Model Choice Boundary

Sourcing, Delegation, and Assurance for Runtime Model Selection in Enterprise AI

James E. Murphy

ORCID 0009-0001-2217-4306

Working draft for discussion · v1.0 · 14 September 2026

jamesemurphy.com/ai-economics/model-routing

Disclosure. The author has no financial relationship with any company named or referenced in this paper. Product descriptions are drawn from public documentation as of the access dates given in the references and illustrate the range of current arrangements without assessing or endorsing any offering. No vendor reviewed or funded this work.

Contents

Abstract

Enterprises increasingly serve AI workloads through routing systems that allocate each request among models, providers, endpoints, service tiers, compute budgets, and fallback paths at runtime. Technical routing research establishes that such selection is feasible but takes the eligible set, the objective, and the authority structure as given. This paper treats the resulting management problem as one of decision rights. It introduces the **Model Choice Boundary**: the allocation of decision rights between an enterprise and one or more intermediaries over which eligible inference configuration performs a workload, under what objectives and constraints, using what evidence, and subject to whose residual authority. The framework separates the sourcing boundary, which concerns who operates the routing mechanism, from the authority boundary, which concerns who controls the decision. It decomposes model choice into meta decision rights over Eligibility, Objective, and Change, an operational Allocation right, an Intervention right, and adjacent rights over Evidence, Learning and Data, and Exit, and allocates them through workload routing mandates using five determinants. It separates baseline workload risk from incremental routing exposure and derives a four level assurance stack. Enterprises can delegate high frequency Allocation without surrendering strategic control, but only where retained rights are backed by mandate separating evidence and feasible correction. A sharp corollary follows: a mandate that names several goals without ordering them has delegated Objective, whatever the agreement says.

1. Introduction: Model Selection Became a Runtime Decision

Consider an enterprise that has adopted a commercial router for a customer support workload and believes it controls the routing objective. Its agreement states that requests are to be routed for quality, cost, latency, and reliability. On most requests those goals do not conflict. On the requests where they do, something must decide how much quality is worth how much latency, and the agreement does not say. Whatever supplies that ordering at runtime, whether a scoring function inside the router, a benchmark weighting chosen by the vendor, or an aggregate of other customers' spending, is exercising a right the enterprise believes it retained. The same gap opens inside the enterprise's own evaluation dashboard: a vector of quality, cost, latency, and reliability scores becomes a ranking only through weights, priorities, or constraints that no evaluator supplies neutrally, and whoever supplies them decides what the router is permitted to find.

This paper argues that runtime model selection has become a decision rights problem, that the rights involved are separable, and that several of them are commonly held by parties other than the ones who believe they hold them. The corollary above is the paper's most immediately checkable claim. The framework that produces it explains why it holds, which other rights behave the same way, and what an enterprise must be able to observe and do for a right retained on paper to be real.

1.1 From design time selection to runtime model choice

In the first wave of enterprise generative AI, a workload was typically associated with one principal model or serving configuration, and that choice changed through periodic evaluation or replatforming. Multimodel portfolios, gateways, cascades, and learned routers now move the selection into production execution. This paper uses **model choice** to mean the runtime allocation of a workload or request among eligible **inference configurations**, where a configuration can include model and version, provider and endpoint, region, service tier, reasoning or compute budget, and fallback or escalation path. The object is broader than a model name. Economic evaluation reaches the same conclusion from the measurement side, because price and realized performance vary across providers serving a nominally identical model (Erol et al., 2025), and a live black box audit distinguished 14 of 15 same model provider pairs by quantization and kernel differences alone (Zhang, Zhang, and Qin, 2026).

Per instance selection among algorithms is not new (Rice, 1976), and routing research has established that selecting among models, escalating across tiers, and allocating under budget constraints are technically feasible (Ong et al., 2025; Hu et al., 2024; Li, Gao, and Lakshmanan, 2026). What is new is the selector. It can now be an internal or external intermediary allocating demand among competing inference suppliers while holding its own information, objectives, economic incentives, and control over the evidence used to evaluate the allocation. The shift from persistent assignment to runtime selection therefore changes who can exercise the choice, what incentives shape it, and what evidence the enterprise needs to determine whether the choice remained appropriate.

1.2 Sourcing the mechanism and sourcing the decision

Enterprise discussion often frames routing as a sourcing choice: build a router, adopt an open source gateway, or buy a routing service. Sourcing matters, but it does not determine who controls model choice. Two boundaries must be separated.

The **sourcing boundary** asks who provides, operates, or hosts the routing mechanism. The **model choice authority boundary** asks who holds the rights to define, constrain, execute, change, observe, and terminate the decision. The two can move together, but they need not. An enterprise can buy commercial routing software while retaining authority over the eligible pool, the objective, evaluation assets, thresholds, change approval, intervention, route evidence, and exit; the intermediary then supplies execution inside an enterprise defined mandate. An enterprise can also self host routing software while relying on external rankings, classifiers, benchmark weights, evaluation sets, or telemetry that materially shape the effective decision; the code is internal, and important dimensions of authority are not.

Sourcing still matters in two indirect ways. Whoever operates the mechanism usually holds the policy state, telemetry, and route history through which rights are exercised, verified, and moved, and control of those assets shapes outside options even when the specified rights sit with the enterprise. And sourcing changes the feasible incentive system: external provision exposes routing to stronger market like performance incentives, while internal provision supports lower powered incentives, tighter direct controls, and more balanced treatment of poorly measured dimensions (Holmström and Milgrom, 1994). Where consequential quality is hard to measure, that difference bears on the efficient arrangement on its own.

Table 1 condenses the public documentation of 2026 routing arrangements into three intermediary types. Three features stand out. Platform integrated routers already separate Eligibility, which the customer sets, from Allocation, which the platform performs. The largest independent intermediaries earn a share of routed spend and are increasingly owned by financial intermediaries, and their documentation shows the Objective right being supplied by the market where the customer has not supplied it. And a third group of routers is being built by firms that also supply candidate models. The documentation establishes that the rights can be separated in practice, and it shows where several of them sit when nothing has been negotiated. It does not establish how enterprises use these controls or whether the resulting governance works.

Table 1. Three routing intermediary types and the rights separation their documentation shows

Type	Examples	What the customer controls per documentation	What the intermediary exercises	Governance feature
Platform integrated: routing is a feature of a cloud or data platform	Amazon Bedrock intelligent prompt routing; Microsoft Foundry Model Router; Snowflake Cortex AI Gateway dynamic routing	The eligible models, providers, or pairs; a routing criterion; exclusion of newly introduced models until added	Request level Allocation inside the customer's set	Eligibility and Change retained by configuration; Exit is a platform migration
Take rate: an independent gateway earning a share of routed spend, increasingly owned by financial intermediaries	OpenRouter, under agreement to be acquired by Stripe; Ramp Router	Allowed models and providers; residency and retention policies; a cost tier or a routing strategy, including customer weighted benchmarks in one case	Allocation; ranking of candidates by other customers' aggregate spend where the customer has set no objective; sampling and comparison of alternatives against its own route	Objective can be supplied by the market rather than the customer; traffic and telemetry retained by default; counterparty independence subject to acquisition
Affiliated: the router operator also supplies candidate models	Cursor Router, launched July 2026; a Meta router reported in development	Optimization modes that shift the cost and capability balance; other rights not examined here	Allocation among candidates that include the operator's own models	Alignment turns on whether affiliated candidates are favored once the authorized objective is accounted for

Sources: platform integrated, AWS (2026), Microsoft (2026), Snowflake (2026); take rate, OpenRouter (2026a, 2026b, 2026c), Stripe (2026), Bloomberg (2026), Ramp (2026a, 2026b); affiliated, Sawers (2026). Public documentation describes available functionality, not realized governance. Product facts are as observed on the access dates given in the references; the six arrangement comparison from which this table is condensed is held with the author's market notes.

Because mechanism sourcing and decision authority can be separated, build versus buy is too coarse as the primary governance question. The precise question concerns the mandate under which model choice is exercised.

1.3 Research question and construct

The paper asks: **how should an enterprise allocate rights over model choice when runtime selection among AI inference configurations is executed by a routing intermediary?** The primary case is an economically distinct intermediary allocating among configurations supplied by independent, affiliated, or vertically integrated providers. Internal and self hosted routing are limiting cases of the same framework, because internalization attenuates some interorganizational conflicts without eliminating objective mismatch among internal functions, information asymmetry, weak measurement, or dependence on external decision inputs.

The answer is organized around the **Model Choice Boundary**: the allocation of decision rights between an enterprise and one or more intermediaries governing which eligible AI inference configuration performs a workload, under what objectives and constraints, using what evidence, and subject to whose residual authority. The central thesis is that model choice is a bundle of separable rights; that routing allows enterprises to delegate high frequency Allocation without surrendering strategic control; and that the appropriate boundary depends on where relevant knowledge resides, how capable and aligned the intermediary is, how consequential error is, and how effectively delegated authority can be independently verified. A second claim qualifies the first: formal retention of a right is not real control over it. No routing mandate can specify the right response to every future model release, workload state, provider change, or evaluation failure, so residual authority over unanticipated states is consequential; and an enterprise can hold that authority formally while losing effective control if the intermediary holds the relevant information and the enterprise cannot measure outcomes or compare alternatives. A routing vendor exercising substantive discretion is therefore more than a software supplier. It is a delegated decision maker operating inside a bounded mandate whose continued alignment must be independently assessable.

1.4 Contributions and plan

The paper makes four contributions. It defines the Model Choice Boundary and separates it from the sourcing boundary, making visible the two off diagonal cases that build versus buy obscures. It decomposes runtime model choice into separable rights and derives the determinacy requirement on Objective stated above. It makes the workload routing mandate the primary unit at which delegated discretion is specified, with an enterprise policy floor beneath it and a portfolio result that shows why individually acceptable mandates do not eliminate shared exposure. And it integrates delegation with incremental routing risk and assurance: a scoping rule that separates routing exposure from baseline workload risk, a conditional allocation result showing why average quality and cost are insufficient evidence, a formal versus real authority condition, and a mandate relative identification result that yields a four level assurance stack.

Section 2 defines the boundary, the mandate, and the rights. Section 3 places the mandate at the workload level. Section 4 defines incremental routing exposure. Section 5 gives the decision rule for allocating rights. Section 6 develops real authority and assurance. Section 7 returns to sourcing. Section 8 discusses implications, the status of the claims, and the research agenda.

2. The Model Choice Boundary

A router does more than execute a comparison among models. It operates inside prior decisions about which configurations may compete, what objective is optimized, what evidence counts, when ordinary discretion ends, and who can revise or terminate the arrangement. Static assignment bundles those decisions into periodic architecture review. Dynamic routing separates them and lets different parties exercise them at different frequencies, with different information and different incentives. This section defines the division of that authority, the instrument through which it is granted, and the rights it contains.

2.1 Definition

> The Model Choice Boundary is the allocation of decision rights between an enterprise and one or more intermediaries governing which eligible AI inference configuration performs a workload, under what objectives and constraints, using what evidence, and subject to whose residual authority.

Two clarifications fix the construct. First, the boundary is a boundary of authority, discretion, and residual control, not a physical or architectural boundary, and it is not determined by where the routing software runs. It is not independent of architecture either. Architecture shapes which allocations of authority can be implemented, observed, changed, and transferred at reasonable cost: an enterprise can contract for Objective and still be unable to express it against a fixed opaque scoring function, or contract for Evidence and find it weak where the serving layer exposes no stable linkage between a request, a model version, and a policy version. Authority and architecture are analytically distinct and operationally interdependent. Second, the boundary is not one line. It can be drawn differently across workloads and differently across rights within one workload: an enterprise may delegate request level selection broadly for one workload, narrowly for a second, and not at all for a third, while retaining the same authority over objectives across all three.

The decomposition applies established logic to a new object. Fama and Jensen (1983) separate decision management from decision control and decompose organizational decisions into initiation, ratification, implementation, and monitoring; information technology governance has long defined itself as the allocation of decision rights and accountability across distinct decision domains (Weill and Ross, 2004), and already treats contemporary governance as a mosaic of rights dispersed across organizational boundaries that insourcing versus outsourcing cannot describe (Tiwana, Konsynski, and Venkatraman, 2013). What is claimed here is the routing specific instance: the rights that runtime inference allocation creates, the mandate that bounds high frequency Allocation, the separation of incremental exposure from baseline workload risk, and the conditions under which a retained right remains real.

Adjacent work operates at other levels: design time specification of a delegated autonomy boundary for agentic systems, framed as policy rather than contract (Arora, Vogelsang, and Sharma, 2026); control plane reference architectures (Futurum Group, 2026); decision sovereignty over commercially supplied models (Wei and Shu, 2026); intelligent choice architectures, the lineage for the term meta decision rights used here (Schrage and Kiron, 2025); and the allocation of human and machine participation into framing, acting, and learning (Sebastian, Weill, Haskamp, and vom Brocke, 2026). Two features separate the present object from all of these. The counterparty is economically distinct from the enterprise, which is why Incentive Alignment, Verifiability, Evidence, and Exit carry weight that an allocation between people and systems inside one organization does not require. And the system executes the supplier allocation rather than presenting it to a person for decision.

2.2 The routing mandate

> A routing mandate is the bounded authority granted to an intermediary to allocate workloads among eligible inference configurations within an enterprise authorized envelope.

The mandate is narrower than a supplier contract. A supplier agreement governs price, security, service levels, liability, data handling, and termination; the routing mandate governs the part of the relationship concerned with model choice discretion. At minimum it specifies the workload in scope; the eligible configuration set; the authorized objective and the quality, security, residency, regulatory, and other constraints that bind it; the decisions excluded from the grant; the discretion permitted within it; fallback and escalation behavior; evidence obligations; review cadence; the change process; and the enterprise's intervention and revocation rights. None of this requires a separate legal document. The mandate is the governance specification; supplier terms, platform configuration, policy as code, and operating procedures are the surfaces through which it becomes enforceable and executable.

