FRONTIER.

FRONTIER RESEARCH / DECISION-GRADE ANALYSIS

Rent the Models. Own the Rules: The Enterprise Guide to AI Token Routing

A 56-profile market guide to model access, intelligent selection and enterprise AI gateways. Why accepted outcomes—not cheap tokens—should govern the economics, and how to gain speed and choice without outsourcing control.

FR-031 / Dion Hinchcliffe / Frontier Research

REPORT PROFILE

Built to change a decision.

A 56-profile market guide to model access, intelligent selection and enterprise AI gateways. Why accepted outcomes—not cheap tokens—should govern the economics, and how to gain speed and choice without outsourcing control.

REPORT	**FR-031**
AUTHOR	**Dion Hinchcliffe**
PUBLISHED	**Frontier Research**
PROFILE	**frontier-ro-0.1**

This PDF is a publication rendering of the canonical Frontier Research Object. The web report remains the live source of truth; citations, figures, methodology, and rights metadata travel with this edition.

Executive brief: the router is not the strategy

Rent the models. Own the rules.

AI token routing is becoming the distribution system for enterprise intelligence. It can collapse weeks of supplier onboarding into a model-selection decision, let developers compare capabilities without scattering provider credentials across projects, and bring spending, access and execution evidence into one place. The strategic dividend is not the cheapest token. It is faster access to capabilities the organization can actually use, with a credible way to change suppliers when technology, economics or trust changes.

That dividend comes with a catch. A router is an intermediary with discretion. It may choose the model, the hosting provider, the region, the retry path, the amount of reasoning, the cache and sometimes the transformation applied to a prompt. Once that happens, the enterprise has delegated part of its procurement policy, data-access policy and application behavior to another decision system. A convenient API can become an inconvenient concentration of power.

Frontier's position is therefore deliberately asymmetric: open the exploration lane; constrain the execution lane. Give research and development teams broad, governed access to models and modalities. Promote only evaluated routes into production, with explicit eligibility rules, bounded fallbacks, verified outcomes and an independently enforceable exit. A production router should optimize inside the enterprise's authority, never redefine it.

This report extends The Token Crash Is Coming. Your AI Bill Is Going Up., which argued that routing is becoming real-time procurement and that completed outcomes—not falling token prices—are the correct economic unit. It also builds on The Autonomous Enterprise and The Future of Work in the AI Era: delegation creates value only when the organization can set boundaries, observe results and intervene. [1] [2] [3]

The six findings that matter

1. A model marketplace, an intelligent selector and a policy gateway are different products. They increasingly overlap, but buying one does not mean obtaining all three. A catalog of video models does not prove quality-aware video routing. A gateway that load-balances a named model does not necessarily know which model should perform a task.
2. General AI access is the innovation engine; enterprise routing is the operating discipline. These are buying modes, not permanent vendor classes. Broad-access platforms are adding enterprise controls while security and API-management vendors are incorporating adaptive selection. Evaluate the specific deployment, not the category label.
3. Quality-adjusted, policy-compliant completion dominates unit price. Retries, failed tools, expired media jobs, human review and rework can erase an apparent inference discount. A fast HTTP response that produces unusable work is negative productivity.
4. The research is promising, not a blank check. RouteLLM demonstrates substantial savings in particular benchmark settings. A much broader 2026 benchmark finds that several newer approaches fail to reliably beat a simple baseline. Model complementarity is real; a router's ability to exploit it must be demonstrated. [4] [5]
5. The market is consolidating around money, security and infrastructure. Stripe announced an agreement to acquire OpenRouter; Palo Alto Networks launched Prisma AIRS AI Gateway after closing its Portkey acquisition; DigitalOcean acquired Katanemo Labs and its Plano infrastructure. The strategic asset is the decision point between enterprise demand and model supply. [6] [7] [8]
6. The most important procurement test is controlled failure. When a route is unavailable, expensive, unsafe or no longer permitted, does the system honor the original constraints—or quietly broaden them to keep returning answers?

   Reading route. Sections 1–3 define the market and its differentiation; sections 4–7 establish the evidence, economics and control architecture; sections 8–10 provide evaluation, contracting and deployment guidance. The directory profiles 56 named offers/projects, with additional supply-side and research adjacencies. Inclusion is not an endorsement, a security certification or a claim that every reseller in the world is covered. Evidence cutoff: 2 November 2026.

1. What exactly is being routed?

“Token router” is useful market shorthand, but technically imprecise. Most of the products in this report route requests, sessions or workflow steps, not individual tokens within a model. They should not be confused with mixture-of-experts routing inside a neural network, OAuth access-token routing, or GPU schedulers assigning inference work among replicas.

A complete route has more coordinates than a model name:

Task -> model/version -> hosting provider -> region -> credential/tenant -> runtime parameters -> permitted fallback -> verification policy.

The choice of reasoning budget, maximum output, tool schema, cache eligibility and modality is part of the effective execution contract. Two requests carrying the same marketing model name can reach different hosting paths, retention policies or implementation settings. Conversely, two differently named gateways may ultimately depend on the same upstream capacity.

