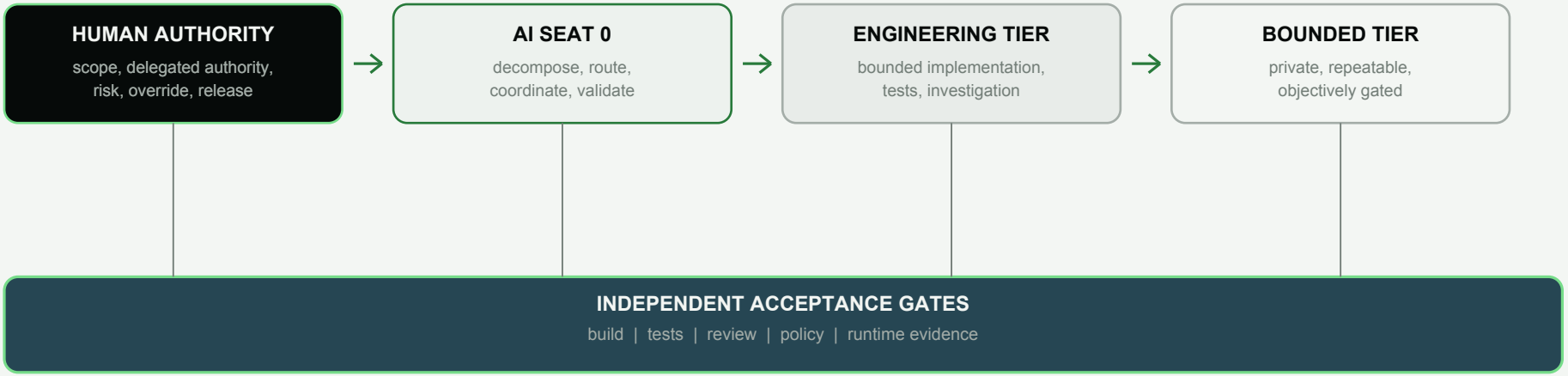
Workers receive bounded contracts. They do not acquire independent authority to change scope, accept risk, release work, or bypass the control tier.

7. Reference Architecture

7.1 Role model

Role	Owns	Does not own	Candidate
Accountable Human Authority	Objective, delegated authority, risk acceptance, override, release	Routine generation, orchestration, or constant supervision	Human operator
AI control plane - Seat 0	Decomposition, route selection, coordination, evidence synthesis, candidate admission, escalation recommendation, validation ownership	Authority expansion, human risk acceptance, publication, or release	Sol
Engineering worker tier	Bounded investigation, implementation, tests, documentation, local validation	Routing policy, scope expansion, final acceptance, or release	DeepSeek-class models
Bounded / local execution tier	Bounded transformation, extraction, retrieval-side work, deterministic checks, low-risk implementation proven by gates	Ambiguous judgment, cross-system authority, final validation, or release	Local models and deterministic tools

7.2 Control flow



The diagram is a responsibility flow, not proof that any runtime implements or enforces it **[FACT | V1-C035]**.

7.3 Dispatch record

Every material route should record:

Field	Required content
Work identity	Project, repository, base state, task ID, and task class
Objective	Accepted outcome and explicit exclusions
Authority	Allowed effects, data boundary, and human owner
Candidate tiers	Tiers considered and current availability evidence
Route decision	Selected tier and reason
Quality gate	Executable or review-based acceptance condition
Cost inputs	Model, compute, operator, validation, retry, and failure costs
Escalation rule	Observable condition that returns control upward
Runtime evidence	Requested route separately from observed route
Result	Accepted, rejected, failed, interrupted, or inconclusive

7.4 Escalation

Escalation is owned by the control plane. A worker returns evidence that it cannot complete the contract; it does not independently select a more powerful provider, expand scope, or create an unbounded delegation tree **[HYPOTHESIS | V1-C012]**.

This preserves the role topology `Accountable Human Authority -> AI Seat 0 -> bounded worker tiers` even when the cheapest admissible dispatch starts with a local tier. Candidate providers remain replaceable mappings beneath those role contracts.