Operational completeness. For workload w, let D_w denote the region of workload states in which ordinary delegated Allocation is authorized, let $\mathcal{E}_w$ denote the evidence or decision criterion that supports delegation over D_w, and let $I_w(x)$ denote the enterprise specified Intervention regime for states outside D_w. A mandate is operationally complete only when all three are defined: a region of ordinary discretion, a basis for treating that region as delegable, and a regime for the complement. The enterprise need not decide each exception at runtime. It must decide in advance what authority applies when the evidence supporting ordinary Allocation no longer does. The evidentiary standard can vary: a residency rule or a prohibited provider list may define part of D_w deterministically, while a claim that routing inside D_w satisfies a calibrated loss bound should be supported by evidence calibrated to that loss. The framework requires the claimed scope of discretion to be supported by evidence appropriate to the claim, not a probabilistic certificate for every mandate.

The instrument has established forms: authorization frameworks for delegated software agents that attenuate scope and specify expiry, revocation, and audit records (South et al., 2025); the delegated investment mandate, which binds an expert manager to an eligible universe, an objective, and monitoring and termination rights (ICGN and GISD, 2022), and whose narrowness is itself an agency cost response, since wider discretion weakens the link between performance and the manager's effort (He and Xiong, 2013); intermediate allocations of authority in which an agent acts without prior approval while the principal preserves a costly right to reassert control (Aghion and Tirole, 1997); and stop conditions treated as specification commitments in agentic requirements engineering (Arora, Vogelsang, and Sharma, 2026). The routing mandate specializes these for high frequency model choice by specifying the evidence supported region of ordinary discretion, the Intervention regime outside it, and the workload as the unit at which both are set.

2.3 Rights over model choice

Model choice is often described as one decision: the router selects a model. For governance that description is too coarse. The decision decomposes into rights that differ in frequency, information requirements, economic leverage, and suitability for delegation, in the way platform governance distinguishes strategic from implementation rights (Tiwana, Konsynski, and Bush, 2010). The decomposition is functional rather than a ranking, and it is not claimed to be exhaustive for every deployment. Table 2 is the canonical description; the text below adds what each right requires.

Table 2. Model choice rights

Right	What it decides	Structural tendency	Evidence required for the right to be real
Eligibility	Which configurations may compete: model, version, provider, endpoint, region, service tier, compute budget, fallback	Enterprise authority with intermediary input; joint approval or bounded delegation within inherited constraints	Reliable model, provider, endpoint, and policy identity; detection of unauthorized changes to the candidate set
Objective	What routing is authorized to optimize and which constraints bind, with the tradeoffs among goals resolved	Usually enterprise authority, with intermediary input where utility is enterprise specific	Measures that correspond to the authorized utility, including downstream review, remediation, and business outcome where the objective includes them
Change	Revision of the eligible pool, thresholds, objective, evaluations, fallback and classification logic	Enterprise approval, joint approval, or bounded delegated change	Version history, drift evidence, comparative evaluation, visibility into model, provider, classifier, and policy changes
Allocation	Which eligible configuration serves a request or workload instance	Often delegable when the intermediary has a material capability advantage	Route attribution and enough request context to test consistency with the mandate
Intervention	Override, escalation, narrowing, suspension, reversion when ordinary discretion should end	Enterprise retained or enterprise defined, including automated regimes	Measurement timely enough to trigger reversion while it can still reduce loss; a reversion path tested under the conditions that trigger it
Evidence	Route records, policy version, model identity, scores, evaluation outcomes, downstream attribution	Generated by an intermediary but requiring sufficient enterprise access and independence	Evidence whose provenance and reliability suffice to support the other retained rights
Learning and Data	Who may retain traffic and learn from prompts, outputs, labels, feedback, and telemetry	Retain, share, or price transfer according to strategic value	Records of reuse and retention sufficient to test compliance with the allocation
Exit and Portability	Exit is the formal right to terminate or replace; portability determines whether supporting assets can move	Enterprise retains Exit; portability governs its cost and effectiveness	Comparative evaluation and portable evaluation sets, route history, policy, and configuration sufficient to identify and move to an alternative

Structural tendencies are priors evaluated right by right in Section 5. Accountability and Assurance are cross cutting functions rather than additional rights.

Meta decision rights: Eligibility, Objective, Change

Three rights shape the architecture within which repeated Allocation occurs, and their leverage is multiplicative: one exercise alters many subsequent allocations without any further decision. That makes them structurally different from selecting one configuration. It does not make retention a rule. They tend toward enterprise retention because their exercise depends on enterprise specific knowledge and on consequences the enterprise bears, and inherited regulatory constraints bind regardless of the evaluation.

Eligibility determines which configurations may compete. It is an input control over the authorized set, in the sense in which control theory distinguishes input, behavior, and outcome controls (Ouchi, 1979; Kirsch, 1997). Authorized is not the same as available: outages, quota exhaustion, rate limits, or degradation can narrow which authorized configurations are executable without the Change right having been exercised. Allocation therefore operates over the intersection of the authorized and the currently feasible, and when no acceptable authorized configuration remains feasible the situation is no longer ordinary Allocation but the Intervention regime.

Objective defines what routing is authorized to optimize and which constraints bind that optimization. It is distinct from Allocation: the intermediary can still make every runtime choice while the objective determines what counts as a good one and so shapes the distribution of later allocations. Multitask incentive logic explains why this is a governance right rather than a configuration detail. When some performance dimensions are easier to measure than others, strong incentives on the measured dimension distort effort away from valuable but harder to measure ones (Holmström and Milgrom, 1991). A policy that minimizes inference price subject to a narrow benchmark can perform well on the observable objective while worsening review burden, remediation, or customer outcomes the objective omits, and the instruments of response, which include weakening the incentive, restricting scope, and changing the discretion granted, are one instrument system rather than separate controls (Holmström and Milgrom, 1994). Mandate breadth and incentive intensity are therefore joint design variables. Section 4 gives the point analytical form.

Retaining Objective also carries a **determinacy requirement**. A mandate that states several goals without specifying how their tradeoffs are resolved has not retained the right, because the missing preference ordering becomes delegated discretion the router must supply on every request where the goals conflict. Determinacy does not require a scalar formula. Hard constraints, ordered priorities, tie break rules, or a scalar objective over a constrained set can each resolve the tradeoffs the enterprise means to keep, and a dimension the enterprise treats as noncompensatory should remain a constraint or separate threshold rather than disappear into an average. Evaluation makes the requirement concrete, because collapsing a vector of quality, cost, latency, and reliability measures into a single ranking requires weights that no evaluator supplies neutrally, so metric aggregation exercises Objective whether or not anyone treats it as a governance decision. Two 2026 instances show the requirement being met and being displaced. A published enterprise router elicits criterion weights from a three person panel of the customer's own executives and adds a risk veto so that business risk cannot be compensated by cost or speed, which is the determinacy structure made explicit (Nowak, 2026). A commercial auto router instead ranks candidates for each task type by the platform community's aggregate spend share over a trailing seven day window, so that the ordering an enterprise has not supplied is supplied by the revealed preferences of others (OpenRouter, 2026a).

Change modifies the rules under which model choice occurs: adding or removing configurations, altering thresholds, revising the objective, changing the evaluation regime or classification logic, and responding to a change in provider affiliation. Because routing policies operate in a moving environment, the Change right determines who adapts the system to that movement, on whose approval, and model upgrades should be regression tested against authority boundaries and escalation conditions rather than aggregate accuracy alone (Arora, Vogelsang, and Sharma, 2026). Rules that live only inside a mutable intermediary controlled component have migrated outside the enterprise's approval process, a condition the accountability boundaries literature calls rule debt (Hydari and Muzaffar, 2026), and retaining Objective or Change has little content if the enterprise cannot identify, version, approve, and revise the rules that generate repeated Allocation.

The operational right and the Intervention right

Allocation determines which eligible configuration serves a particular request. It is the principal high frequency right that routing technology exercises, and delegation can be especially attractive when the intermediary has broad and rapidly refreshed information about performance, availability, latency, capacity, and price. Delegating Allocation does not require delegating Eligibility, Objective, or Change.

Intervention is the right to change who decides: to override a route, force escalation, narrow the pool, suspend dynamic routing, or revert to a static baseline. It is separated from the meta decision rights because it performs a different function. They define or revise the ordinary architecture; Intervention governs the return of control when a state falls outside the conditions under which ordinary delegation remains appropriate, and $I_w(x)$ is the regime applied outside D_w. A retained Intervention right is operationally real only if the reversion path remains executable under the conditions that trigger it: failover concentrates traffic, retries amplify load, and rate limits bind precisely when reversion is invoked, so fallback capacity and recovery behavior belong in mandate testing rather than being inferred from the existence of a fallback rule. Intervention does not require a human on every exception; the enterprise can exercise it in advance through an automated fallback, a deterministic policy, a review queue, or suspension, and the relevant question is who has authority to define the reversion mechanism.

Information and property rights

Evidence is the right to receive and control the information needed to reconstruct and evaluate model choice: route records, model and version identity, endpoint and fallback facts, confidence scores where produced, evaluation outcomes, downstream attribution, and the policy version in force at each decision. Evaluation results belong to it under the same discipline as routes, since a score is interpretable only alongside the versioned protocol that produced it. The right to observe a metric is distinct from the right to use it to alter future Allocation: where the same signal retrains a classifier, revises rankings, rewards the intermediary, or updates policy, control of it also engages Objective, Learning and Data, and Change, and those rights form one adaptive loop in any router that learns from what it is measured on. The sovereignty literature names the failure from the state side as audit asymmetry, in which system behavior is visible only through provider controlled interfaces (Wei and Shu, 2026).

Learning and Data determines who may retain traffic; use prompts, outputs, and labels; learn from outcome telemetry; improve the router; improve affiliated models; and aggregate across customers. It extends beyond privacy because routed traffic can improve the router, support cross customer learning, or improve an affiliated model, so its allocation affects future capability and competitive position. Section 7 develops the economics.

Exit and Portability governs the enterprise's ability to move the decision and its supporting assets elsewhere. Exit is the formal right to terminate, replace, or reclaim the arrangement. Portability is the enabling condition that determines how costly and effective that right is to exercise, by making routing policy, evaluation sets, route history, telemetry, model mappings, thresholds, and workload configuration reusable elsewhere. Exit preserves an outside option when Change, Intervention, or contractual remedy is insufficient.

Accountability and assurance

Accountability is not another right. It is the obligation that attaches to every delegated right.

> Every delegated model choice right should be paired with an identified enterprise owner who holds authority to ratify the delegation, monitor its exercise, and revise or withdraw it, and who carries evaluable responsibility for its consequences and continuing suitability.

This pairing follows from coupling delegated decision management with retained ratification and monitoring (Fama and Jensen, 1983), and it is consistent with field evidence in which deployed agents are assigned a named business owner accountable for outcomes and learning (Sebastian, Weill, Haskamp, and vom Brocke, 2026). Right specific ownership needs one further element. Distributing eight rights across architecture, risk, data, procurement, and platform functions can leave every right owned and the result unowned, so right level ownership should roll up to a named workload business owner accountable for whether the mandate is still producing the outcome the workload exists to produce.

Assurance is the independent function through which delegated authority is tested. It asks progressively whether the route occurred as claimed, whether it complied with the mandate, whether the outcome met the required bar, and whether the allocation remained aligned with enterprise interests relative to eligible alternatives. Section 6 develops it.

2.4 Own, Delegate, Verify

The decomposition separates three verbs that routing discussions conflate.

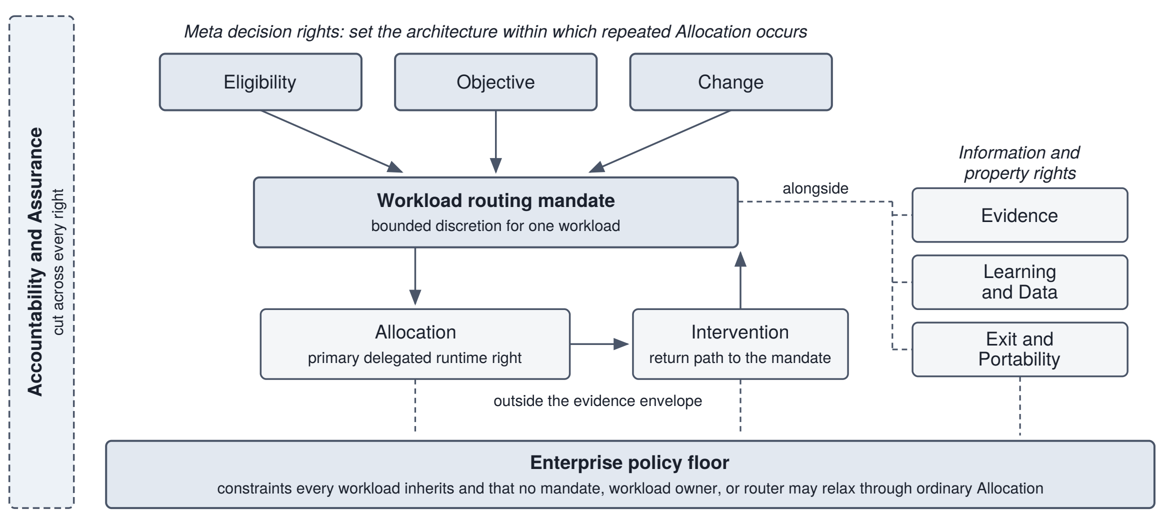
Own means retain residual authority over a dimension of model choice when the mandate or contract does not settle the state encountered. Incomplete contracts theory distinguishes specified contractual rights from residual rights of control where contingencies cannot all be written into the contract (Grossman and Hart, 1986; Hart and Moore, 1990), and no present mandate can specify every future model release, provider acquisition, pricing change, quality regression, or evaluation failure. Routing is a distinctive contracting environment: executable policy and automated route evidence make some dimensions unusually contractible, while model evolution, proprietary serving behavior, and semantic output quality keep others costly to specify or verify. Own is governance shorthand rather than ownership of a nonhuman asset. Each domain can contain specified rights, delegated discretion, and residual authority at once, and an enterprise owns Eligibility when it can remove a provider from the approved set even though a third party operates the router, owns Change when material revisions require its approval, and owns Exit when it can terminate and take the decision assets with it.