Four mechanisms deserve separate treatment. Deterministic routing uses explicit rules, priorities or weights. Predictive routing estimates which model will meet a task requirement before generating an answer. Cascading tries a model, evaluates the answer and escalates when necessary. Portfolio or ensemble execution obtains multiple answers or delegates subtasks, trading extra work for possible quality gains. A fifth, provider routing, keeps a model largely fixed while choosing where to run it. These mechanisms have different costs, failure modes and audit requirements. RouteLLM, FrugalGPT and current gateway documentation illustrate why the distinction is substantive. [4] [9] [10]

The multimodal trap

A single API key is not a single execution model. Text completion may stream; image creation returns one or more assets; video often creates an asynchronous job with cancellation, polling and storage semantics. Speech can require continuous low-latency streaming. A coding agent may retain tool results, reasoning state and cache context across dozens of calls. Media can be billed per image, output second, resolution or compute time rather than a comparable text-token unit. Replicate's prediction API, Segmind's workflow model and WaveSpeed's media API make these differences visible. [11] [12] [13]

Therefore, model accessibility and adaptive routing coverage must be recorded separately. LLMGateway, for example, documents dynamic routes for chat/responses/messages rather than every media endpoint. A supplier can be an excellent media access partner without yet offering reliable autonomous selection among video models. [14]

For an enterprise, the broader destination is an AI workload exchange: one procurement and control surface, several modality-specific execution contracts. Calling all of that “tokens” is convenient until finance or security needs to explain what actually happened.

2. Two markets, one converging control point

General AI routing: maximize the speed of discovery

The general-access market reduces integration and purchasing friction. OpenRouter, Hugging Face Inference Providers, Requesty, Eden AI, AI/ML API, Commonstack and other aggregators make it easier to try multiple providers. Vercel and Cloudflare connect access to developer platforms. fal, Replicate, Segmind and WaveSpeed emphasize the rapidly changing media ecosystem. Their relevance is broader than small developers: they can be the fastest way for enterprise teams to establish whether a capability is useful at all. [15] [16] [17] [18] [19] [20] [21] [22] [23] [11] [12] [13]

The early return is option value. A team can test a new specialist against its existing baseline without waiting for a separate account, SDK, procurement workflow and secrets configuration. Consolidated purchasing can reduce operational friction. Bring-your-own-key arrangements preserve provider relationships but do not magically eliminate upstream onboarding. Both are useful; they solve different problems. [24]

The risk is the friction removed from experimentation also being removed from data disclosure. An employee able to select almost any model may be able to select almost any processor of sensitive information. A broad sandbox must therefore be broad in capabilities, not indiscriminate in data access. Sanitized test corpora, scoped identities, short-lived credentials, spend limits and approved processor rules belong at the front door.

Enterprise AI routing: maximize the safe delegation envelope

The enterprise market adds enforceable identity, organizational budgets, data boundaries, approved model/provider sets, change control, operational evidence and contractual accountability. Prisma AIRS/Portkey, LiteLLM, Bifrost, TrueFoundry, TensorZero, Braintrust, Respan, LangDB, Orq.ai and nexos.ai approach that requirement from different directions. API/security platforms such as Kong, Tyk, Gravitee, WSO2 and F5 bring an existing control discipline; cloud and data platforms embed routing into their own estates. [25] [26] [27] [28] [29] [30] [31] [32] [33] [34] [35] [36] [37] [38] [39]

Enterprise does not mean “everything on premises.” Nor does general-purpose mean “unsafe.” A managed platform with enforceable regional restrictions and reliable operations may be preferable to a poorly maintained private proxy. An open-source gateway gives the enterprise control over deployment; it does not supply an operations team, an incident process or a guarantee that software artifacts are trustworthy.

The most durable distinction is who retains authority. Does the enterprise own eligibility, policy versions, outcome evidence and escape routes? Or does it merely select a service tier and hope the provider's optimization remains aligned?

3. The differentiation map: buy the job, not the buzzword

Frontier segments the market by the primary job a buyer is hiring it to perform. These are analytical placements based on documented product emphasis, not market-share rankings or scored performance results. Products overlap; a placement does not imply missing capabilities elsewhere.



Figure 1. General AI routing differentiation map: 29 named offers and projects grouped into access markets, adaptive selection, media platforms and developer platforms. Qualitative categories, not rankings.

Figure 1. General AI and innovation-led routing. Broad-access exchanges compete on reach and onboarding; media platforms on specialized execution and workflow; adaptive selectors on deciding which model or budget should handle a task. Developer-platform entry points and open distribution projects connect these groups. The directory provides current names, URLs and qualification questions. Logos/icons identify products, not endorsement.



Figure 2. Enterprise AI routing differentiation map: 27 named offers and projects grouped into governed platforms, API/security fabric, cloud/data estates and open distribution. Not a production certification.

Figure 2. Enterprise routing by prime differentiation. Governance/evaluation platforms, API/security fabrics and cloud/data estates start with different assets. Choose the operating control you need before comparing catalogs. Portkey is grouped with Prisma AIRS; Plano with DigitalOcean, avoiding false supplier diversity.

Three procurement consequences follow. First, a marketplace and an enterprise gateway can coexist: the enterprise gateway owns policy and credentials while an approved marketplace provides eligible supply. Second, a decision-only selector can sit beside the gateway rather than becoming another full data-plane intermediary. Nadir documents that separation; it is a useful architectural option, not evidence that its recommendations will generalize to every workload. [40]

Third, provider diversity must be measured underneath the logos. An apparent two-vendor strategy can share the same model, cloud region, quota account or safety processor. Conversely, one cloud's model catalog can contain several model suppliers without providing independence from that cloud's identity, networking or billing failure domain. The correct comparison is the execution path and its correlated dependencies.

Consolidation changes the neutrality question

Stripe's August announcement is an agreement to acquire OpenRouter; this report does not infer closing from the announcement. Palo Alto Networks' July GA announcement explicitly says its Portkey acquisition had closed. DigitalOcean's April announcement describes a completed Katanemo acquisition. Those different transaction states matter. [6] [7] [8]

Our inference is that routing is becoming strategic infrastructure because it controls both demand allocation and the evidence used to justify that allocation. Financial infrastructure wants to optimize consumption and margins. Security infrastructure wants to authorize machine work. Clouds want to keep workloads within their delivery systems. None is inherently disqualifying. All make incentive transparency, policy portability and a change-of-control clause more valuable.