8. Falsifiable Hypotheses

H0 - Routing irrelevance

Heterogeneous tiered routing does not reduce total cost per accepted engineering outcome relative to the best preregistered homogeneous baseline, after accounting for orchestration, validation, retries, latency, and operator effort, subject to non-inferiority requirements for quality, authority compliance, and delivery time **[HYPOTHESIS | V1-C051]**.

```
# H0 decision thresholds - bind before data collection
materiality_threshold: preregistered
quality_noninferiority_margin: preregistered
delivery_time_noninferiority_margin: preregistered
```

H1 - Tiered allocation

On matched engineering tasks, the tiered architecture will have lower total cost per accepted outcome than control-tier-only, worker-tier-only, or local-tier-only execution at a comparable accepted-outcome rate **[HYPOTHESIS | V1-C023]**.

H2 - Concentration of high-judgment value

The marginal value of the AI Seat 0 control plane will concentrate in decomposition, ambiguity resolution, evidence review, integration, and validation rather than routine artifact generation **[HYPOTHESIS | V1-C024]**.

H3 - Bounded local value

Local models will be economically admissible for task classes with bounded context, low authority, reversible effects, and objective gates, but not for all engineering work **[HYPOTHESIS | V1-C016]**.

H4 - Escalation policy

A calibrated escalation policy will produce a better cost-quality frontier than always-strongest, always-cheapest, and fixed provider rules **[HYPOTHESIS | V1-C025]**.

H5 - Private-workload validity

Provider performance differences measured on private, held-out project work will differ materially from public leaderboard differences **[HYPOTHESIS | V1-C026]**.

H6 - Evidence-gated quality

Routing policies that require objective evidence before acceptance will produce fewer false-completion admissions than policies that use model confidence or self-report alone **[HYPOTHESIS | V1-C027]**.

9. Validation Plan

9.1 Baselines

Every routing experiment should compare identical task draws across:

- AI control-plane role only (current candidate: Sol);
- engineering-worker role only (current candidate: DeepSeek);
- local-tier only;
- always strongest;
- always cheapest;
- fixed rule routing;
- proposed tiered routing; and
- proposed tiered routing with one component ablated.

These baselines are part of the proposed method, not evidence that each route is currently available **[OPINION | V1-C019]**.

9.2 Primary outcome

The primary quality unit should be an accepted engineering outcome: the requested artifact satisfies its preregistered build, test, review, evidence, and authorization gates **[OPINION | V1-C014]**.

Model preference, prose quality, and self-reported completion may be secondary observations but cannot replace the acceptance result.

9.3 Economic outcome

The primary economic unit should be total cost per accepted outcome, including:

- control-tier usage;
- worker-tier usage;
- local compute and energy;
- failed and rejected runs;
- retries and recovery;

- operator-active effort;
- validation and review;
- latency or queue penalties when material; and
- infrastructure and tooling cost.

Counting only worker tokens will understate orchestration cost **[HYPOTHESIS | V1-C013]**.

9.4 Experiment record

Long-horizon experiments should record each run with a schema equivalent to:

```
experiment_id: ORCH-EX-007
status: preregistered | complete | failed | interrupted | inconclusive
hypothesis_refs: [H1, H2]
project_class: mature-private-repository
task_class: bounded-implementation
task_ids: [...]
holdout: true
topology:
human_authority: accountable-human-authority
control_plane: ai-seat-0
engineering_worker_tier: engineering-worker-role
bounded_execution_tier: bounded-execution-role
candidate_mapping:
control_plane: sol
engineering_worker_tier: deepseek
bounded_execution_tier: local
versions:
repository_base: <commit>
agent_system: <version-or-digest>
routing_policy: <version-or-digest>
router: <version-or-digest>
shim: <version-or-digest>
provider_adapter: <version-or-digest>
model_candidates: [...]
provider_endpoints: [...]
harness: <version-or-digest>
prompts: <digest>
tools: [...]
environment: <digest>
```

```
operator:
identity_class: <class>
prior_exposure_to_task: null
prior_exposure_to_repository: null
prior_exposure_to_benchmark_corpus: null
intervention_count: null
acceptance_gates: [...]
baselines: [...]
cost_ledger:
control_tier: null
worker_tier: null
local_compute: null
operator_active: null
validation: null
retries: null
failures: null
runtime_evidence:
requested_routes: [...]
observed_routes: [...]
result:
accepted_outcome: null
metrics: {}
limitations: [...]
next_experiment: null
```

Missing values remain `null`; they are not inferred.