Delegate means grant bounded discretion under a routing mandate. Delegation can be broad or narrow, and it differs from surrender: graduated authority that tiers actions into those executed autonomously, those proposed pending approval, and those prohibited regardless of instruction is established practice in adjacent specification work (Arora, Vogelsang, and Sharma, 2026). The intermediary may make millions of Allocation decisions while the enterprise retains the authority that defines the choice set, the objective, the evidence obligations, and the reversion conditions.

Verify means establish, using evidence sufficiently independent of the delegated act, whether the authority was exercised as intended. Retaining a contractual right does not establish that the intermediary complied with it, and receiving an operator authored log does not establish that the allocation remained aligned with enterprise utility.

> Own defines residual authority. Delegate defines permitted discretion. Verify provides assurance over delegated exercise.

Accountability stays outside the triad: it identifies who inside the enterprise is responsible for the continuing suitability of the delegation. The separation makes the two off diagonal cases of Section 1.2 visible, and it implies that there is rarely one enterprise wide boundary. Rights can be allocated differently across workloads and reallocated as evidence, capability, or consequence changes, which is why the next section treats the workload as the unit at which mandates are designed.



Authority flows downward from the meta decision rights to runtime Allocation; the floor bounds all of it from beneath.
Rights are analytically separable; their optimal allocation need not be independent.

Figure 1. Model choice as a bundle of separable rights. Authority flows downward: the meta decision rights over Eligibility, Objective, and Change define the workload routing mandate, and Allocation is the primary delegated runtime right, with Intervention returning control to the mandate when a state falls outside the evidence envelope. Evidence, Learning and Data, and Exit and Portability sit alongside the mandate as information and property rights. The enterprise policy floor beneath is the set of constraints every workload inherits and no mandate or router may relax. Accountability and Assurance cut across every right.

3. The Workload Mandate and the Enterprise Policy Floor

There is no enterprise wide best router for the same reason there is no enterprise wide best hosting model: the appropriate architecture depends on the workload. Different workloads carry different quality requirements, latency tolerances, data constraints, consequences, evaluation evidence, and degrees of reversibility, so the same enterprise can rationally run different mandates, eligible pools, and sourcing arrangements at once. What transfers from infrastructure placement is the workload level perspective, which begins with requirements rather than products and separates a workload decision from an enterprise standard. What does not transfer is the governance model, because routing operates at request frequency, semantic output quality is harder to observe than infrastructure performance, routed traffic generates learning for the intermediary, and the intermediary has its own incentives over which configuration is selected. Routed AI therefore delegates recurring judgment about quality and economic utility, and workload placement must be paired with explicit decision rights and assurance.

3.1 Two levels and a design rule

The framework separates an **enterprise policy floor** from the **workload routing mandate**. The floor contains rules that bind all workloads or broad classes of them: prohibited providers, residency, security controls, minimum auditability, retention conditions, route logging requirements, intervention rights, and change control. It denotes constraints that workload owners and routing intermediaries are not authorized to relax through ordinary Allocation.

A design rule governs what belongs at which level: standardize the governance primitives that must compose across workloads, and localize the decisions whose value depends on workload specific knowledge. Identifiers, minimum route evidence schemas, policy and evaluation versioning, provider qualification, a common Intervention interface, portability requirements, conflict disclosure, and incident records belong in the first category, because portfolio review and substitution are impossible when each workload records them differently. Objective, eligible subset, quality thresholds, evidence envelope, consequence assumptions, and review intensity above the enterprise floor belong in the second. Each level states its own authority, the decisions reserved to the enterprise, and the escalation path when they conflict.

The workload routing mandate then specifies the discretion available for workload w:

$$M_w = \{E_w, U_w, C_w, I_w, A_w\},$$

where E_w is the eligible set of inference configurations, U_w the enterprise authorized objective, C_w the inherited constraints, I_w the Intervention and escalation regime, and A_w the required assurance. Runtime Allocation is authorized only within the mandate. The notation is compact by design: its purpose is to separate five things that are usually collapsed into one routing policy, namely what can be chosen, what the router is supposed to optimize, what optimization cannot trade away, what happens when ordinary discretion should end, and what evidence is required to establish that the delegation remains acceptable. The architecture has three levels. The floor establishes common constraints, the mandate translates workload requirements and enterprise rules into bounded discretion, and the intermediary exercises Allocation inside it. A workload owner may have authority to tune routing within the mandate while lacking authority to weaken a residency rule or reduce an enterprise wide logging requirement; routing should not silently turn an optimization problem into a mechanism for bypassing upstream controls.

Two classes of input stay separate. Workload characteristics, such as data classification, quality and latency requirements, existing controls, reversibility, evaluation availability, and strategic sensitivity, determine what the routing policy must preserve, and the framework inherits them from existing governance processes rather than reproducing a risk taxonomy. Governance determinants, developed in Section 5, determine who should control and exercise that policy.

3.2 Workload certification does not imply portfolio independence

An enterprise can assess each workload separately and still miss exposure created by shared components. With portfolio loss $L_{\text{portfolio}} = \sum_w L_w$, expected loss composes linearly, but dispersion does not:

$$\mathrm{Var}(L_{\text{portfolio}}) = \sum_w \mathrm{Var}(L_w) + 2\sum_{w<v} \mathrm{Cov}(L_w, L_v).$$

Two portfolios can have the same expected loss workload by workload and different portfolio variance because their losses move together differently. Routing creates or amplifies that dependence when workloads share a model family, provider, endpoint, routing classifier, policy, gateway, or control plane. Recovery policy is a channel that is easy to miss because primary routes look diversified: workloads with distinct primary configurations become coupled when they share a fallback path, and one broad outage can trigger simultaneous failover onto it. Ownership and affiliation add a further channel when apparently separate services become subject to one strategic decision. Ten individually acceptable mandates that depend on one model family, one classifier, or one gateway can therefore produce positively correlated losses while every workload remains compliant. The identity is standard and variance is only one concentration measure; the lineage runs through algorithmic monoculture, where shared models homogenize outcomes across decisions (Bommasani et al., 2022). The governance implication is the point:

> Workload level mandate adequacy does not by itself establish portfolio level assurance when workloads share failure channels.

A workload owner evaluating only w cannot internalize dependence with v unless it is imported into the analysis, so some concentration controls belong above the individual mandate: limits or review thresholds on shared exposure, or a portfolio review that identifies common dependencies, tests correlated failure scenarios, and decides whether diversification, fallback independence, or additional Intervention capability is warranted.

3.3 Heterogeneous mandates as a prediction

Once the workload mandate is the unit of design, heterogeneous routing architectures inside one enterprise are an expected result rather than an inconsistency. A high volume commodity workload can support a commercial dynamic router with a broad Allocation mandate where enterprise specific knowledge is limited, the capability advantage is material, and outcomes are readily verifiable. A proprietary workload can use commercially sourced gateway mechanics while the enterprise retains the taxonomy, Objective, evaluations, thresholds, and Evidence rights that depend on firm specific knowledge. A high consequence or poorly measurable workload can justify a narrow pool, a tight evidence envelope, intensive assurance, or static assignment for some states. Static assignment is one possible mandate outcome, not the definition of good governance, and a commercial router can be appropriate for a high consequence workload if capability and assurance are strong. Variation in workload requirements and inherited constraints should therefore produce heterogeneous mandates and sourcing configurations within the same enterprise, and hybrid sourcing is a prediction of the framework rather than an anomaly. An enterprise level build or buy label obscures exactly this: the governance question is whether each workload's arrangement preserves the required rights, constraints, evidence, and exit options.

4. Incremental Routing Exposure

Dynamic routing does not create the underlying risk of using AI for a workload. A fraud decision can be wrong without a router; sensitive information can be exposed by a statically assigned model; a hallucination can occur when every request goes to the same approved configuration. Treating all such failures as routing risk would make the Model Choice Boundary indistinguishable from a general AI risk framework. The narrower question is what changes because a model choice layer now sits between the workload and the configuration that serves it, and the relevant comparison is with the enterprise's approved model choice policy absent dynamic routing, not with a hypothetical world in which the strongest model handles every request.

> A risk that exists materially unchanged under static assignment is baseline workload risk. Where routing changes its probability, distribution, control, or observability, only that incremental change falls within the framework.

4.1 Counterfactual baseline and three categories

Let x denote a workload or request state, which can include contemporaneous operational information such as provider health, latency, queue state, or quota where the policy conditions on it. Let $B(x)$ be the configuration selected under the enterprise's approved static policy, $R(x)$ the configuration selected by the routing layer, and $L(m, x)$ the task level loss when state x is served by configuration m, measured in the workload's own quality terms. A basic measure of selection loss is

$$\Delta L_{\text{selection}} = E\left[L(R(x),x) - L(B(x),x)\right].$$

The baseline is not "always use the most accurate model." An enterprise may have approved different static configurations for different workloads, jurisdictions, or tiers, and the counterfactual isolates the effect of introducing dynamic choice rather than imposing a technical optimum. An error the static and routed systems make identically, at the same rate and under the same controls, is not an incremental routing loss.

Table 3. Baseline workload risk versus incremental routing exposure

Category	Definition	Illustrative examples	Governance treatment
Baseline workload risk	Exposure that exists materially unchanged without dynamic routing	Substantive model error, hallucination, sensitive data mishandling by an approved configuration	Inherit existing workload controls; do not relabel as routing risk
Routing induced exposure	Exposure created by the model choice layer	Misallocation, unauthorized fallback, policy drift, hidden eligibility change, classifier manipulation, model substitution	Govern through the mandate, control rights, and assurance
Routing amplified exposure	Preexisting exposure whose probability, distribution, control, or detectability changes under routing	Broader provider exposure, request level quality variation, common gateway dependence, harder outcome attribution	Isolate and govern the incremental component relative to the approved static baseline

Routing induced exposure is no longer only theoretical. The routing control plane can be attacked directly (Shafran, Schuster, Ristenpart, and Shmatikov, 2025), and measurement studies of live gateways and reseller endpoints now document silent downgrades, model switches, and billing deviations (Lin et al., 2026), identity verification failures at close to half of the third party wrapper endpoints examined, including wholesale substitution of a much smaller model for a frontier one (Zhang et al., 2026), and a commodity router population numbering in the hundreds (Liu et al., 2026). Whether any of this affects a given enterprise is an empirical question; that the layer fails in these ways at scale is not.

4.2 Three mechanisms

Incremental exposure enters through three mechanisms that can coexist and are not assumed to be statistically independent or additive.

Selection risk asks a narrow question: assuming the mandate is appropriate and the routing layer behaves as authorized, does dynamic selection create greater enterprise loss than the approved static counterfactual? The technical literature reports average quality and cost tradeoffs, and that evidence is necessary but insufficient, because routing changes the conditional assignment distribution. A small average gap between two configurations does not establish that the requests sent to the cheaper one have a small loss gap; a low frequency subtype can have little effect on an aggregate benchmark while carrying large conditional loss if the router systematically selects the wrong configuration for it.

Consider two configurations, an enterprise approved reference S and a lower cost eligible C. Let $r(x) \in [0,1]$ be the probability that the policy assigns state x to C, let $d(x) = E[L(C,x) - L(S,x) \mid x]$ be the signed incremental loss from using C, let $c(x) \geq 0$ be an enterprise consequence weight, and let $h(x) = c(x)d(x)$. The consequence weighted selection loss attributable to routing to C is $E[r(X)h(X)]$, and by the covariance identity

$$E[rh] = E[r]E[h] + \mathrm{Cov}(r,h).$$

The first term is the loss of a nonselective benchmark that sends the same average share of requests to C independently of state. The second isolates conditional selection. If $\mathrm{Cov}(r(X), c(X)d(X)) > 0$, the policy disproportionately assigns C to states with higher consequence weighted incremental loss and performs worse on this measure than nonselective assignment at the same average use of C; if the covariance is negative, conditional selection improves the distribution. The relevant object is $c(x)d(x)$, not consequence alone: a high consequence request creates no adverse term if C has no material disadvantage for it, and a modest weight can matter when the gap is large.[1] The nonselective allocator is an analytical benchmark; the governance baseline remains $B(x)$. Instance dependent cost sensitive learning, learning to defer, and budget constrained routing already formulate allocation under heterogeneous errors and costs (Castiñeiras Rella et al., 2025; Madras, Pitassi, and Zemel, 2018; Li, Gao, and Lakshmanan, 2026). The contribution is the governance reading of an elementary decomposition: it makes explicit why the Objective right determines whether enterprise consequence, downstream loss, and full economic cost enter the routing criterion, and it gives empirical work a target, which is not how often a cheaper model is used but whether that use is concentrated where consequence weighted loss is, and whether any such concentration is justified by a matching concentration of savings. The result is static over a given workload distribution or evaluation window and does not model how routing itself changes future demand, capability, capacity, or prices.

Control risk begins when the routing layer does not behave as authorized: hidden fallback or substitution, unauthorized policy changes, expansion of the eligible set, evidence manipulation, adversarial rerouting, or self preferencing inconsistent with the authorized objective. These failures can occur when every model performs exactly as expected, so evidence about model quality does not bound them. The verification and measurement literatures cited above sit at this layer, and Section 6 addresses the corresponding assurance.