Fireworks Nexus illustrates the convergence: a supplier of inference capacity now also selects between open and closed coding models through FireRouter, while exposing user budgets and cost-per-merged-PR reporting. This is an instructive combination of supply, selection and evidence—not proof that the supplier’s preferred route will win on an enterprise’s own work. [41]

4. What the research proves—and what it does not

The academic case begins with model complementarity. The best model in aggregate is not necessarily the best eligible model for every task. Routing has value when cheaper, faster or specialized models can satisfy requirements that would otherwise be assigned to a general frontier model. But a theoretical oracle that knows every model's outcome in advance is not a deployable router. The research problem is predicting enough of that advantage without spending it all on classification, redundant execution or errors.

A compact evidence ledger

FrugalGPT, 2023. Chen, Zaharia and Zou reported up to 98% cost reduction while matching the best individual model in studied settings, or improved accuracy at matched cost. The important contribution is the cascade and budget-allocation idea. The percentage depends on the historical model set, data and objective; it is not a 2026 enterprise savings estimate. [9]

RouteLLM, 2024/2025. Ong and colleagues trained routers using preference data and reported more than a twofold cost reduction in some benchmark settings without reducing response quality. Transfer to different strong/weak model pairs is encouraging. It is not proof that an unconstrained, many-provider router will preserve legal, security or agentic outcomes. [4]

LLMRouterBench, Findings of ACL 2026. The benchmark covers more than 400,000 instances from 21 datasets, 33 models and 10 representative routing baselines. It confirms complementarity but finds several recent approaches, including commercial routers, do not reliably outperform a simple baseline. It also finds diminishing returns from enlarging the model ensemble relative to careful curation. The buyer implication is direct: compare against your best permitted single model and simple rules, not only against an expensive straw man. [5]

TwinRouterBench, May 2026 preprint. Its static track contains 970 router-visible prefixes from 520 instances across five workloads. Its dynamic harness supports the 500-case SWE-bench Verified suite, but the paper reports a 100-case held-out evaluation. The contribution is testing whether a model substitution at one agent step preserves final task resolution and realized spend. The harness capacity must not be reported as the paper's measured sample size. [42]

RouterWorld, October 2026 preprint. The authors assemble 371,000 single-turn responses and 76,000 multi-turn trajectories, evaluating 13 baselines. Its key contribution is partial observability: production routers see outcomes only for routes actually taken, and those choices change future training data. Latency is explicit, but its hardware-grounded annotations are not measurements of every commercial provider's live queue or outage behavior. This research was performed at Amazon. [43]

SCRouter, October 2026 preprint. Its benchmark covers 38,869 utility/security queries, 12 models and 12 budgets—144 action combinations—plus 2,753 out-of-domain queries. It argues that utility and security prefer different model/budget choices. Reported gains use Oracle Gap Recovery; they are not percentage-point gains in enterprise accuracy. The practical lesson is that a single “quality” score can conceal the property that matters most. [44]

These studies are not a universal vendor leaderboard. They use different model generations, prices, datasets, judges, budgets and notions of success. We have not run a new head-to-head production benchmark for this report. Treat the literature as evidence for mechanisms and evaluation design, not as transferable service-level guarantees.

The stronger academic argument: constrain first, optimize second

Let a workload have context x, enterprise policy P and operational state s. Define the eligible action set A(x,P,s) as the intersection of permitted models, versions, providers, regions, identities, modalities, tool capabilities, retention conditions and minimum service requirements. An action includes its fallback and reasoning budget, not just the first selected model.

The router should select an action only from A. Within that set, it can minimize expected total execution cost subject to a lower bound on accepted task success and an upper bound on completion latency. Alternatively it can maximize expected business utility after cost, with the same hard policy boundaries. These are Frontier's design formulations, not a claim that all surveyed products implement such an optimizer.

If A is empty, the correct result is rejection, deferral or an explicitly authorized human path. It is not a more permissive provider. A data-residency prohibition should not become a soft penalty that can be outweighed by a cheaper token.

For an agent, optimizing each call independently can be wrong. A locally cheaper selection can invalidate a tool sequence, lose cache reuse, break reasoning-state compatibility or cause repeated corrective work. The objective needs to be final task success over a trajectory. TwinRouterBench motivates this distinction; RouterWorld adds the warning that the evidence collected under the current routing policy is selected, not neutral. [42] [43]

This yields a testable proposition: a router is valuable only when its selection benefit exceeds its own operating, decision and error costs within the permitted action space. More models increase theoretical choice; they also increase qualification work, policy complexity and statistical uncertainty. The enterprise needs a curated executable portfolio, not a museum of model IDs.

Production evidence: governance is not inherently a slow path

A 26 October 2026 preprint by PayPal engineers describes the internally operated Cosmos LLM Gateway. The authors report service to more than 12,000 developers and roughly 200 billion tokens daily, with 99.999% availability to business. Their synthetic-upstream core-path benchmark reports 4 ms median and 11 ms p99 overhead above 2,200 requests per second on two 4-vCPU instances. These are operator-reported results, not an independent SLA audit or a normalized comparison with commercial gateways. The architectural lesson is more durable than the headline numbers: keep declarative authorization, brokered credentials, versioned policy and durable usage attribution separate from latency-sensitive streaming translation. This supports governed production routing—not unconstrained model substitution. [45]

5. Cost per outcome: stop buying impressive denominators

The wrong metric is dollars per million tokens in isolation. The better metric is all-in cost per accepted, policy-compliant outcome delivered within its service objective. “Accepted” must be defined by the business or a validated verifier, not retrospectively by the router claiming savings.

A practical accounting definition is:

Cost per accepted outcome = all execution, routing, verification, rework and allocated operating cost / accepted outcomes.