9.5 Provider evaluation rule

Provider evaluation reports must be time-bound and role-specific. They should state model version, provider, observation date, price snapshot, task corpus, harness, sample size, dispersion, failures, and validity interval. A provider "review" without those fields is an opinion, not an evaluation **[OPINION | V1-C022]**.

10. Research Cycle Status

Hypothesis

Volume 1 registers a null hypothesis plus six falsifiable hypotheses covering routing relevance, tiered allocation, high-judgment control value, bounded local value, escalation, private-workload validity, and evidence-gated quality. None is presented as validated.

Experiment

No original experiment or benchmark was run for Volume 1 **[FACT | V1-C029]**. The experiment schema and comparison baselines are proposed protocols for future empirical work.

Evidence

The evidence in this volume consists of the cited external literature and directly verifiable properties of this document. External empirical results remain `REPORTED` until reproduced by this series **[INFERENCE | V1-C021]**.

Limitations

The principal limitations are the absence of original measurements, uncertain transfer from public studies to private engineering work, unmeasured time-variation in provider attributes, runtime-attestation gaps, private-data constraints, and operator effects. The full limitations are stated in Section 12.

Next Experiment

The next empirical step is a preregistered pilot using a small held-out task sample, identical acceptance gates, complete attempt retention, and at least three baselines. The pilot should test measurement feasibility before any provider-ranking claim or production routing rule is admitted **[OPINION | V1-C050]**.

11. Claim Register

Claim ID	Summary	Epistemic status	Claim type	Source scope / maturity	Validation path
V1-C001	Routing and cascades can reduce cost at selected benchmark quality	REPORTED	Empirical claim	External / mixed peer-reviewed papers and preprints	Economic-routing study and long-horizon experiments
V1-C002	Reported routing performance is task-, model-, budget-, target-, and method-dependent	REPORTED	Empirical claim	External / mixed peer-reviewed papers and preprints	Benchmark study and long-horizon experiments
V1-C003	SWE-bench evaluates repository issues through codebase edits and patch evaluation	REPORTED	Descriptive claim	External / peer-reviewed benchmark paper	Benchmark study
V1-C004	The NeurIPS task-horizon study reports a relationship between model success and skilled-human task duration	REPORTED	Descriptive claim	External / peer-reviewed conference paper	Long-horizon experiments
V1-C005	DORA reports that organizational systems mediate AI delivery outcomes	REPORTED	Descriptive claim	External / industry report	Long-horizon experiments
V1-C006	Public scores do not directly determine private-project performance	INFERENCE	Evidence-scope inference	Mixed / external evidence; locally unvalidated	Benchmark study, long-horizon experiments, and provider evaluation
V1-C007	No study matching every defined full-stack criterion was located by the bounded search	UNKNOWN	Research-gap claim	External / bounded, non-systematic search in Appendix A	Ongoing review and long-horizon experiments
V1-C008	Role contracts can reduce architectural provider coupling	INFERENCE	Design inference	Internal / candidate architecture	Provider evaluation
V1-C009	Accountable Human Authority remains responsible for scope, delegated authority, risk, override, and release	OPINION	Normative principle	Internal / candidate architecture	Governance specification
V1-C010	Cheapest-admissible routing improves economic efficiency	HYPOTHESIS	Economic hypothesis	Internal / untested	Economic-routing study and long-horizon experiments
V1-C011	Provider candidates map to role contracts and do not constitute a measured ranking	FACT	Document definition	Internal / verified document	Long-horizon experiments and provider evaluation