Mandate specification uncertainty exists when the router behaves exactly as authorized and the mandate itself proves incomplete for the state encountered in production: its objective, constraints, eligible set, workload classification, or evidence base did not anticipate that state. This is the routing layer expression of incomplete contracting. If the intermediary expands the eligible set without authority, the failure is control; if the enterprise approved the set but later discovers that a consequential production state was not represented in the evidence supporting it, the

[1] A cost matched version replaces equal routing share with equal expected savings. With $s(x) > 0$ the savings from assigning x to C, the nonselective allocator with the same expected savings uses C with probability $q = E[rs]/E[s] \in [0,1]$, and the loss difference becomes $D = \mathrm{Cov}(r,h) - \frac{E[h]}{E[s]}\mathrm{Cov}(r,s)$. Where $E[h] \geq 0$ and the router concentrates cheap Allocation in high savings states, the second term works in its favor; a positive association between routing and consequence weighted loss works against it.

failure is specification, and the remedy is a narrower mandate, better evidence, or a different Intervention regime rather than stronger enforcement.

4.3 The evaluated region and the routing evidence envelope

Evaluation evidence can characterize known selection risk only within the states and alternatives it represents. Within an evaluated region, observed differences among configurations can bound the losses from choosing one rather than another, and shadow evaluation, challenger models, and downstream attribution sharpen the estimate. The gap between the routed portfolio and the static baseline says nothing about control failures or about states outside that evidence, and three consequences follow. Support is protocol bound: a measured difference characterizes exposure only for the scenario distribution, configurations, adaptation procedures, metrics, and pipeline that produced it (Liang et al., 2023). The object of evidence must match the claim: general benchmarks can inform Eligibility, while evidence used to justify delegated Allocation must evaluate the routing policy over the workload distribution and objective for which discretion is claimed, because a router can allocate poorly among individually well evaluated models. And an economic objective is not self policing: on forgiving task formats, a strategy no better than random can register a lower cost per successful outcome than a capable one unless the objective excludes degenerate strategies and penalizes unreliability (Erol et al., 2025). Whoever specifies the objective decides which strategies the routing system is permitted to find.

The **routing evidence envelope** is the evidentiary side of operational completeness:

> The routing evidence envelope is the region of workload states for which the enterprise has sufficient evidence to justify the authorized level of dynamic routing discretion.

Inside the envelope, ordinary Allocation proceeds under the mandate. Near its boundary, sparse evidence, novelty, drift, or a new model version may justify a narrower pool, additional evaluation, or closer monitoring. Outside it, the enterprise specified Intervention regime governs, and that regime need not route every uncertain request to a human or to the most capable model: it can select a safer static tier, narrow the pool, require review for a request class, or suspend dynamic routing until evidence improves. Reject rules, selective prediction, learning to defer, and risk controlled decision systems formalize withholding or redirecting a prediction when evidence is inadequate (Chow, 1970; Madras, Pitassi, and Zemel, 2018; Mozannar and Sontag, 2020; Angelopoulos et al., 2024); the envelope operates one level up, governing not whether a model should answer but whether the intermediary remains authorized to choose among configurations for this state. Evidence supporting an envelope should identify its material gaps as well as its covered regions, and the envelope is a calibrated statistical region only where the mandate expressly relies on a procedure that supplies one.

> **P1.** Specification uncertainty. As uncertainty increases that a workload lies within the evidence supporting the routing policy, the rational scope of unreviewed or unconstrained Allocation decreases, requiring narrower eligibility, escalation, or stronger assurance.

The response can itself be automated. The proposition says only that the ordinary Allocation mandate should not silently continue unchanged as confidence in its supporting evidence falls. Consequence is a moderator rather than a further directional proposition: it raises the cost of a wrong Allocation and therefore the value of capability, alignment, evidence, and Intervention, without determining by itself whether Allocation should be retained or delegated.

A stylized case ties the mechanisms together. An enterprise moves a support workload from static assignment onto dynamic routing that sends requests classified as simple to a cheaper configuration. A small subtype is linguistically simple, carries materially higher business consequence, and differs from the cases in the evaluation evidence. The router sends it to the cheaper configuration; average benchmark quality stays close to the approved level, inference cost falls, and conditional loss on the subtype rises. Selection exposure is the positive association between cheap Allocation and consequence weighted loss; specification uncertainty is the mismatch between the authorized region and its evidence; observability is the enterprise's easier view of spend than of downstream loss; and Intervention is what should govern once the request lies outside the evidence supporting ordinary Allocation. The example is illustrative, not evidence that enterprises commonly route consequential requests to weaker models.

Box 1. Three insufficiency results

The paper's formal statements share one structure: evidence aggregated at one level does not bound the distinction a retained right requires at another.

Across workloads. $\mathrm{Var}(\sum_w L_w) = \sum_w \mathrm{Var}(L_w) + 2\sum_{w<v} \mathrm{Cov}(L_w, L_v)$. Workload by workload adequacy does not bound portfolio dispersion when shared components make losses covary (Section 3.2).

Across request states. $E[rh] = E[r]E[h] + \mathrm{Cov}(r, h)$. Average quality and average cost do not bound consequence weighted selection loss when the policy concentrates cheap Allocation where the incremental loss is (Section 4.2).

Across policies. If two routing policies π_1, π_2 induce the same distribution of observable evidence, $P_Z^{\pi_1} = P_Z^{\pi_2}$, while the mandate treats them differently, $M(\pi_1) \neq M(\pi_2)$, then every test built from Z has power equal to its size. Evidence that cannot separate the policies the mandate distinguishes cannot support the right that distinguishes them (Section 6.2).

None of the three is new mathematics. Their role is to make three governance objects precise: the portfolio review, the Objective right, and the sufficiency of evidence.

5. Where Should the Boundary Be Drawn?

Section 2 made the rights separable and Section 4 showed that their delegation cannot be evaluated from average quality and cost. The remaining question is which rights to retain, share, or delegate. Five variables serve as the decision engine: **Knowledge location, Capability, Incentive Alignment, Verifiability, and Consequence**. They apply to each right separately, they are not additive scores, and they do not form a maturity ladder. Meta decision rights often depend on enterprise specific knowledge and consequences, which favors retention; Allocation often benefits from rapidly changing market information and technical capability, which favors delegation. Both tendencies are priors. One boundary lies outside the test: where law, regulation, contractual commitment, or the enterprise policy floor imposes a nonnegotiable constraint, that requirement limits the feasible mandate before the delegation decision is made.

5.1 Knowledge location

Decision rights should sit where the costly to transfer knowledge needed to exercise them resides (Jensen and Meckling, 1992). In routing that knowledge is split. An intermediary may hold superior general market knowledge: current latency, availability, price movement, capacity, failure states, and comparative performance across a broad portfolio. The enterprise holds specific knowledge: what the workload means, which request classes matter, what downstream failure costs, how review and retry labor is incurred, what customer and regulatory consequences attach to an error, and which tradeoffs are acceptable. Split knowledge can justify split rights, and the split is right specific: Eligibility can turn on proprietary workload and regulatory knowledge while Allocation turns on provider conditions, and Change can require both, with the intermediary detecting a market shift first and the enterprise deciding whether it warrants revision.

> Knowledge locality. The greater the enterprise specificity and transfer cost of the knowledge required to exercise a right over model choice, the stronger the case for retaining that right or tightly constraining its delegation.

The argument is a collocation argument, not an internalization rule: internal teams can lack market breadth and external intermediaries can be given enterprise specific information. Delegation changes which knowledge the enterprise needs without eliminating the requirement, since a client's peripheral knowledge of an outsourced domain complements outcome based control rather than being displaced by it (Tiwana and Keil, 2007). And where a right depends on both sides at once, as Change does, assigning it jointly supplies no integration; that requires shared domain knowledge and a working mechanism through which the two bodies of knowledge are actually combined (Kearns and Sabherwal, 2006).

5.2 Capability

Capability asks whether the intermediary exercises the right materially better or faster than the enterprise could. It is distinct from knowledge, since a party can hold information and lack the systems, evaluation process, or scale to use it. For Allocation it has two components that move independently. Selection capability concerns whether the intermediary chooses an appropriate configuration for the request state and rests on classifier discrimination, generalization to the workload distribution, and comparative evidence. Execution capability concerns whether it can realize that choice under actual conditions and rests on end to end latency including routing overhead, tail behavior, burst handling, retry and fallback behavior, and recovery. A strong selector can be a weak runtime delegate and the reverse, so offline routing accuracy does not establish production capability for a workload whose binding constraint is tail latency under burst. The routing literature also shows that no single model is best across tasks, which is the case for routing at all, and that a well tuned nearest neighbor router can match or outperform far more elaborate learned ones (Hu et al., 2024; Li, 2025), so capability must be established for the relevant workload and right rather than inferred from the sophistication of a product. And capability is insufficient alone: a highly capable intermediary is a poor delegate if its incentives diverge or its behavior cannot be verified, while weak enterprise capability can make delegation valuable even at high consequence, provided the mandate is narrow and assurance is strong.

5.3 Incentive Alignment

The determinant is incentive alignment in the agency sense: congruence between the delegate's economic payoff and enterprise utility over the delegated right, read as a residual after contractual incentives, pricing, affiliation, provider relationships, and data value have been accounted for. It is not business and information technology strategic alignment, a distinct construct. The paper uses Alignment as shorthand for the incentive sense. Alignment matters most for Objective and Allocation, because if the intermediary chooses both what counts as value and which configuration delivers it, a conflict enters twice.

The three intermediary types in Table 1 carry different residual divergences. A **platform integrated** router earns primarily from the platform relationship rather than from the route, which limits route level conflict while creating exit costs that Section 7 examines. A **take rate** router earns a share of routed spend, and a spend management or payments operator additionally holds the customer's spend data, so its interest in volume and in learning from traffic is structural even where it has no candidate of its own. An **affiliated** router is also a candidate supplier. The mechanisms behind these divergences are classical. Where inference cost is readily metered and application specific quality is not, external provision strengthens the supplier's incentive to reduce cost, and cost reduction becomes excessive precisely when the supplier does not internalize quality that is difficult to contract on (Hart, Shleifer, and Vishny, 1997). Where the router captures value from more than one side, the structure of payments across sides shapes conduct, so the enterprise facing fee does not characterize routing incentives (Rochet and Tirole, 2003). A model of a provider routing between a cheaper and a slower reasoning model for a user who can re prompt or abandon derives a misalignment gap where the provider's ranking by cost per successful pass diverges from the user's ranking by value net of latency, and finds that in nearly all regimes the provider optimal policy is a static assignment (Mahmood, 2026), which supports treating static assignment as a legitimate mandate outcome rather than a governance failure. That model's principal can only leave; the rights of Section 2 are an enterprise's substitutes for a consumer's power to walk away. A dual role platform model adds that traffic allocated to a supplier affects future quality through learning by serving, and that self preference alters those flows (Zhang, Bao, and Li, 2026).

These mechanisms establish why alignment belongs in the boundary decision. They do not justify assuming bias in every integrated intermediary, and affiliation enters only through the payoff divergence it creates, in either direction. Lower alignment also does not mechanically require internalization; it changes the terms of delegation.

> Incentive Alignment. Greater divergence between intermediary and enterprise utility narrows the routing mandate or increases the assurance required for a given level of discretion.

The formulation deliberately allows two responses: reduce delegated discretion, or hold discretion constant and invest in assurance. Which is efficient depends on Verifiability. The same residual logic applies to Objective itself: a formally retained Objective is weak when the measures and rewards governing delegated exercise systematically favor a different one, so governance design should test consistency among the authorized Objective, the measures used to evaluate routing, and the payoff attached to them.

5.4 Verifiability

Verifiability asks whether the enterprise can independently establish that delegated model choice remained within the mandate and aligned with its interests. Hidden action is less costly when the principal can observe or infer what the delegate did, and the relevant property is informativeness rather than volume: a signal earns value by adding information about the delegated action conditional on evidence already held, so a noisy measure that separates the intermediary's choices from exogenous shocks can contribute more than a precise one that does not (Holmström, 1979). The key word is independently. Evidence authored only by the intermediary is disclosure, weakest exactly where the question is whether the intermediary followed the declared route, changed policy, substituted models, or favored an affiliated option. Verifiability is also right specific: a record that Model A served a request establishes execution without showing that Model A was the best eligible choice, and a signed policy establishes the authorized rule without showing that it remains appropriate.

> **P2.** Verifiability. Holding other determinants and verification costs constant, greater independent verifiability expands the set of model choice discretion that can rationally be delegated.

The proposition cuts against a common intuition that stronger measurement is a reason to bring decisions inside the enterprise. Better measurement can make external delegation safer, because it reduces information asymmetry and lets the enterprise retain real authority while granting more operational discretion. Two qualifications follow from the multitask logic already used. Verifiability does not improve uniformly: route identity, latency, and inference cost become cheap to verify long before semantic quality and downstream consequence do, and a partial improvement can be worse than none if it expands delegation on measurable dimensions while the consequential unmeasured ones carry the widened discretion, so an enterprise should ask which dimension improved before treating better evidence as a reason to delegate more. And verification costs, including the exposure of workload content, customer data, and proprietary policy that collecting and sharing assurance evidence can create, belong in the same calculation. The proposition is a feasible set result: independent evidence makes a broader set of delegated arrangements governable, conditional on the other determinants.

5.5 Consequence

Consequence is inherited from the workload and business context. It is not a routing layer risk; it is a weight on incremental routing exposure. Higher consequence raises the cost of errors in every right and the value of accurate evidence and timely correction, and where material it includes reversibility, timing, and recoverability, which can change the value of Intervention even at similar expected loss. The intuitive response, less delegation, is too simple. A high consequence workload can rationally use substantial dynamic routing when the intermediary has superior capability, the mandate is tight, incentives are aligned, and outcomes are verifiable; a low consequence workload can still justify retention if the enterprise holds unique knowledge and the intermediary adds little. Consequence therefore moderates: it usually increases the value of narrower eligibility, stronger assurance, better escalation, and more conservative confidence thresholds without determining who holds Allocation. Treating it this way also keeps the framework from becoming a second risk classification, since it can inherit an existing consequence rating and ask only how much routing discretion that consequence justifies.