The numerator includes failed calls, paid retries, speculative branches, media generations rejected by reviewers, tools, retrieval, safety/classifier calls, storage, human review and the attributable operating platform. The denominator excludes incomplete, invalid or late work, but those attempts' costs stay in the numerator. Report rejected/deferred requests and total demand separately so a router cannot appear efficient merely by refusing difficult work. This extends Frontier's token-economics thesis and aligns with the FinOps Foundation's emphasis on use-case economics rather than isolated tooling cost. [1] [46] [47]



Figure 3. Hypothetical 1000-task comparison: a 4-cent call yields 5.47 cents per accepted outcome versus 10.86 cents for a 2-cent call after retries, verification and acceptance. Not observed vendor performance.

Figure 3. A cheaper call can buy a more expensive result. Original illustrative scenario, not observed vendor performance. Per 1,000 submitted tasks, Route A spends $36 on model attempts plus $40 on verification/rework and produces 700 accepted outcomes: $0.1086 each. Route B spends $42 plus $10 and produces 950 accepted outcomes: $0.0547 each. B's nominal call price is twice A's; its cost per accepted outcome is about 50% lower. The example holds demand fixed and includes costs from unsuccessful tasks.

The arithmetic is intentionally transparent. Route A assumes $0.02 per attempt and 1.8 attempts per submitted task. Route B assumes $0.04 and 1.05 attempts. The assumed verification/rework costs are $0.04 and $0.01 per submitted task respectively. These are not market prices or a forecast. They demonstrate why reliability, acceptance and rework can dominate a lower input price.

Performance figures need their operating envelope

Maxim's Bifrost benchmark reports 11 microseconds of gateway overhead at 5,000 requests per second on an AWS t3.xlarge, excluding upstream response time. Its separate mock-response comparison reports 0.99 milliseconds of overhead. Both are vendor-run measurements, and neither means an AI answer arrives in microseconds. Authentication, policy plugins, logging, streaming, request size and deployment topology must be reproduced before a buyer uses the numbers for capacity planning. [48]

Not Diamond's Code launch claims more than 20% lower inference cost without quality degradation; its public documentation is early access. Snowflake reports internal examples of up to threefold token efficiency for a dbt pipeline and 25% greater token efficiency for the same number of pull requests. These are useful hypotheses for a pilot, not directly comparable measures of cost per accepted outcome. Token efficiency alone omits human review, elapsed time and failure recovery. [49] [50] [51]

A serious scorecard should therefore show first-pass acceptance, eventual acceptance, task completion by deadline, p95/p99 full completion time, invalid tool-call rate, policy violations, retries per submitted task, cost per accepted outcome and quality by workload slice. Keep time to first token, gateway overhead and upstream generation speed as diagnostic metrics. None is a substitute for successful work.

FinOps is becoming an authorization function

A centralized spend panel is valuable, but after-the-fact accounting is insufficient when agents can create hundreds of calls. Bind budgets to users, workloads, projects and environments; reserve estimated capacity for long jobs; reconcile actual usage; cap retries and tool loops; and preserve a route-level cost trail. Warn before hard limits, but never silently change a high-consequence workflow to an unqualified model simply to stay under budget.

One additional calculation matters: routing can reduce per-outcome cost while increasing total spend because more useful work becomes feasible. That is not automatically failure. The CIO and CFO need separate targets for unit economics, demand growth, realized value and cash exposure. A cost dashboard that cannot distinguish those effects will encourage the wrong decisions. [1]

6. The Routing Constitution: delegate choice, retain authority

A routing partner should be able to improve performance without becoming the institution that decides what the enterprise is allowed to expose. Frontier recommends a Routing Constitution: an enterprise-owned, versioned set of non-negotiable boundaries that surrounds the vendor's optimization logic.