Claim ID	Summary	Epistemic status	Claim type	Source scope / maturity	Validation path
V1-C012	Control-plane-owned escalation is preferable to nested worker authority	HYPOTHESIS	Governance hypothesis	Internal / untested	Long-horizon experiments and governance specification
V1-C013	Full orchestration cost changes routing economics	HYPOTHESIS	Economic hypothesis	Internal / untested	Economic-routing study and long-horizon experiments
V1-C014	Accepted engineering outcome is the primary quality unit	OPINION	Method commitment	Internal / normative	Benchmark study
V1-C015	Runtime evidence should outrank confidence and configuration requests	OPINION	Evidence principle	Internal / normative	Long-horizon experiments and governance specification
V1-C016	Local value is bounded by task class and objective gates	HYPOTHESIS	Task-class hypothesis	Internal / untested	Long-horizon experiments
V1-C017	Provider attributes should be recorded as time-indexed observations with explicit validity intervals	OPINION	Method commitment	Internal / normative; no longitudinal provider study in this volume	Economic-routing study and provider evaluation
V1-C018	Evaluated routers may collapse toward strongest-model selection as budget rises	REPORTED	Empirical claim	External / preprint	Long-horizon experiments
V1-C019	Future empirical work should preregister and fully report experiments	OPINION	Protocol commitment	Internal / normative	Benchmark study, long-horizon experiments, and provider evaluation
V1-C020	NIST AI RMF is voluntary and uses Govern, Map, Measure, and Manage; the GAI Profile is a companion resource	FACT	Descriptive claim	External / official government framework and profile	Governance specification
V1-C021	External results remain REPORTED until reproduced under this research protocol	INFERENCE	Taxonomy rule	Internal / definition	Future empirical work
V1-C022	Provider claims require time-bound independent evaluation	OPINION	Evaluation rule	Internal / normative	Provider evaluation
V1-C023	Tiered allocation lowers cost per accepted outcome	HYPOTHESIS	Research hypothesis	Internal / untested H1	H1 in long-horizon experiments
V1-C024	AI Seat 0 value concentrates in high-judgment stages	HYPOTHESIS	Research hypothesis	Internal / untested H2	H2 in long-horizon experiments
V1-C025	Calibrated escalation improves the cost-quality frontier	HYPOTHESIS	Research hypothesis	Internal / untested H4	H4 in long-horizon experiments
V1-C026	Private-workload provider deltas differ from leaderboards	HYPOTHESIS	Research hypothesis	Internal / untested H5	H5 in benchmark study, long-horizon experiments, and provider evaluation
V1-C027	Evidence gates reduce false completion	HYPOTHESIS	Research hypothesis	Internal / untested H6	H6 in long-horizon experiments
V1-C028	The current SWE-bench FAQ lists 500 engineer-verified solvable instances in SWE-bench Verified	REPORTED	Descriptive claim	External / official benchmark documentation	Benchmark study
V1-C029	Volume 1 contains no original benchmark	FACT	Document property	Internal / verified document	Long-horizon experiments
V1-C030	Private-project redaction can limit reproducibility	INFERENCE	Method-risk inference	Internal / research design	Benchmark study and long-horizon experiments
V1-C031	Judge and operator behavior can change measured economics	HYPOTHESIS	Confounder hypothesis	Internal / untested	Long-horizon experiments
V1-C032	Latency and concurrency can change routing economics	HYPOTHESIS	Economic hypothesis	Internal / untested	Economic-routing study and long-horizon experiments
V1-C033	Local compute can carry material non-token costs	HYPOTHESIS	Economic hypothesis	Internal / untested	Economic-routing study and long-horizon experiments
V1-C034	The cited literature does not validate the proposed hierarchy	INFERENCE	Evidence-scope inference	Mixed / cited literature	Long-horizon experiments and provider evaluation
V1-C035	The architecture diagram is descriptive, not runtime proof	FACT	Document property	Internal / verified document	Governance specification

Claim ID	Summary	Epistemic status	Claim type	Source scope / maturity	Validation path
V1-C036	Prior work recommends transparent, reproducible evaluation artifacts	REPORTED	Descriptive synthesis	External / peer-reviewed research literature	Benchmark study, long-horizon experiments, and provider evaluation
V1-C037	Selective task publication would bias the corpus	INFERENCE	Bias inference	Internal / research design	Benchmark study and long-horizon experiments
V1-C038	Roles should be specified before model candidates	OPINION	Design principle	Internal / normative	Governance specification
V1-C039	Providers should be admitted only when they satisfy the role contract	OPINION	Admission principle	Internal / normative	Provider evaluation
V1-C040	Actual runtime identity cannot be inferred from a configuration request	INFERENCE	Evidence rule	Internal / runtime-evidence distinction	Long-horizon experiments and governance specification
V1-C041	Governance constraints should be part of routing architecture	OPINION	Design principle	Internal / normative	Governance specification
V1-C042	Workflow context can reverse an expected productivity gain in some evaluated settings	INFERENCE	Evidence inference	External / mixed empirical sources	Long-horizon experiments
V1-C043	The orchestration system should be evaluated with the model	INFERENCE	Evaluation inference	External / industry framing	Long-horizon experiments
V1-C044	Reviewed routing results are conditional on specified evaluation configurations; arbitrary scaffold sensitivity is unmeasured here	INFERENCE	Method-scope inference	External / routing literature	Benchmark study
V1-C045	Multi-turn engineering adds costs absent from many routing studies	INFERENCE	Research-gap inference	Mixed / literature synthesis	Economic-routing study and long-horizon experiments
V1-C046	A randomized preprint reported increased completion time in its studied repositories	REPORTED	Empirical claim	External / preprint	Long-horizon experiments
V1-C047	Existing routing results do not establish multi-turn engineering economics	INFERENCE	Evidence-scope inference	External / literature synthesis	Economic-routing study and long-horizon experiments
V1-C048	Routing should use observed task-class capability rather than brand prestige	OPINION	Design principle	Internal / normative	Provider evaluation and governance specification
V1-C049	Internal reproduction does not by itself establish external validity	INFERENCE	Taxonomy rule	Internal / definition	Future empirical work
V1-C050	The first empirical study should be a preregistered measurement-feasibility pilot	OPINION	Proposed experiment	Internal / normative	Measurement-feasibility pilot
V1-C051	Heterogeneous routing does not reduce total cost per accepted engineering outcome relative to the best preregistered homogeneous baseline under specified non-inferiority conditions	HYPOTHESIS	Null hypothesis	Internal / untested H0	Preregistered pilot and field experiments