5.6 Decision rule

The five variables are applied to each right for each workload, and the output is a posture rather than a score. Four postures suffice. **Retain**: the enterprise holds final authority and grants no substantive discretion, though the intermediary can and often should supply input. **Joint approval**: material exercise requires affirmative authority from both parties, or each holds a defined veto. **Delegate narrowly**: the intermediary exercises the right within a constrained mandate, pool, or evidence envelope. **Delegate broadly**: the intermediary receives substantial discretion subject to retained accountability, evidence, Intervention, and Exit. Two guards apply. Where one right's value depends on another, as with Objective and Evidence or Eligibility and silent version Change, the bundle is assessed jointly. And decision input differs from decision authority, so broad input is compatible with narrow authority; a rights allocation is easier to make real where the architecture exposes controls and evidence surfaces corresponding to it, and the decomposition does not assume that any product does.

Table 4. Boundary determinants and directional implications

Variable	Condition favoring more delegation	Condition favoring retention or a tighter mandate
Knowledge location	Relevant knowledge concentrated with the intermediary or cheaply transferable to it	Enterprise specific knowledge is critical and costly to transfer
Capability	Material intermediary performance or speed advantage on the relevant component	Weak, unproven, or no capability advantage
Incentive Alignment	Incentives closely track enterprise utility	Material payoff divergence, including affiliation or commercial arrangements that create a relevant conflict
Verifiability	Strong independent evidence and outcome observability on the dimensions the mandate cares about	Opaque execution or weak independent evidence
Consequence	Lower downside from an incorrect exercise	Higher downside, raising the value of tighter discretion and assurance

The table records a reading at a point in time, because the determinants are not exogenous to the allocation chosen. Formal authority changes each party's incentive to acquire and communicate information (Aghion and Tirole, 1997), so Knowledge and Verifiability respond to the boundary; the composition of the eligible set determines which verification mechanisms are available; delegated Allocation determines which comparative evidence exists later; and the allocation of Evidence and Exit rights changes both sides' investment incentives, so today's boundary conditions tomorrow's Capability. A sound application allocates rights, asks whether the allocation degrades the determinants it assumed, and revises where it does.

Three cases show the variables at work. Allocation for a generic internal workload, with intermediary advantage, limited conflicts, observable outputs, and modest consequence, supports broad delegation. Objective for a customer facing workload whose downstream economics the intermediary cannot observe favors retention even when the same intermediary is excellent at Allocation, so the rights split. The third case runs the variables on an internal delegate, because the framework claims to govern authority rather than the legal boundary of the firm. Consider a platform or financial operations team that operates the router, owns the route telemetry, and is measured on inference spend. Knowledge favors that team on provider conditions and the business unit on workload consequence, so the split has the same shape as the external case. Alignment is weaker than the internal setting suggests, because a unit rewarded on a cheap to measure dimension while controlling the evidence that would reveal the expensive one is the multitask problem in its purest form. Verifiability is weaker still, for a reason the external case lacks: there is no counterparty whose records can be demanded, no audit clause, and no contractual remedy, so assurance depends on separating the evaluation function from the operating function. The response is not to insource further, since the delegate is already internal. It is to separate evaluation from operation, set the objective outside the operating unit, and require comparative evidence the operating unit does not produce. Internalization attenuates some conflicts, creates others, and removes the contractual instruments available for the ones it creates.

The structural tendencies can also reverse: an enterprise with little internal expertise may let a vendor configure much of Eligibility for a generic workload it can independently test and cheaply exit, while a high consequence institution may retain parts of Allocation through deterministic policy routing behind an external gateway. Discretionary enterprise constraints, unlike inherited legal ones, carry an opportunity cost, because a narrower pool reduces agency exposure while suppressing the value of the delegate's knowledge and capability. Read together, the five determinants are one economic tradeoff: Knowledge and Capability capture the cost of poor decisions when a right sits far from the relevant knowledge and ability, Alignment and Verifiability capture the agency and control costs created when discretion sits with a party whose payoff can diverge and whose behavior is costly to evaluate, and Consequence scales the loss attached to error on either side. That is why no universal direction toward retention or delegation exists.

6. Real Authority and Assurance

Aghion and Tirole (1997) distinguish formal authority, the right to decide, from real authority, effective control over the decision, and make the information structure the source of the difference. Applied to model choice, an enterprise can retain contractual authority over every right in Table 2 and still lack control. It may retain Eligibility without reliable knowledge of which version or endpoint served a request; Objective without the downstream measurement the objective is supposed to optimize; Change without evidence that a model update degraded performance; Intervention while discovering failures only after the harm is irreversible; Exit while lacking the portable evaluation assets and comparative evidence needed to identify an alternative. Eligibility carries a less visible version of the same problem, because approving a menu is not shaping it: an enterprise can hold formal approval over every addition while the intermediary determines which models it integrates, which candidates it evaluates, and which comparisons reach the enterprise at all, so curation settles much of the decision one stage before Allocation.

The difficulty is sharpest for learned routers, whose behavior must be inferred from inputs, outcomes, and tests rather than read from a policy. Control theory predicts the consequence: where behavior observability is low, input and outcome controls carry more of the burden (Ouchi, 1979; Kirsch, 1997; Choudhury and Sabherwal, 2003; Tiwana, 2008). Eligibility functions as input control, the outcome and alignment levels below carry the monitoring burden, and architecture itself, through externally controlled allowlists and independent enforcement points, can implement part of the regime before supervision is invoked (Tiwana, Konsynski, and Venkatraman, 2013). None of this makes learned routers less governable than deterministic ones; a rules system nobody measures can be governed worse than a learned one with strong evidence and intervention.

> **P3.** Real authority. Formal retention of a model choice right will not yield effective control absent evidence sufficient to distinguish the states that matter to that right and feasible capability to correct, constrain, or exit when those states require action.

The proposition is right specific, and the corrective capability can take different forms: technical denial, Intervention, Change, contractual remedy, or Exit. Measurement is necessary for most retained rights and sufficient for none. Its implication runs in a direction that is easy to miss. Weak measurement makes external delegation unsafe; stronger measurement can therefore support more delegation rather than pushing toward internalization, which is the mechanism behind P2. Internalization does not solve the measurement problem, since an enterprise operating its own router can still lack outcome attribution, comparative evaluation, or evidence that a classifier remains appropriate. The rest of this section states what sufficient evidence means and what levels it comes in.

6.1 Disclosure versus verification

> Disclosure is the routing operator's statement of what it did. Verification is the enterprise's independent establishment of the claims relevant to the rights it has retained: that the route occurred as claimed, that it complied with the mandate, that outcomes met the required threshold, and that the allocation remained aligned with enterprise interests.

Disclosure is useful and insufficient. Per request route records, identifiers, fallback facts, scores, and policy versions are necessary inputs, but a log created inside the operator controlled service is evidence supplied by the party being evaluated, and its weight depends on the claim and on the operator's ability and incentive to shape it. Supplier self control, such as internal route monitoring and regression gates, reduces the probability of failure without establishing the claims a retained right depends on. Agency cost has three parts, the principal's monitoring expenditure, the agent's bonding expenditure, and the residual loss that remains after both (Jensen and Meckling, 1976), and intermediary supplied attestation, receipts, audit commitments, and service credits are bonding whose credibility changes how much enterprise funded monitoring is worth buying.

Independence is relative to the component in dispute, and assurance is threat model relative. A signature establishes who produced a record and that it has not changed since, which is provenance rather than truth: a provider signing its own route record produces authenticated disclosure, strong as evidence to a downstream auditor and weak as evidence against the provider itself. Evidence that settles accidental misconfiguration can be insufficient against a strategic intermediary able to shape both the service and the telemetry describing it, and naming the adversary is part of specifying the requirement rather than an accusation. Where the mandate relies on independent testing, assurance also requires practical testing rights: access to run the authorized tests, preserve the evidence, reproduce a material finding, and escalate it through an agreed remediation path (Longpre et al., 2025). An audit clause paired with rate limits or prohibitions on repeated probing leaves verification formal in exactly the sense P3 gives the word.

6.2 Mandate relative identification

Box 1 states the identification result: if two feasible routing policies induce the same distribution of the evidence Z observable to the enterprise while the mandate property M distinguishes them, then any test built from Z has power equal to its size, and the enterprise cannot separate them with that evidence. Depending on the retained right, M can be route identity, authorization status, outcome quality, or relative enterprise utility. The statement is close to a definition, and its value is in what it rules out.

> Evidence is sufficient for exercising a retained model choice right only if it separates the policy differences that matter to that right.

It is a necessary condition, structural rather than finite sample, and adjacent to the observation that incompatible delegation assignments in agentic systems can be indistinguishable under standard execution traces (Mishra and Sharad, 2026). Real authority additionally requires feasible correction, so the relationship is directional: real authority over a right requires both mandate separating evidence and feasible corrective capability. The result also sets the reach of the assurance system. If the mandate distinguishes only which configuration served a request and whether it was authorized, evidence for route integrity and mandate compliance suffices; if it also distinguishes policies by outcome or by relative utility, the same evidence is inadequate even when perfectly accurate.

Exact equivalence is a limiting case. In practice, policies induce evidence distributions that are similar rather than equal, and sufficiency becomes a question of statistical power, sample size, and audit cost, which the gateway auditing literature now quantifies. Detecting fractional substitution, in which a gateway diverts a share ϵ of requests to a cheaper backend, costs on the order of $1/\epsilon$ queries where the substitute leaves a separating tail in the output distribution and on the order of ϵ^{-2} where it does not, and several frontier to frontier dilutions remain undetectable at five to ten percent even after hundreds of queries (Zhang, Zhang, and Qin, 2026). Operational sufficiency therefore requires precision and uncertainty appropriate to the consequence and to the claim being certified, not only the structural condition.

6.3 Four levels of assurance

The levels are claims the enterprise may need to establish, not control modes; the mechanisms supporting them can be recombined without reallocating any right.

V1, Route integrity: did the claimed inference configuration actually serve the request?

Relevant evidence includes route receipts, identifiers, fallback records, cryptographic attestation, and black box tests. The route receipt has been proposed as a compact per answer record from which a relying party can reconstruct material routing decisions without exposing proprietary internals (Schmalbach, 2026); this paper specifies the authority structure such an artifact serves and the conditions under which it is evidence rather than disclosure. Statistical tests establish identity only under restricted threat models, and each is channel specific: behavioral equality tests need a reference implementation of the claimed model and so apply to open weight candidates (Zhu et al., 2026), text only audits can be evaded by adaptive routers that recognize probes (Zhang, Zhang, and Qin, 2026), and agentic serving stacks that emit structured actions rather than text weaken the text channel altogether (Wang et al., 2026). Hardware attestation is the stronger route where provider support exists (Cai, Shi, Zhao, and Song, 2025), provided the evidence binds the request to the measured serving configuration rather than attesting that an approved model exists somewhere in a fleet. The identity to be established is itself mandate relative: a provider and model identifier suffices where the mandate distinguishes nothing finer and is insufficient where the enterprise retains rights over version, tier, compute level, or fallback. Because the available instruments depend on the eligible set, Eligibility partly determines what can later be verified, and where no credible V1 mechanism exists the mandate should say so. V1 does not establish that the route was permitted: a truthful record of an unauthorized model is still a truthful record.

V2, Mandate compliance: was the route permitted under the authorized mandate? V2 compares the observed route and policy state with the authority defined in the mandate, using a signed and versioned routing policy, the approved pool, the objective and constraint configuration, and tamper resistant records linking a request to the policy version in force when Allocation occurred. A control plane record of an authorized version does not establish request level enforcement, because staged rollout, stale caches, and recovery replicas can leave serving processes enforcing a previous revision with no misconduct involved; V2 evidence should establish which policy the serving process actually enforced, which gives Change a concrete dependency on rollout observability. V2 detects a route outside Eligibility, an unauthorized fallback, or an unapproved change. It does not establish that the authorized policy produced an acceptable outcome, and a router can comply perfectly with an incomplete mandate.

V3, Outcome assurance: did routed outputs continue to meet the enterprise's required threshold? Mechanisms include enterprise evaluation sets, shadow execution, challenger models, drift monitoring, matched loss comparison, downstream attribution, and human review where consequence requires it. The outcome should follow the authorized Objective rather than default to benchmark accuracy, and Section 4 shows why an average can miss loss concentrated in economically important states. V3 changes the independence requirement: if the provider that benefits from a route supplies the evaluation set, judge, weights, and reported score, no independent outcome assurance has been created. It also carries an attribution boundary. A routed outcome is jointly produced by the router, the provider, the serving infrastructure, the evaluator, and the enterprise workflow, so a miss does not identify which actor caused it, and the remedy may be a different router, provider, eligible set, objective, evaluation, or workflow. V3 can pass while V4 fails.

V4, Relative alignment: would another route that was both authorized and operationally feasible under the same decision state have materially improved enterprise utility? V4 asks whether the delegated allocator remained aligned relative to relevant alternatives, using challenger models, counterfactual sampling, and shadow comparison, with utility measured net of full economic cost where observable, and where affiliation is material, whether routing systematically favors affiliated configurations after accounting for the authorized objective. It is not a hindsight optimality test: outputs are stochastic and decision time information is incomplete, so the question is whether the policy systematically sacrifices enterprise utility against the alternatives available at decision time, not whether it won every request. Comparison also filters shocks common to the eligible alternatives, the informational rationale for relative performance evaluation (Holmström, 1982), which is what lets V4 evidence inform the attribution question V3 leaves open. V4 is where procedural compliance and economic alignment separate most clearly: a router can truthfully identify the serving model, stay inside the authorized pool, and keep outputs above a floor while systematically sending traffic to an option more profitable for the intermediary and less valuable to the enterprise, and V1 through V3 do not distinguish that policy from an aligned one. V4 is required only where relative alignment is part of the retained authority.