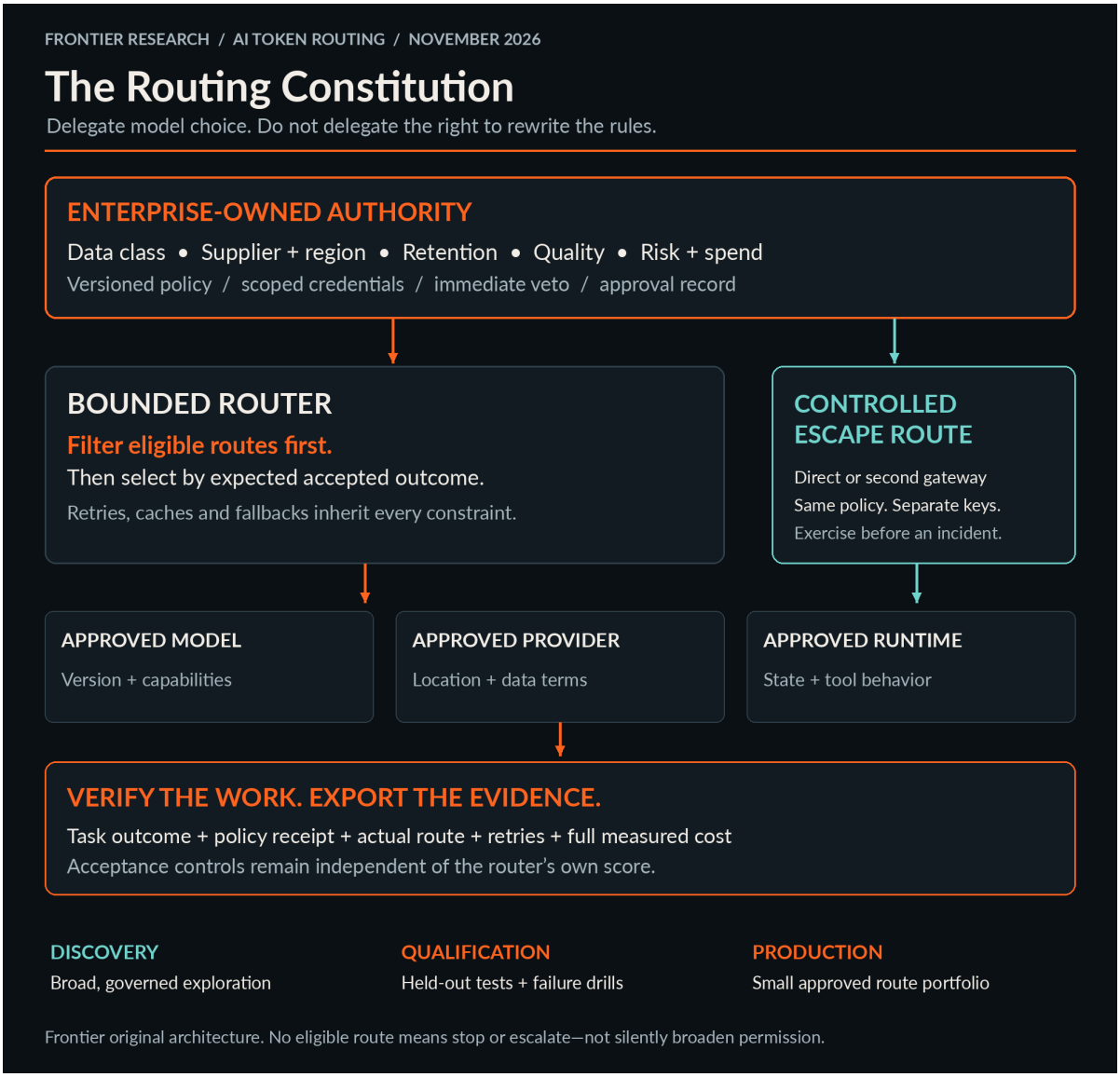


Figure 4. The Routing Constitution: enterprise-owned eligibility and authority govern both the bounded optimizer and independent escape path; verified outcomes and decision receipts close the loop.

Figure 4. The Routing Constitution. Enterprise identity, policy, approval and outcome evidence remain outside the supplier's discretion. The router selects only among eligible execution paths. A separately controlled direct route preserves continuity without bypassing the same safeguards. No eligible route means fail closed or defer—not broaden the trust boundary.

Three environments, three kinds of freedom

The discovery sandbox should make experimentation easy: many approved capabilities, synthetic or appropriately sanitized inputs, disposable projects, scoped credentials, clear spend limits and no authority to write to production systems. Record results so learning survives a developer's project. A one-key experience is useful; a shared organization-wide secret is not.

The qualification sandbox is production-like without production authority. Replay authorized representative workloads, test provider substitutions, inject failures, verify region and retention restrictions, compare output schemas, and run the real identity, logging and safety configuration. Shadow routing should initially recommend a route without executing it; parallel live tests should use data and actions explicitly authorized for that experiment.

The production lane uses a deliberately small approved portfolio. Pin versioned route policies and qualify model/provider combinations. Require a release process for changes to model aliases, guardrails, cache behavior and fallback order. A signed supplier changelog is useful evidence; it does not eliminate regression testing.

Separate credentials, budgets, storage and administrative privileges across all three environments. A production sandbox is not merely a label on the same key. For highly sensitive workloads, separate deployments or accounts may be necessary to obtain the required failure and data boundaries.

Seven non-negotiable controls

Eligibility before optimization. Filter by data class, processing location, retention/training terms, approved supplier, required tool behavior, model version and workload risk before comparing price or predicted quality. Reapply the filter to retries, caches, safety services and nested routers.

Credential containment. Use scoped virtual or workload identities at the gateway, keep upstream credentials in a managed vault where supported, rotate and revoke them independently, and limit each environment's authority. Bring-your-own-key reduces some commercial dependency; it does not mean the intermediary cannot observe requests or misuse credentials it handles.

Decision receipts. Preserve request/workflow ID, policy version, selected model and provider, region where available, reason for selection, fallback sequence, retries, latency, billable units and outcome verification. Minimize prompt retention; metadata may still be sensitive. Require export in an operationally usable format.

Bounded recovery. Distinguish transport failure before a side effect from a timeout after a tool or media job may have executed. Apply workflow-level deadlines, idempotency keys where supported, retry budgets, exponential backoff with jitter and circuit breakers. Never replay a side-effecting action just because an HTTP acknowledgment was lost.

Independent veto. The organization must be able to revoke a provider, model, region or credential immediately without waiting for the router's optimizer to learn. Test that revocation wins against cached policies, fallback and ongoing sessions.

Controlled escape. Maintain at least one independently approved delivery path for critical workloads. It may be a direct provider integration or a second gateway. The escape route must enforce the same data, safety and identity rules; bypassing the intermediary must not mean bypassing governance.

Outcome evidence. Connect routing decisions to business acceptance and incident history. Retain enough independent test data to challenge the vendor's claims. A router trained on its own self-selected traffic can become confident in the choices it already makes.

7. Designing explicitly for loss of control

The selector is also a data processor

A routing classifier, embedding service, semantic cache, safety filter, trace store and media CDN can each see different parts of a workload. “The final model runs in the EU” is not a complete data-flow statement if an earlier classifier or later log export processes the content elsewhere. Inventory the whole path, including support access and exceptional retention. “Zero retention,” “no training” and “no human access” are distinct commitments, not interchangeable labels.

Ask vendors to describe processing per component and per feature. Require changes to subprocessors or exceptional handling to pass the same review as changes to the inference provider. An open or local classifier may reduce disclosure, but only if the rest of the execution path also fits policy.