12. Limitations

12.1 No original benchmark

Volume 1 contains no project measurement. It cannot establish cost savings, provider quality, admissible task boundaries, or the superiority of the current candidate mapping **[FACT | V1-C029]**.

12.2 Literature-to-engineering transfer

The routing studies reviewed here principally evaluate bounded query-level selection or cascades. Multi-turn agent delivery adds control-plane cost, state, tools, retries, validation, and human intervention not jointly measured by those routing results (Chen et al.; Ding et al.; Hu et al.; Ong et al.) **[INFERENCE | V1-C045]**.

12.3 Task and project selection

LCFE, Fernain Jobs, or any other real project may not represent other repositories, organizations, languages, or regulatory environments. Publishing only successful tasks would further bias the corpus **[INFERENCE | V1-C037]**.

12.4 Private data and reproducibility

Real-project evidence can conflict with privacy, customer, security, or commercial constraints. Redaction can remove context needed for independent reproduction **[INFERENCE | V1-C030]**.

12.5 Time-varying provider attributes

This publication contains no longitudinal provider evidence and therefore makes no empirical claim that provider behavior or prices drift. As an internal method commitment, a valid provider evaluation must snapshot model aliases, versions, endpoints, prices, rate limits, observed behavior, and collection dates, and must assign a validity interval rather than assuming stability **[OPINION | V1-C017]**.

12.6 Requested versus actual runtime

A configured model name is not proof of the model that executed. Without authoritative runtime metadata, actual provider, model, and reasoning remain `UNKNOWN` **[INFERENCE | V1-C040]**.

12.7 Judge and operator effects

Human review and model judging can introduce bias. Operator experience, intervention style, familiarity with the repository, and tolerance for rework can change measured economics **[HYPOTHESIS | V1-C031]**.

12.8 Latency and concurrency

Parallel agents may reduce elapsed time without reducing total compute or review. Queueing, rate limits, context switching, and orchestration delay must be reported separately from active effort **[HYPOTHESIS | V1-C032]**.

12.9 Local cost accounting

Local inference is not free. Hardware depreciation, energy, heat, maintenance, model preparation, storage, and operator time may dominate low-volume workloads **[HYPOTHESIS | V1-C033]**.

12.10 External validity

An internally reproduced `EXPERIMENT` is still not externally validated. Findings remain conditional on task, system, operator, environment, and time **[INFERENCE | V1-C049]**.

13. Future Work

Economic routing

Define the formal objective, cost ledger, quality constraint, latency term, privacy constraint, and admissibility rule. Compare token cost, hosted compute, local compute, operator-active effort, and cost of failure without collapsing them into one unsupported unit.

Agent engineering benchmark

Publish task classes, repository selection, task sampling, held-out policy, harness, environment, acceptance tests, human baselines, contamination checks, and statistical reporting rules.

Long-horizon multi-agent experiments

Build the structured corpus. Each entry must expose topology, duration, project class, versions, costs, failures, acceptance evidence, limitations, and the next experiment. Diary-style narrative may accompany the record but cannot replace it.