6.4 Designing the evidence

The stack states what must be established, not how to obtain it against an intermediary that may detect testing or shape the evidence. Four requirements follow from the sections above; the mechanics belong to the technical literature they cite.

Test the hard case. Partial and request conditional substitution, rather than wholesale replacement, is what an assurance program should be built to detect, and the gateway auditing literature formalizes it as a mixture whose routing fraction may depend on the request and on interaction history (Zhu et al., 2026; Zhang, Zhang, and Qin, 2026). Where an intermediary can condition behavior on detection of an audit, the audit method should be assumed known and unpredictability placed in the sampled challenge rather than in obscurity of method.[2]

Condition and adjust. Comparative evidence should be conditioned on economically relevant request classes rather than pooled, because a policy can look strong in aggregate while performing poorly where consequence concentrates, and it must be adjusted for the fact that the router selected the traffic: unadjusted averages by route confound execution quality with the allocation policy that assigned requests, so matched, shadow, or randomized comparison is required to separate the two. Under a fixed budget, coverage belongs where consequence weighted exposure, specification uncertainty, or weak observability is greatest, with some randomized coverage elsewhere.

Keep the evaluator independent in provenance as well as operation. V3 and V4 evidence can fail through validity, reliability, holdout integrity, or governance independence, and these are not substitutes: an enterprise owned evaluation set has no affiliation problem and can still be overfit by repeated router tuning, while a third party evaluation can be statistically untouched and economically tied to a candidate. Governance independence is where affiliation enters a second time. A router that samples production requests to a candidate model, compares cost and latency itself, and hosts the customer's quality evaluations is supplying part of the V3 and V4 evidence about its own Allocation and controlling the sampling that produces the rest, and that shape

[2]Where compensation, renewal, or continuation depends discontinuously on measured thresholds, the measurement system itself creates incentives to distort behavior around the threshold, so a metric's incentive effects should be tested alongside its reliability.

is now a mainstream product pattern (Ramp, 2026a; Maxim AI, 2026). Adding a monitoring layer adds a strategic actor whose incentives must be governed, organizational separation is not incentive independence (Tirole, 1986), and model based judges carry measured biases of their own (Shi et al., 2025).

> The enterprise evaluation system should be sufficiently independent of the routing candidates, in provenance as well as operation, that evaluation does not become another channel of affiliation bias.

Establish that assurance is affordable. Allocation is also an information acquisition policy: by determining which configurations receive production traffic, it determines which comparative evidence exists later, so persistent concentration makes the evidence base path dependent even when every route is compliant. Shadow evaluation, challenger traffic, and counterfactual sampling restore coverage, and they are not free. Detecting a consequence weighted association at usable power generally requires paying for the more expensive configuration on some share of traffic, and that share is drawn from the savings dynamic routing was adopted to capture. Whether the two are compatible for a given workload is open, and an enterprise should establish that relative assurance is affordable before relying on a mandate that requires it.

6.5 Mandate adequacy review and the institutional analogues

V1 through V4 evaluate delegated exercise against the mandate. Whether the mandate itself remains appropriate is a separate periodic question that revisits the eligible set, objective, quality floor, evidence envelope, and downstream outcomes, and asks whether the evidence that originally supported D_w still does. Three further objects follow from the framework's dynamics: architecture and governance fit, since evidence surfaces and component ownership can change while the written mandate stays the same; the evaluation basis, because a benchmark can saturate or lose value through tuning; and intermediary capability, which grows as a router accumulates workload data and can justify wider delegation but never expands authority on its own, being evidence for exercising Change rather than a substitute for it. Review should also refresh $B(x)$, since after enough model turnover no static policy is current and an unrefreshed baseline makes the selection term notional, and where feasible it should test the eligible set against credible alternatives outside the current mandate, because V2 can show perfect compliance with an obsolete policy and V4 relative alignment within a set whose membership should be reconsidered. Mandate review exercises Change; it is not a fifth level.

Two institutional analogues clarify the design without claiming legal equivalence. Best execution supplies the request level discipline: order routing rules require reasonable diligence and, where execution is not reviewed order by order, regular and rigorous review of routing against competing markets, and the obligation cannot be transferred to another person (FINRA, 2026). Execution quality there is already multidimensional (Harris, 2002), review is statistical while the protected object remains the individual order, and an arrangement is not allowed to validate itself against only the venues it already uses. Payment for order flow, internalization, and preferencing map onto provider rebates, affiliated routing, and favored provider relationships, and in each case the question is whether the inducement moved the route away from the objective. The analogy stops where routed AI lacks a consolidated benchmark that makes heterogeneous outcomes comparable. The delegated investment mandate supplies the relationship level logic of bounded authority, monitoring, revision, and termination (He and Xiong, 2013; ICGN and GISD, 2022).

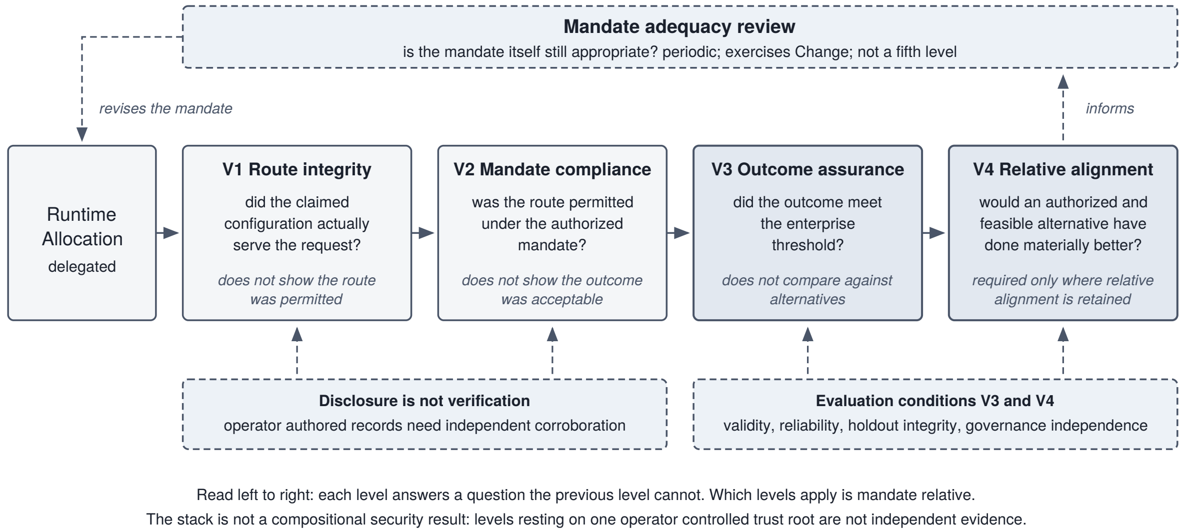


Figure 2. Assurance for delegated model choice. Read left to right, each level answers a question the previous level cannot: V1 establishes Route integrity, V2 Mandate compliance, V3 whether outcomes meet the enterprise threshold, and V4 Relative alignment against alternatives that were authorized and feasible, where the mandate requires that distinction. Operator authored records need independent corroboration to support V1 and V2; evaluation validity, reliability, holdout integrity, and governance independence condition V3 and V4. Periodic mandate adequacy review sits above the four levels and exercises Change.

The stack is not a compositional security result. A route log, a signed receipt, a policy version record, and a dashboard can be four artifacts resting on one operator controlled system, and V3 and V4 can depend on one evaluator; defenses an adaptive adversary can address together do not yield the strength of their sum (He et al., 2017). The enterprise should examine the joint evidence dependencies of its assurance design, which is the assurance analogue of the portfolio argument in Section 3.

7. Sourcing the Router Without Surrendering the Decision

> You do not need to own the router to own the decision. But owning the decision on paper is not the same as controlling it in practice.

Sections 5 and 6 supplied the conditions behind both halves of that statement. This section examines the interaction between the sourcing boundary and the authority boundary directly, and asks which assets are worth retaining when the mechanism is sourced externally.

7.1 Sourcing and authority as separate variables

Table 5. Sourcing boundary by model choice authority boundary

Routing mechanism	Enterprise retains substantial model choice authority	Intermediary holds substantial model choice authority
Internally sourced or self hosted	Internal mechanism with internal authority	Self hosted mechanism with externally shaped effective authority
Externally sourced	Bought mechanism with retained strategic authority	Broad outsourced model choice authority

The diagonal cases are familiar; the off diagonal cases are the paper's subject. In the lower left, the enterprise buys the mechanism and retains authority: the intermediary performs Allocation, maintains integrations, and supplies live market information while the enterprise controls the eligible set, Objective, Change, evidence requirements, Intervention, and Exit, so commercial routing functions as delegated implementation within an enterprise owned mandate. In the upper right, the enterprise operates the mechanism and relies on external inputs that shape the route: a vendor ranking, a managed evaluation set, a third party classifier, or an external telemetry layer that controls which outcomes are visible. Control of code and control of decision logic are separable, and make is increasingly a hybrid category. The rows are idealized endpoints, since one workload can combine enterprise owned policy with an externally operated gateway and externally served models, and the columns still ask which party controls each right across that hybrid. A gateway hosted as a foundation open source project rather than a vendor product (Linux Foundation, 2025) has the authority properties of any self hosted mechanism and one distinct feature, Exit, because it cannot be acquired out from under its adopters.

In either off diagonal case the practical value of the arrangement depends on more than contract language. A contractual right to change the eligible pool is weak if the enterprise cannot compare alternatives, and a right to terminate is weak if the evaluation assets and telemetry needed to switch are nonportable. Buying preserves strategic authority only when authority is both allocated and operationally supported.

7.2 What is worth retaining

Some routing capabilities are generic and reusable across firms: API normalization, retries, basic failover, connector maintenance, availability monitoring, and commodity cost telemetry. Sourcing them is efficient where external providers have scale or specialization advantages. Other assets encode enterprise specific knowledge or future advantage: workload taxonomies, proprietary evaluation data, error and failure economics, business outcome telemetry, policy thresholds, reviewer feedback, domain specific labels, and the ability to intervene and change policy. Resource based and dynamic capabilities logic separates the two by asking which assets determine how the enterprise defines utility, distinguishes acceptable from unacceptable routes, and learns from production, and which capabilities let it sense change and reconfigure as the model market moves (Barney, 1991; Teece, Pisano, and Shuen, 1997). The principle is not that every enterprise must build these assets internally. It is that strategically specific assets should remain sufficiently controlled, portable, and observable that sourcing the mechanism does not surrender the capability they support, which shifts the sourcing question from which router to own toward which knowledge, evidence, and policy assets must remain under enterprise control for the enterprise to adapt.

7.3 Portability, specific investment, and a self reinforcing feedback

Portability links Evidence and Exit. Evaluation sets, route histories, policy mappings, outcome labels, and telemetry integrations become relationship specific when they depend on one intermediary's formats, identifiers, or model pool, and as they do switching becomes more costly. Hold up logic makes the governance concern precise: when a party must make a specific, partly nonrecoverable investment before later renegotiation or switching, anticipated loss of bargaining position reduces the incentive to invest. Information technology outsourcing research links asset specificity to vendor lock in and shows how asset transfer creates one sided switching costs (Shawosh and Berente, 2019; Shivendu, Zeng, and Gurbaxani, 2020), and recent work models residual control over delegated AI execution as affecting relationship specific investment under incomplete contracts (Zhang and Xu, 2026). The investment being discouraged here is exactly the measurement capability P3 requires: an enterprise will hesitate to build rich evaluation and outcome attribution around one intermediary if those assets cannot be exported.

The channel runs in both directions, as the property rights tradition insists. Stronger enterprise control protects enterprise specific investment in evaluation and telemetry while reducing the intermediary's return on enterprise specific adaptation and tuning; broader delegation reverses the tradeoff. One direction is self reinforcing and deserves naming. Delegating Allocation gives the intermediary traffic, learning, and telemetry. Its Capability rises while the enterprise's independent Verifiability falls, because the comparative evidence that would ground a challenge is increasingly generated by the party that would be challenged, and applying the five variables again in the next period, both movements argue for wider delegation. The normative guard is that capability growth is evidence for exercising Change rather than authority that expands on its own, but the guard does not defuse the mechanism. What defuses it is preserving an independent evidentiary position: portable evaluation assets, retained route and outcome history, and comparison against alternatives outside the current mandate.

Three observations condition the recommendation. Portable route records, standardized identifiers, exportable evaluation sets, and outcome telemetry that survives a vendor change reduce specificity, and cost telemetry standardization now has an institutional home in a vendor neutral foundation formed in August 2026 to define measurement standards for AI economics (Linux Foundation, 2026); evaluation portability is more demanding, because reproducing a comparison elsewhere requires the protocol and provenance behind it. The recommendation assumes the intermediary is severable: where routing is a feature of a cloud or data platform, exit is a platform migration dominated by data gravity, identity, and committed spend, and portability effort belongs on the assets that can move independently, principally the evaluation protocol, the policy definitions, and the outcome telemetry. And a negotiated allocation is only as durable as its counterparty. In May 2026 an independent routing intermediary was acquired by a security platform vendor (Palo Alto Networks, 2026), and on August 19, 2026 the largest independent intermediary announced an agreement to be acquired by a payments company; the price, reported as above seven billion dollars, is attributed to people familiar with the transaction and was not disclosed by either party (Bloomberg, 2026; Stripe, 2026; OpenRouter, 2026c). The announcement commits that routing decisions will remain driven by the user's interest and will not bend to any parent company, which is a bonding statement in the sense of Section 6.1 whose credibility is exactly what V4 evidence exists to test. In a consolidating market the horizon that matters is the counterparty's independence, which can be shorter than the contract term.