One key is also one unusually valuable target

Centralization can remove scattered provider secrets and improve revocation. It also creates a privileged concentration of credentials, prompts, quotas and business context. LiteLLM's canonical October investigation documents a compromise of the v1.107.0 release pipeline. The appropriate inference is not that every installation or all current versions are unsafe. It is that the routing layer's build and release chain deserves the protection assigned to other high-authority infrastructure. [52]

Pin verified artifacts, segregate build/signing/release permissions, minimize gateway credentials, control outbound connections, test rotations and maintain clean recovery procedures. This directly echoes Black October: restoring trusted authority is more consequential than merely applying a patch. Self-hosting changes who operates the boundary; it does not remove the software supply chain. [53]

The route can change the application without a deployment

A supplier can update a model alias, selection classifier or “balanced” policy while the application's code stays unchanged. That can alter refusal behavior, output structure, tool selection, latency and cost. Treat a material routing-policy change as an application release. Require version pinning or advance notice, canary traffic, rollback and a defined maximum period for retaining the previous behavior.

An aggregate quality score can hide concentrated harm. Test languages, long-context work, rare but high-impact decisions, security-sensitive prompts, tool-heavy sessions and modality-specific edge cases separately. A model that improves average prose quality may worsen a regulated extraction task. SCRouter's distinction between utility and security is a warning against letting a single scalar decide all of this. [44]

The optimizer can be manipulated

Rerouting LLM Routers demonstrates that routing decisions themselves can be steered by adversarial inputs, including toward expensive models. The result establishes a control-plane integrity problem distinct from whether the final model answers safely. Enterprises should test whether untrusted prompt content can override cost tiers, supplier eligibility or administrative instructions; routing rules should not take policy authority from the text they are classifying. [54]

The business response is bounded authority, not merely a stronger prompt. Separate trusted policy inputs from user content, impose per-identity and per-workflow ceilings, monitor unusual escalation patterns and maintain a deterministic safe path for anomalous traffic.

API portability is not behavioral portability

The cheapest route can destroy a valuable cache, require a different prompt template, mishandle a tool result or fail to preserve provider-specific reasoning state. A common request schema lowers integration cost; it does not make every field or behavior semantically identical. Qualification should test unsupported parameters, context truncation, structured outputs, multimodal fidelity, reasoning budgets and whether adapters silently drop fields.

This matters especially for agent sessions. A provider substitution mid-trajectory may invalidate assumptions the agent made earlier. Prefer session affinity when appropriate, and allow migration only at explicit checkpoints that reconstruct the necessary state. Evaluate final task completion after the switch, not just whether the next call returns text.

Economic incentives can become invisible policy

A router may earn through subscription fees, credit margins, routing charges, volume commitments or other commercial arrangements. Those models are not inherently improper. They do create different incentives. Ask whether ranking incorporates rebates, owned capacity, commercial commitments or reseller margins, and whether the buyer can exclude those factors or inspect their effect.

One invoice does not mean one accountable supplier. Define who is responsible for data processing, unauthorized rerouting, service failure, credits and incident notification. A refund for an unsuccessful API request is not compensation for a damaged business process. Do not let a small credit replace a meaningful continuity obligation.

The next lock-in may be evidence, not inference

Models can be replaced. A history of accepted outcomes, evaluation datasets, route experiments, tuned policies and approval decisions can be much harder to reconstruct. An apparently neutral router can become sticky through the evidence required to operate safely. The enterprise should own exportable policies, evaluation corpora, labels and decision history, with an exercised migration procedure.

That is the less obvious strategic risk: routing can make model switching easier while making the routing system harder to replace. Buy its learning, not permanent dependence on its memory.

8. How to select a routing partner without buying a benchmark trick

Start with three workload classes, not a vendor demonstration. Choose a high-volume, machine-verifiable task; a multi-step agent workflow; and a sensitive workload with hard data constraints. Add a creative/media workload when it is material to the business. Define acceptance before showing suppliers the test.

For each class, compare four baselines: the best permitted single model; a simple deterministic routing policy; the candidate's adaptive routing; and, where appropriate, a verified cascade. The baseline must be commercially and operationally realistic. Comparing a router against an unnecessarily expensive model is not a useful business case.

Use representative held-out data with a separate calibration set. Report uncertainty, not just point estimates. For paired tasks, compare success and cost under matched conditions; for nondeterministic agents, repeat enough trials to expose variability. Apply stratification to rare high-consequence failures rather than letting easy tasks dominate the average. Keep a time-based holdout to reveal model and workload drift.

Do not route real customer data to extra vendors merely to obtain a counterfactual benchmark. Use authorized replay, synthetic equivalents or a specifically approved experiment. A decision-only shadow route measures predicted choices, not actual cost savings or completion quality. Promotion requires executed evaluation in the permitted environment.

The failure test vendors should expect

Simulate a provider outage, exhausted quota, slow streaming, a timeout after a tool side effect, an unavailable verifier and a sudden loss of the preferred region. Remove one approved provider while a session is active. Change a model alias. Trigger the budget ceiling. Confirm that the system does not silently discard a required parameter or move confidential data to an unapproved path.

For media, include duplicate webhook delivery, cancellation after execution starts, expired output URLs, partial assets and a generation that passes technical checks but fails creative acceptance. Measure when the organization learns of the failure and whether it can avoid paying twice or repeating a consequential action.

A supplier unwilling to expose these boundaries is not ready to own a critical path. A supplier that documents them clearly may be the better choice even when its headline model list or per-token price is less impressive.

The procurement scorecard

Frontier recommends gating eligibility before scoring commercial attractiveness. Fail the procurement when required security, residency, authority or recovery conditions are unmet. Among passing candidates, weight accepted outcome quality, operational reliability, time to integrate/qualify new capabilities, evidence and policy portability, and all-in economics according to workload. Do not publish a universal weighted ranking: a research sandbox and a high-impact production workflow have different utility functions.