Provider evaluation

Evaluate candidates from OpenAI, DeepSeek, Alibaba/Qwen, Moonshot/Kimi, Anthropic/Claude, Google/Gemini, and other providers against the role contracts and benchmark. Provider chapters remain time-bound evaluation reports rather than reviews.

Agent system governance

Translate validated findings into an RFC-like specification for roles, authority, receipts, evidence, escalation, provider replacement, and human acceptance. Candidate design preferences that lack evidence remain labeled.

Works Cited

Autio, Chloe, et al. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. NIST AI 600-1, National Institute of Standards and Technology, July 2024, https://doi.org/10.6028/NIST.AI.600-1.

Becker, Joel, et al. "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity." *arXiv*, 12 July 2025, arXiv:2507.09089, https://doi.org/10.48550/arXiv.2507.09089. Preprint.

Chen, Lingjiao, et al. "FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance." *Transactions on Machine Learning Research*, 2024, https://mlanthology.org/tmlr/2024/chen2024tmlr-frugalgpt/.

Ding, Dujian, et al. "Hybrid LLM: Cost-Efficient and Quality-Aware Query Routing." *The Twelfth International Conference on Learning Representations*, 2024, https://proceedings.iclr.cc/paper_files/paper/2024/hash/b47d93c99fa22ac0b377578af0a1f63a-Abstract-Conference.html.

DORA. *State of AI-Assisted Software Development 2025*. Version 2025.2, Google Cloud, 2025, https://dora.dev/research/2025/dora-report/.

Gebru, Timnit, et al. "Datasheets for Datasets." *Communications of the ACM*, vol. 64, no. 12, 2021, pp. 86-92, https://doi.org/10.1145/3458723.

Hu, Qitian Jason, et al. "RouterBench: A Benchmark for Multi-LLM Routing System." *arXiv*, 18 Mar. 2024, arXiv:2403.12031, https://arxiv.org/abs/2403.12031. Preprint.

Jimenez, Carlos E., et al. "SWE-bench: Can Language Models Resolve Real-world Github Issues?" *The Twelfth International Conference on Learning Representations*, 2024, https://proceedings.iclr.cc/paper_files/paper/2024/hash/edac78c3e300629acfe6cbe9ca88fb84-Abstract-Conference.html.

Kwa, Thomas, et al. "Measuring AI Ability to Complete Long Software Tasks." *Advances in Neural Information Processing Systems 38*, 2025, https://doi.org/10.52202/085713-3086.

Lai, Guannan, and Han-Jia Ye. "When Routing Collapses: On the Degenerate Convergence of LLM Routers." *arXiv*, 3 Feb. 2026, arXiv:2602.03478, https://arxiv.org/abs/2602.03478. Preprint.

Liang, Percy, et al. "Holistic Evaluation of Language Models." *Transactions on Machine Learning Research*, 2023, https://mlanthology.org/tmlr/2023/liang2023tmlr-holistic/.

Mitchell, Margaret, et al. "Model Cards for Model Reporting." *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, 2019, pp. 220-29, https://doi.org/10.1145/3287560.3287596.

Ong, Isaac, et al. "RouteLLM: Learning to Route LLMs with Preference Data." *arXiv*, 26 June 2024, arXiv:2406.18665, https://arxiv.org/abs/2406.18665. Preprint.

Pineau, Joelle, et al. "Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program)." *Journal of Machine Learning Research*, vol. 22, no. 164, 2021, pp. 1-20, https://www.jmlr.org/papers/v22/20-303.html.

SWE-bench. "Frequently Asked Questions." *SWE-bench*, https://www.swebench.com/SWE-bench/faq/. Accessed 13 Aug. 2026.

Tabassi, Elham. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1, National Institute of Standards and Technology, Jan. 2023, https://doi.org/10.6028/NIST.AI.100-1.

Research Interpretation

The Works Cited sources support studying heterogeneous cost-quality tradeoffs, routing, transparent evaluation, real-repository tasks, long-horizon behavior, reproducibility, and lifecycle governance. They do not validate the proposed candidate mapping as a provider ranking, prove that cheaper orchestration preserves quality on Joseff Betancourt's projects, or establish a universal routing policy **[INFERENCE | V1-C034]**.