> Portability is not only an exit convenience. It can preserve the value of investing in the evidence needed for real model choice authority.

7.4 Learning and Data

Prompts, outputs, route decisions, labels, reviewer feedback, retries, and downstream outcomes can improve future routing, an affiliated model, or an intermediary's market wide capability through learning across customers. A dual role platform model formalizes the mechanism: traffic allocated to an outside supplier affects both current demand and future quality through learning by serving, and platform self preference alters those flows (Zhang, Bao, and Li, 2026). The mechanism rests on this single formal treatment within the reviewed record and is corroborated qualitatively from the market side: investors in the largest independent router describe its moat as the accumulated dataset of prompts, model assignments, contexts, and results, and describe the future of the product as the router running the comparative evaluation and recommending the switch (Menlo Ventures, 2026). That description places the intermediary in control of both Allocation and the V4 evidence about it.

> Learning and Data rights. The greater the competitive or capability value of routing traffic, outcome telemetry, and evaluation labels, the stronger the case for retaining those rights or explicitly governing and valuing their transfer in the mandate.

Data need not stay internal and value need not be a cash price; sharing can create value through reciprocal access, service commitments, or restrictions on downstream use. The point is that the transfer is an economic choice rather than an incidental term inside a generic data policy, that it interacts with Alignment because an intermediary that learns from traffic sent to an affiliated model faces a different incentive from a neutral one, and that it interacts with Exit because accumulated learning deepens dependency even when raw data remain exportable. A routing vendor that exercises substantive discretion should accordingly be governed less like a conventional software supplier and more like a delegated decision intermediary operating under a bounded mandate and an independently verifiable standard of alignment, with the degree of control contingent on the five determinants rather than presumed.

8. Discussion

8.1 Five management questions

The framework changes the managerial question from which router to buy to which model choice rights a workload should delegate, under what mandate, and with what evidence. For each workload, management should be able to answer five questions.

1. **What existing constraints must routing preserve?** Discretion begins inside the feasible set left by data classification, residency, prohibited providers, quality floors, required review, and policy floor rules, and should not silently relax them.

2. **Which rights does the enterprise retain, and does the mandate resolve what those rights decide?** For Objective in particular, a written statement of retention is not evidence of retention. Ask whether the mandate resolves the tradeoffs it claims to keep, and who supplied the weights in the evaluation that ranks candidates.
3. **What discretion is the intermediary actually being granted?** State it in terms of workload, eligible configurations, authorized objective, evidence envelope, excluded decisions, fallback, and escalation.
4. **How will the enterprise know when the mandate or its evidence base is no longer adequate?** Drift, new model versions, changed affiliation, unexplained deterioration, weaker comparative performance, and movement outside the evaluated region each need a detection path.
5. **Can the enterprise independently verify, intervene, change, and exit?** Where any of the four exists only in the contract, the corresponding right is formal rather than real.

Each retained or delegated right then needs an operating mechanism, a named authority, a required input, an evidence path, and an exception route. Table 6 illustrates the pattern without prescribing any particular allocation, and two implementation principles accompany it. Existing mechanisms can carry rights, so the framework does not require a new governance apparatus where procurement, vendor management, security, change control, and platform operations already exist. And runtime Intervention, which needs a technically executable mechanism, differs from governance exception handling, which decides on unanticipated states that fall outside the evidence supporting ordinary Allocation.

Two tests show whether the arrangement operates rather than merely exists: the people expected to run it can say who may add or remove a configuration, who sets the objective, what the intermediary may change without approval, what triggers Intervention, and who can terminate; and the governance system is measured separately from the router, through Change and Intervention cycle time, evidence completeness, exception aging, and portability test results, kept as operating measures rather than a score. The purpose of a mandate is not maximum retained control. It is to make the ordinary path fast because authority has already been settled, while exceptions and reversion remain explicit. Nor is the purpose of the framework to produce dynamic routing. Static assignment, or continuation of the existing model choice architecture, remains a legitimate outcome where the incremental value of delegation does not exceed its economic, control, and assurance costs, and the static counterfactual in Section 4 exists to make that comparison rather than to assume its answer.

Table 6. Illustrative implementation matrix for model choice rights

Right	Typical mechanism	Authority holder	Required input	Evidence path	Exception route
Eligibility	Approved configuration registry, signed allowlist, gateway policy	Enterprise, with intermediary proposals	Provider qualification, workload constraints, evaluation evidence	Registry version history, route records showing registry compliance	Emergency removal or suspension procedure
Objective	Mandate document, policy as code, evaluation specification	Enterprise, with intermediary input	Workload utility, constraints, tradeoff priorities, consequence weights	Evaluation protocol version, objective configuration, drift monitoring	Objective review triggered by outcome deterioration
Change	Change approval process, version control, staged rollout	Enterprise approval or joint approval; bounded delegation within approved classes	Proposed change, regression evidence against authority boundaries and escalation conditions	Policy version history, rollout observability, approval records	Rollback and freeze authority
Allocation	Router, classifier, cascade, or deterministic policy	Intermediary within mandate	Request context, eligible set, objective, live provider conditions	Route records linking request, configuration, and policy version	Fallback, escalation, or suspension on trigger
Intervention	Kill switch, circuit breaker, static fallback, review queue	Enterprise, or enterprise defined automated regime	Trigger conditions, monitoring thresholds, tested fallback capacity	Intervention logs, recovery records, fallback tests under load	Escalation to workload owner and accountable executive
Evidence	Route receipts, attestation, telemetry export, evaluation records	Intermediary generates, enterprise controls access and retention	Evidence schema, provenance requirements, independence requirements	Tamper resistant records, independent verification results	Evidence gap escalation and audit rights
Learning and Data	Data processing terms, retention configuration, use restrictions	Enterprise, with negotiated transfer	Data classification, competitive sensitivity, reciprocal value	Retention and reuse records, audit of stated use	Revocation and deletion rights
Exit and Portability	Portable policy, evaluation sets, route history, standardized identifiers	Enterprise	Portability requirements at contracting, periodic portability tests	Portability test results, export records	Termination and migration plan

The matrix is illustrative; allocations follow from the five determinants applied per workload and right.

> If the router makes a decision the enterprise remains accountable for, the enterprise should be able to identify the mandate, observe the decision, test the outcome, change the rules, and reclaim the decision when necessary.

Procurement is the implementation of that prior governance decision, and it changes vendor comparison: a technically strong router with weak portability, opaque policy changes, or limited independent evidence can be less attractive for a workload whose retained rights depend on those capabilities than a less sophisticated mechanism that preserves them.

8.2 Status of the claims

The framework, propositions, and analytical results are conceptual. The paper does not claim to invent decision rights, delegation, algorithm selection, cost sensitive allocation, or verification, nor that separating sourcing from authority is new in information systems governance; the mathematics is elementary and serves to make governance objects precise. P1, P2, and P3 are directional propositions specific to this framework and are the ones later work should translate into prespecified hypotheses; the knowledge locality, incentive alignment, and Learning and Data statements apply established results to a new object. The five determinants carry no weights. Public product and market records establish that the relevant control arrangements exist and where several rights sit by default; they do not show how enterprises allocate authority, how often a control is used, or whether it is effective. No independent outcome study of enterprise model routing was located, and available case studies are vendor narratives; control layer integrity, by contrast, is now field measured at the gateway and reseller level, and that distinction should be kept. The paper also does not establish which recommended allocations are commercially available. Analytical separability is not commercial separability, and an enterprise may find that the arrangement it should want, such as holding provider contracts directly while delegating the mechanism, is one no vendor whose margin depends on routed tokens offers. Claims of differentiation are bounded by the search actually conducted.

8.3 Boundary conditions

The framework applies to runtime allocation among eligible inference configurations. A fixed cascade whose transitions are predetermined by the mandate is part of one configuration; where the intermediary chooses adaptively among stages, those transitions are Allocation. Physical scheduling such as replica selection or batching lies outside the boundary unless it alters an attribute the mandate governs. The paper does not govern tool selection, agent orchestration, or autonomous action, and the reason is stronger than scope: the Section 4 result treats the consequence weight as a function of the request state, and in an agentic trajectory it is not, because what an output costs depends on what later steps do with it. That breaks a premise of the result rather than narrowing its range, and it makes trajectory level mandates the successor object; practitioners now make the same observation from the deployment side (Menlo Ventures, 2026). The framework confines itself to governance that can be made explicit, since relational mechanisms such as trust and repeated interaction can complement it without removing the need for mandate separating evidence where independent assurance is required. It does not assume that routing is opaque, adversarial, or misaligned; those enter as contingent variables. And it rejects a universal best model premise: the counterfactual is the enterprise's approved static policy, not an imagined policy that always selects the most capable configuration.

8.4 Research agenda

Three tasks follow, in rough order of cost. First, commercial feasibility: for named routing products, determine from public terms which of the rights treated here as separately allocable are separately purchasable, including direct provider contracting, explicit allow listing, exact version pinning, objective transparency, and portable route history. This establishes the frontier against which every recommendation in the paper should be read, since analytical separability is not commercial separability. Second, real versus formal authority: map who actually controls each right in deployed enterprises, operationalize the gap with right specific measures of observability, comparative evaluation capability, switching capability, evaluation independence, and time to Intervention, and test whether formal retention predicts control only when backed by separating evidence and feasible correction, which is P3; the same fieldwork can test the heterogeneity prediction and distinguish the existence of an intermediary conflict from evidence that it moves Allocation against enterprise utility. Third, assurance economics: for a given savings rate, effect size, and traffic volume there should exist a minimum shadow rate below which a consequence weighted association cannot be detected, and it is not obvious that this rate is always smaller than the saving. Where it is not, relative assurance is unaffordable precisely for the workloads whose thin margins make routing attractive, and the economically rational well governed policy for those workloads is not to route dynamically at all. Cross sectional tests of all three should be paired with process evidence that traces how incidents, accumulated evidence, and capability growth actually move rights and mandate scope over time (Sabherwal and Robey, 1995).

9. Conclusion

Model routing adds an optimization component to enterprise AI architecture while changing the frequency, locus, and potentially the ownership of model choice authority. A decision once bundled into periodic architecture review can now be exercised repeatedly at runtime by an intermediary whose information, incentives, and evidence differ from the enterprise's. Three conclusions follow. Sourcing and authority are distinct: who operates the mechanism does not determine who owns the decision, in either direction. The boundary is workload specific: the primary unit for specifying delegated discretion is the workload routing mandate, resting on an enterprise policy floor and subject to a portfolio review of shared exposure. And delegation requires real authority and assurance: formal retention of a right matters only when the enterprise has evidence that separates the states the right cares about and a feasible way to act on them, which route disclosure alone cannot supply. The framework does not prescribe internalization or presume that routing raises risk; its narrower claim is that routing changes the governance object because repeated model choice is now exercised inside an explicit or implicit mandate. One consequence is immediately checkable. A right is retained only if the mandate resolves what that right decides, so an enterprise that names several goals without ordering them, or lets a multi metric evaluation be collapsed into a ranking by someone else, has delegated Objective whatever its agreement says.

> Model choice can be delegated without being surrendered, but only when the enterprise can define the mandate, observe its exercise, and reclaim authority when the evidence no longer supports discretion.

References

Aghion, Philippe, and Jean Tirole. 1997. “Formal and Real Authority in Organizations.” *Journal of Political Economy* 105(1): 1-29.

Amazon Web Services. 2026. “Understanding Intelligent Prompt Routing in Amazon Bedrock.” Amazon Bedrock documentation. Accessed August 30, 2026.

Angelopoulos, Anastasios N., Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. 2024. “Conformal Risk Control.” *The Twelfth International Conference on Learning Representations* (ICLR 2024). arXiv:2208.02814.

Arora, Chetan, Andreas Vogelsang, and Abbi Sharma. 2026. “Specifying the Delegated-Autonomy Boundary: Requirements Engineering for Agentic AI.” *Proceedings of the 41st IEEE/ACM International Conference on Automated Software Engineering (ASE 2026), New Ideas and Emerging Results Track*, Munich, Germany, October 12-16, 2026. arXiv:2607.17225.

Barney, Jay B. 1991. “Firm Resources and Sustained Competitive Advantage.” *Journal of Management* 17(1): 99-120.

Bloomberg. 2026. Sun, Yazhou, Natasha Mascarenhas, and Paige Smith. “Stripe Finalizes Deal to Acquire AI Startup OpenRouter for Over $7 Billion.” Bloomberg News, August 16, 2026.

Bommasani, Rishi, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. 2022. “Picking on the Same Person: Does Algorithmic Monoculture Lead to Outcome Homogenization?” *Advances in Neural Information Processing Systems* 35: 3663-3678.

Cai, Will, Tianneng Shi, Xuandong Zhao, and Dawn Song. 2025. “Are You Getting What You Pay For? Auditing Model Substitution in LLM APIs.” arXiv:2504.04715.

Castiñeiras Rella, Jorge, Gerda Claeskens, Ricardo Cao, and Juan M. Vilar. 2025. “Instance-Dependent Cost-Sensitive Parametric Learning.” *Neurocomputing* 615: 128875.

Choudhury, Vivek, and Rajiv Sabherwal. 2003. “Portfolios of Control in Outsourced Software Development Projects.” *Information Systems Research* 14(3): 291-314.

Chow, C. K. 1970. “On Optimum Recognition Error and Reject Tradeoff.” *IEEE Transactions on Information Theory* 16(1): 41-46.

Erol, Mehmet Hamza, Batu El, Mirac Suzgun, Mert Yüksekgönül, and James Zou. 2025. “Cost-of-Pass: An Economic Framework for Evaluating Language Models.” arXiv:2504.13359.