Obtain the actual security report scope, contractual data-processing terms, incident process, supported deployment matrix and service-level definitions. Check exclusions for upstream provider failures. Clarify whether an availability credit applies to the gateway accepting requests, a model producing output or the end-to-end service. Those are materially different promises.

Require export rights, transition assistance, revocation controls, notice of material policy/model changes, independent route-level audit evidence and a tested emergency path. For a managed platform, clarify where the control plane and data plane run. For a self-hosted gateway, define who patches, signs, upgrades, monitors and restores it.

9. A 90-day operating plan

Days 1–30: make access visible and safe

Inventory model accounts, keys, direct integrations, shadow spending and AI-enabled tools. Appoint an accountable platform owner with security, finance and business outcome partners. Establish data classes and a minimal eligible-model registry. Create the discovery sandbox with separate credentials and no production write authority. Baseline existing tasks before introducing intelligent routing.

The exit criterion is not “we installed a gateway.” It is that the team can attribute a request, its spend, its processor and its permissions, revoke access, and explain which data may enter the sandbox.

Days 31–60: prove two bounded production routes

Select one verifiable text/coding workflow and one additional priority workload. Run the qualification protocol, including the failure tests above. Set completion and cost-per-outcome targets, establish quality floors and cap exposure. Start with a small model portfolio and simple rules. Add learned selection only when it improves the agreed objective over those baselines.

The exit criterion is independent evidence that the selected route stays inside policy, preserves task success and can fail safely. A dashboard showing fewer tokens is not enough.

Days 61–90: institutionalize choice and the ability to leave

Promote qualified configurations through a canary process. Publish a routing policy registry, accountable owners, approved fallbacks and incident runbooks. Reconcile gateway costs to provider invoices and business outcomes. Exercise supplier revocation and the alternate path. Export the policies and evidence, then restore them outside the primary service to expose hidden dependencies.

The exit criterion is an operating capability: teams can discover new models quickly, qualify them reproducibly, deploy them within constraints and replace a routing partner without discarding governance.

When not to add another intermediary

A single stable workload on one qualified model may be better served directly, especially where the provider supplies the necessary identity, private connectivity and monitoring. A low-volume application may not recover the engineering cost of adaptive selection. Extremely sensitive work may require a tightly controlled local or private model with no broad external routing. The objective is not to maximize the number of routers. It is to establish a deliberate model-access and change-control layer proportional to the problem.

For organizations experimenting across many models and modalities, lacking such a layer increasingly creates avoidable friction. The right conclusion is not “buy any router.” It is stop making every team rebuild model access, security and economics from scratch.

10. Vendor and project directory: the practical shortlist universe

The following profiles separate documented differentiators from Frontier's buyer tests. URLs are vendor/project entry points; linked citations identify the evidence used. “Documented offer” means a public product or technical surface was available to inspect, not that every feature is generally available or covered by an enterprise SLA. Current entitlements, certifications, regional processing and contractual terms require diligence.

10.1 Access markets

OpenRouter — Documented offer. Broad model and provider access, consolidated purchasing, provider selection and enterprise workspaces. Its media APIs make the story larger than chat.

Buyer’s test: Distinguish model selection from hosting-provider selection; lock retention, region, provider and fallback conditions. Stripe announced an acquisition agreement, not evidence here of a completed closing. [15] [10] [6]

Website: https://openrouter.ai

Hugging Face Inference Providers — Documented offer. A model-centric access layer connecting the Hub to participating inference providers, using centralized billing or provider credentials.

Buyer's test: The fastest, cheapest and preferred selectors principally select a provider for a named model. Do not mistake this for a business-task quality oracle. [16] [24]

Website: https://huggingface.co/docs/inference-providers

Requesty — Documented offer. Unified model access with routing, observability and team-oriented controls.

Buyer's test: Translate approved-model commitments into explicit model/provider/region tests on every fallback, and inspect the signed service terms rather than a generic compliance badge. [17] [55]

Website: https://www.requesty.ai

Eden AI — Documented offer. Breadth across generative and specialized AI services, useful for comparing multiple types of AI capability.

Buyer's test: Inventory the endpoint-specific interfaces, data processors and failure semantics; a unified commercial layer does not make OCR, audio and generation interchangeable. [18]

Website: https://www.edenai.co

AI/ML API — Documented offer. Multimodal model aggregation plus JevRouter-style selection under one commercial entry point.

Buyer's test: Separate router decision charges, model charges and fallback tiers. Check which modalities actually support automatic selection rather than just access. [19]

Website: https://aimlapi.com

Commonstack / UncommonRoute — Documented offer. Unified upstream access and an open local routing layer for model selection.

Buyer's test: Evaluate actual route disclosure, upstream authorization and support commitments. A routing brand need not own the inference capacity it resells. [20]

Website: https://commonstack.ai

ZenMux — Documented offer. Model and provider routing plus a distinctive compensation proposition for qualifying service problems.

Buyer's test: Treat compensation eligibility, caps and exclusions as contract terms, not insurance against business loss or proof of answer quality. [56]

Website: https://zenmux.ai

LLMGateway — Documented offer. Configurable gateway routes with provider-aware fallbacks and documented compatibility constraints.

Buyer's test: Dynamic routes are documented for chat/responses/messages, not every media endpoint. Verify context, tool and reasoning-state behavior for each route. [57] [14]

Website: https://llmgateway.io

Routeway — Emerging. A model-access and routing service with an enterprise-oriented commercial offering.

Buyer's test: Verify upstream supplier authorization, exact privacy/retention terms, support and workload-specific reliability before sensitive production use. [58]

Website: https://routeway.ai

Abacus.AI RouteLLM API — Documented offer. Unified access, prompt-based model selection and failover alongside the broader ChatLLM offering. Its commercial RouteLLM service is distinct from the academic RouteLLM project.