Publication and Attribution

Publication field	Information
Canonical title	*Provider-Agnostic AI Orchestration: Principles and Design*
Series	*Agent System Research Series*
Volume	1
Version	1.3.0
Status	Publication version
Prepared	13 Aug. 2026
Author and publisher	Joseff Betancourt
Research practice	Fernain Research
DOI	Not assigned
Canonical URL	Pending publication
License	Pending author decision; no license assigned
ORCID	Pending author confirmation; not provided for this publication

Research direction and methodology: Joseff Betancourt. Evidence discovery, analysis, adversarial review, and drafting: AI-assisted. Final evaluation, editorial judgment, and publication responsibility: Joseff Betancourt.

Named provider and model candidates are proposed research subjects, not endorsements or validated rankings.

Appendix A. Bounded Search Method

A.1 Purpose and review class

The search supports a structured research note, not a systematic review or meta-analysis. Its purposes were to verify that every cited source exists, prefer a peer-reviewed final version when one was available, identify source maturity accurately, and test the narrowly stated full-stack evidence-gap claim.

A.2 Dates and search surfaces

The seed review was performed on 10 Aug. 2026. The source and citation audit was refreshed on 13 Aug. 2026. Web search was constrained toward primary or canonical records, including arXiv, OpenReview, ICLR and NeurIPS proceedings, TMLR records, JMLR, ACM and DOI records, NIST publications, the official SWE-bench documentation, and DORA's official research site. Secondary search results were used only to locate a canonical record and were not cited as evidence.

A.3 Query families

Exact-title queries were used for all 16 Works Cited entries. The conceptual search used the following bounded query families, with close title and hyphenation variants:

- `LLM routing cost quality FrugalGPT RouteLLM Hybrid LLM RouterBench`;
- `software engineering agents SWE-bench accepted outcome executable tests`;
- `AI task completion time horizon software tasks`;
- `experienced open-source developer productivity AI randomized study`;
- `HELM model cards datasheets reproducibility machine learning`;
- `NIST AI RMF voluntary Govern Map Measure Manage`;
- `DORA 2025 AI amplifier software delivery`;
- `routing collapse LLM routers budget`;
- `provider-agnostic AI orchestration software engineering benchmark`;
- `accountable human authority AI control plane local models software engineering`;
- `multi-agent orchestration mature repositories cloud local model routing`; and
- `heterogeneous LLM orchestration long-running private repositories empirical study`.

A.4 Inclusion, exclusion, and maturity rules

Sources were included when they directly supported a point-of-use claim and were one of the following: a peer-reviewed final paper, an original benchmark or project document, an official government framework or profile, a first-party industry research report, or a clearly labeled current preprint. Where both an arXiv version and a peer-reviewed final version were located, the final version was cited.

The audit upgraded FrugalGPT to its 2024 TMLR version, Hybrid LLM and SWE-bench to their ICLR 2024 versions, and the task-horizon study to its NeurIPS 2025 version under the final title *Measuring AI Ability to Complete Long Software Tasks*. HELM, Model Cards, Datasheets for Datasets, and the NeurIPS reproducibility report use their peer-reviewed final records. RouteLLM, RouterBench, Becker et al., and Lai and Ye are explicitly retained as preprints.

Sources were excluded when they were only marketing claims, secondary summaries without a canonical record, inaccessible snippets that could not confirm bibliographic identity, or adjacent studies that did not support the specific claim. A source's existence and bibliographic verification do not validate its findings.

A.5 Stopping rule and bounded gap conclusion

Exact-title verification stopped after a canonical record confirmed title, authors, year, publication maturity, and persistent URL or DOI. The full-stack gap search stopped after screening the four final query families in Section A.3 for studies matching every defined dimension: accountable human authority, an AI control plane, economical cloud workers, bounded local workers, long-running tool use, mature private repositories, total cost per accepted outcome, and observed runtime routes.

The search located adjacent work on routing, software-agent benchmarks, human-governed architectures, multi-agent systems, and developer productivity, but no result matching all dimensions. Therefore V1-C007 remains `UNKNOWN` and means only "not located under this bounded method." It does not mean "no such study exists."

A.6 Search limitations

The review was English-language, web-index dependent, time-bounded, and non-exhaustive. It did not search subscription databases comprehensively, perform citation-network saturation, contact authors, assess study quality with a formal risk-of-bias instrument, or reproduce any result. Publication status can change after 13 Aug. 2026 and must be refreshed before publication.