Fama, Eugene F., and Michael C. Jensen. 1983. “Separation of Ownership and Control.” *Journal of Law and Economics* 26(2): 301-325.

FINRA. 2026. “Rule 5310. Best Execution and Interpositioning.” Financial Industry Regulatory Authority.

Futurum Group. 2026. “Futurum Agent Control Plane Framework: A Reference Model for Production AI Agents.” The Futurum Group, April 2026.

Grossman, Sanford J., and Oliver D. Hart. 1986. “The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration.” *Journal of Political Economy* 94(4): 691-719.

Harris, Lawrence E. 2002. *Trading and Exchanges: Market Microstructure for Practitioners.* New York: Oxford University Press.

Hart, Oliver, and John Moore. 1990. “Property Rights and the Nature of the Firm.” *Journal of Political Economy* 98(6): 1119-1158.

Hart, Oliver, Andrei Shleifer, and Robert W. Vishny. 1997. “The Proper Scope of Government: Theory and an Application to Prisons.” *Quarterly Journal of Economics* 112(4): 1127-1161.

He, Warren, James Wei, Xinyun Chen, Nicholas Carlini, and Dawn Song. 2017. “Adversarial Example Defenses: Ensembles of Weak Defenses are not Strong.” *11th USENIX Workshop on Offensive Technologies* (WOOT 17).

He, Zhiguo, and Wei Xiong. 2013. “Delegated Asset Management, Investment Mandates, and Capital Immobility.” *Journal of Financial Economics* 107(2): 239-258.

Holmström, Bengt. 1979. “Moral Hazard and Observability.” *The Bell Journal of Economics* 10(1): 74-91.

Holmström, Bengt. 1982. “Moral Hazard in Teams.” *The Bell Journal of Economics* 13(2): 324-340.

Holmström, Bengt, and Paul Milgrom. 1991. “Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design.” *Journal of Law, Economics, and Organization* 7(Special Issue): 24-52.

Holmström, Bengt, and Paul Milgrom. 1994. “The Firm as an Incentive System.” *American Economic Review* 84(4): 972-991.

Hu, Qitian Jason, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. 2024. “RouterBench: A Benchmark for Multi-LLM Routing Systems.” arXiv:2403.12031.

Hydari, Muhammad Zia, and Farooq Muzaffar. 2026. “Redrawing the AI Map: A Theory of Accountability Boundaries in Agentic Ecosystems.” arXiv:2605.23179.

ICGN and GISD Alliance. 2022. *Model Mandate*. International Corporate Governance Network and Global Investors for Sustainable Development Alliance.

Jensen, Michael C., and William H. Meckling. 1976. “Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure.” *Journal of Financial Economics* 3(4): 305-360.

Jensen, Michael C., and William H. Meckling. 1992. “Specific and General Knowledge, and Organizational Structure.” In *Contract Economics*, edited by Lars Werin and Hans Wijkander, 251-274. Oxford: Blackwell.

Kearns, Grover S., and Rajiv Sabherwal. 2006. “Strategic Alignment Between Business and Information Technology: A Knowledge-Based View of Behaviors, Outcome, and Consequences.”

Journal of Management Information Systems 23(3): 129-162.

Kirsch, Laurie J. 1997. “Portfolios of Control Modes and IS Project Management.” *Information Systems Research* 8(3): 215-239.

Li, Yang. 2025. “Rethinking Predictive LLM Routing: When Simple KNN Beats Complex Learned Routers.” arXiv:2505.12601, revised May 2026.

Li, Yifei, Zihui Gao, and Laks V. S. Lakshmanan. 2026. “WISERouter: LLM Routing with Workload Budget Constraint.” arXiv:2607.23765.

Liang, Percy, Rishi Bommasani, Tony Lee, et al. 2023. “Holistic Evaluation of Language Models.” *Transactions on Machine Learning Research*.

Lin, Guanjie, Yinxin Wan, Shichao Pei, Ting Xu, Kuai Xu, and Guoliang Xue. 2026. “Behavioral Consistency and Transparency Analysis on Large Language Model API Gateways.” In *Proceedings of the 2026 ACM Internet Measurement Conference* (IMC ’26), Karlsruhe, Germany, October 12-16, 2026. arXiv:2604.21083.

Linux Foundation. 2025. “Linux Foundation Announces the Formation of the Agentic AI Foundation (AAIF).” Press release, December 9, 2025.

Linux Foundation. 2026. “Linux Foundation Launches the Tokenomics Foundation to Define the Economics and ROI of AI Value.” Press release, August 4, 2026.

Liu, Hanzhi, Chaofan Shou, Hongbo Wen, Yanju Chen, Ryan Jingyang Fang, and Yu Feng. 2026. “Your Agent Is Mine: Measuring Malicious Intermediary Attacks on the LLM Supply Chain.” arXiv:2604.08407.

Longpre, Shayne, Kevin Klyman, Ruth Elisabeth Appel, et al. 2025. “Position: In-House Evaluation Is Not Enough. Towards Robust Third-Party Evaluation and Flaw Disclosure for General-Purpose AI.” *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267.

Madras, David, Toniann Pitassi, and Richard Zemel. 2018. “Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer.” *Advances in Neural Information Processing Systems* 31: 6150-6160.

Mahmood, Rafid. 2026. “Routing, Cascades, and User Choice for LLMs.” In *The Fourteenth International Conference on Learning Representations* (ICLR 2026), 52692-52714. arXiv:2602.09902.

Maxim AI. 2026. “Bifrost: Open Source LLM Gateway.” Product documentation and repository, maximhq/bifrost. Accessed September 4, 2026.

Menlo Ventures. 2026. Murphy, Matt, and Deedy Das. “Stripe to Acquire OpenRouter: Why Everyone Is Obsessed With Model Routing.” Menlo Ventures Perspective, August 19, 2026.

Microsoft. 2026. “How to Use Model Router for Microsoft Foundry.” Microsoft Learn. Accessed August 30, 2026.

Mishra, Abhinav, and Kumar Sharad. 2026. “Observability for Delegated Execution in Agentic AI Systems.” arXiv:2606.09692.

Mozannar, Hussein, and David Sontag. 2020. “Consistent Estimators for Learning to Defer to an Expert.” In *Proceedings of the 37th International Conference on Machine Learning*.

Nowak, Marcin. 2026. "A Multi-Criteria Decision Framework for Enterprise LLM Routing." *Information* 17(6): 539.

Ong, Isaac, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. 2025. "RouteLLM: Learning to Route LLMs from Preference Data." In *The Thirteenth International Conference on Learning Representations* (ICLR 2025). arXiv:2406.18665.

OpenRouter. 2026a. "Auto Router: Intelligent Model Selection." Product documentation, openrouter.ai/docs/guides/routing/routers/auto-router; and "Model Routing Powered by Wisdom of the Market," announcement, August 10, 2026. Accessed September 12, 2026.

OpenRouter. 2026b. "Presets," "Guardrails" (May 29, 2026), and "In-Region Routing: Keep Your Data in the US or EU" (September 9, 2026). Product documentation and announcements; platform fee terms per openrouter.ai/pricing and the OpenRouter help center article on bring your own key charges. Accessed September 12, 2026.

OpenRouter. 2026c. "OpenRouter is Joining Stripe." Announcement, August 19, 2026. openrouter.ai/blog/announcements/openrouter-is-joining-stripe. Accessed September 12, 2026.

Ouchi, William G. 1979. "A Conceptual Framework for the Design of Organizational Control Mechanisms." *Management Science* 25(9): 833-848.

Palo Alto Networks. 2026. "Palo Alto Networks Completes Acquisition of Portkey." Press release, May 29, 2026.

Ramp. 2026a. Patel, Veeral. "Introducing Router.com: The Best Model Changes. Your Application Shouldn't." Ramp blog, August 19, 2026, ramp.com/blog/router-launch; and "Ramp Launches Router.com to Cut Companies' Rising AI Bills." Press release, August 19, 2026. Accessed September 12, 2026.

Ramp. 2026b. "Router." Product documentation and data retention terms, router.com. Accessed September 4, 2026.

Rice, John R. 1976. "The Algorithm Selection Problem." *Advances in Computers* 15: 65-118.

Rochet, Jean-Charles, and Jean Tirole. 2003. "Platform Competition in Two-Sided Markets." *Journal of the European Economic Association* 1(4): 990-1029.

Sabherwal, Rajiv, and Daniel Robey. 1995. "Reconciling Variance and Process Strategies for Studying Information System Development." *Information Systems Research* 6(4): 303-327.

Sawers, Paul. 2026. "Cursor, Ramp, and Meta Are All Building Model Routers, But Two Have Major Model Ambitions Themselves." The New Stack, July 23, 2026. Accessed September 12, 2026.

Schmalbach, Vincent. 2026. "Model Routing as a Trust Problem: Route Receipts for Adaptive AI Systems." arXiv:2605.01710.

Schrage, Michael, and David Kiron. 2025. "Winning With Intelligent Choice Architectures." MIT Sloan Management Review and Tata Consultancy Services, July 2025.

Sebastian, Ina M., Peter Weill, Thomas Haskamp, and Jan vom Brocke. 2026. "Designing Decision Rights for AI." MIT CISR Research Briefing XXVI-6, June 18, 2026.

Shafran, Avital, Roei Schuster, Thomas Ristenpart, and Vitaly Shmatikov. 2025. "Rerouting LLM Routers." arXiv:2501.01818.

Shawosh, Mazen, and Nicholas Berente. 2019. "Software Development Outsourcing, Asset Specificity, and Vendor Lock-in." *AMCIS 2019 Proceedings*.

Shi, Lin, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. "Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge." In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 292-314.

Shivendu, Shivendu, David Zeng, and Vijay Gurbaxani. 2020. "Optimal Asset Transfer in IT Outsourcing Contracts." *MIS Quarterly* 44(2): 857-905.

Snowflake. 2026. "Snowflake Unlocks Better AI Economics with Dynamic Model Routing, Delivering More Value to Customers." Company announcement, August 18, 2026.

South, Tobin, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Dazza Greenwood, Alan Chan, and Alex Pentland. 2025. "Authenticated Delegation and Authorized AI Agents." arXiv:2501.09674.

Stripe. 2026. "Stripe Agrees to Acquire OpenRouter to Help Businesses Optimize Token Routing and Usage." Newsroom, August 19, 2026. Accessed September 13, 2026.

Teece, David J., Gary Pisano, and Amy Shuen. 1997. "Dynamic Capabilities and Strategic Management." *Strategic Management Journal* 18(7): 509-533.

Tirole, Jean. 1986. "Hierarchies and Bureaucracies: On the Role of Collusion in Organizations." *Journal of Law, Economics, and Organization* 2(2): 181-214.

Tiwana, Amrit. 2008. "Does Technological Modularity Substitute for Control? A Study of Alliance Performance in Software Outsourcing." *Strategic Management Journal* 29(7): 769-780.

Tiwana, Amrit, and Mark Keil. 2007. "Does Peripheral Knowledge Complement Control? An Empirical Test in Technology Outsourcing Alliances." *Strategic Management Journal* 28(6): 623-634.

Tiwana, Amrit, Benn Konsynski, and Ashley A. Bush. 2010. "Research Commentary: Platform Evolution: Coevolution of Platform Architecture, Governance, and Environmental Dynamics." *Information Systems Research* 21(4): 675-687.

Tiwana, Amrit, Benn Konsynski, and N. Venkatraman. 2013. "Information Technology and Organizational Governance: The IT Governance Cube." *Journal of Management Information Systems* 30(3): 7-12.

Wang, Xun, Bihe Zhao, Michael Backes, Franziska Boenisch, and Adam Dziedzic. 2026. "Agent-Prov: Auditing Agentic LLM API Providers via Tool-use Policy Probes." arXiv:2609.00052.

Wei, Peng, and Wesley Shu. 2026. "Preserving Decision Sovereignty in Military AI: A Trade-Secret-Safe Architectural Framework for Model Replaceability, Human Authority, and State Control." arXiv:2604.20867.

Weill, Peter, and Jeanne W. Ross. 2004. *IT Governance: How Top Performers Manage IT Decision Rights for Superior Results.* Boston: Harvard Business School Press.

Zhang, Xufeng, Shenghui Bao, and Han Li. 2026. "Expert Routing and Self-Preferencing on AI Platforms." *Applied Operations and Analytics* 2(1): 1-21. Published online August 11, 2026.

Zhang, Yage, Yukun Jiang, Zeyuan Chen, Michael Backes, Xinyue Shen, and Yang Zhang. 2026. "Real Money, Fake Models: Deceptive Model Claims in Shadow APIs." arXiv:2603.01919.

Zhang, Yuewei, Zhi-Hai Zhang, and Hanzhang Qin. 2026. "Which Model Is Actually Serving You? IRIS: Budgeted Black-Box Auditing of Model Substitution and Routing Dilution in LLM Gateways." arXiv:2607.20860.

Zhang, Yukun, and Kemu Xu. 2026. "Delegation Rights: Property, Agency, and Investment Incentives in the Age of AI Agents." arXiv:2606.31935.

Zhu, Xiaoyuan, Yaowen Ye, Tianyi Qiu, Hanlin Zhu, Sijun Tan, Ajraf Mannan, Jonathan Michala, Raluca Ada Popa, and Willie Neiswanger. 2026. "Auditing Black-Box LLM APIs with a Rank-Based Uniformity Test." In *The Fourteenth International Conference on Learning Representations* (ICLR 2026). arXiv:2506.06975.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author used large language model assistants, including Anthropic's Claude and OpenAI's ChatGPT, for drafting support, source verification support, critical review, and editorial iteration under the author's direction. After using these tools, the author reviewed and edited all content, independently verified the sources and the claims they support, and takes full responsibility for the content of this publication.

Comments are invited; this is an explicitly provisional working draft.

Working draft for discussion. Not investment, legal, or tax advice.