Buyer's test: Verify API credits and endpoint-specific modalities separately from consumer subscription benefits. Test custom routing, chosen-model disclosure, retention and enterprise key/hostname requirements. [59]

Website: https://routellm-apis.abacus.ai/

10.2 Media platforms

fal — Documented offer. Media-specialist model access and execution for image, video, audio and related creative workloads.

Buyer's test: Treat it as a media execution platform, not proof of general-purpose intelligent LLM routing. Measure accepted asset quality, job lifecycle and rights. [23]

Website: https://fal.ai

Replicate — Documented offer. A developer-friendly model and prediction execution surface with model versions and asynchronous jobs.

Buyer's test: Preserve exact versions and input schemas; validate output retention, cancellation, duplicate jobs and enterprise deployment terms. [11]

Website: https://replicate.com

Segmind / PixelFlow — Documented offer. Composable media model workflows that can be published as APIs.

Buyer's test: Workflow-level correctness and asset continuity matter more than whether a single generation request succeeds. Version the complete graph. [12]

Website: https://www.segmind.com

WaveSpeedAI — Documented offer. Broad image, video, audio and 3D endpoint access through a common API.

Buyer's test: Recheck each model license, source of capacity, retention and output quality. Endpoint counts are not counts of independent suppliers. [13]

Website: https://wavespeed.ai

Kie AI — Diligence-first. A fast-moving marketplace for media and additional AI endpoints behind common account utilities.

Buyer's test: For sensitive use, obtain contractual answers on authorized supply, subprocessors, data handling and support. Inclusion is not an enterprise endorsement. [60]

Website: https://kie.ai

10.3 Adaptive selection

Not Diamond Code — Early access. Coding-oriented selection across models and reasoning budgets, with session and cache economics central to its proposition.

Buyer's test: The public Code documentation is early-access/waitlist material. Its 20%+ savings claim is vendor-reported, not a guarantee on the buyer's repositories. [50] [49]

Website: https://code.notdiamond.ai

DigitalOcean Inference Router / Plano — Documented offer. Task-aware managed routing backed by the open Plano data plane and Katanemo routing-model research.

Buyer's test: Qualify the configured model pool and the cost of classification itself. Katanemo is now part of DigitalOcean; do not count it as a separate independent supplier. [61] [8]

Website: https://www.digitalocean.com

Martian — Confirm product scope. Historically known for learned selection; current public gateway documentation offers standardized multi-model access.

Buyer's test: Ask which learned-routing capabilities, models, deployment forms and service guarantees are currently orderable. Historical research results are not current product entitlements. [62]

Website: https://withmartian.com

Nadir — Emerging. Decision-only routing beside an existing gateway, or a managed proxy with optional answer verification and escalation.

Buyer's test: A shadow recommendation is not an executed saving. Verify privacy of classifier inputs and evaluate cascades on full task resolution, including escalation cost. [40]

Website: https://getnadir.com

Neutrino AI — Emerging. Prebuilt multi-model engines and custom enterprise routing propositions.

Buyer's test: Require a current supported-model list, fresh evaluation and explicit operating commitments; public engine descriptions are not comparative reliability evidence. [63]

Website: https://www.neutrinoapp.com

BlockRun / ClawRouter — Diligence-first. A locally evaluated routing layer combined with wallet-based, per-call payment and a managed model-access path.

Buyer's test: The reviewed documentation describes a safety-net override when all models in a tier are excluded. Require an unambiguous fail-closed contract before enterprise use; local selection does not mean prompts never reach the gateway. [64]

Website: https://blockrun.ai

vLLM Semantic Router — Open project. An open, programmable decision layer separating routing policy and signals from model backends, with local, private and cloud options.

Buyer's test: Do not confuse it with the vLLM inference engine. The operator owns classifier deployment, policy correctness, upgrade compatibility and service availability. [65]

Website: https://vllm-sr.ai

TensorOpera — Documented offer. Model routing within a broader agent, model-serving and GPU platform.

Buyer's test: Separate agent selection, model selection and compute placement. Ask for current router evaluations and deployment rights rather than assuming historical research results apply to the full stack. [66]

Website: https://tensoropera.ai

Fireworks Nexus / FireRouter — Documented offer. Coding-focused routing between open and closed models, integrated with Fireworks inference, developer harnesses, user budgets and outcome reporting. Difficult tasks can pass to a closed provider on the customer's key.

Buyer's test: Test the exact model-and-harness route and accepted pull-request outcome. Supplier-published savings, cache rates and model evaluations are not independent proof of router-wide production improvement; preserve your external policy gateway. [41]

Website: https://fireworks.ai/nexus

Arcee Conductor — Reconfirm current offer. Documented task-, domain-, complexity- and language-conditioned selection between smaller Arcee models and larger closed models. It makes the classifier and model-pool concept unusually explicit.

Buyer's test: Public Conductor examples retain older model names and update dates. Reconfirm current orderability, catalog, support and contractual scope rather than treating historical savings claims as current results. [67]

Website: https://docs.arcee.ai/arcee-conductor

Aurelio Semantic Router — Open project. Open-source semantic decision layer routing inputs to handlers, models or tools using vector-space matching and configurable decision components. Useful when teams want to own the selector.

Buyer's test: A routing library is not a managed model marketplace or enterprise gateway. You must supply evaluation, credentials, policy, fallback and operational support; semantic similarity is not outcome verification. [68]

Website: https://www.aurelio.ai/semantic-router

10.4 Governed platforms

Prisma AIRS AI Gateway / Portkey — Documented offer. Palo Alto Networks combines Portkey-origin gateway capability with runtime security, identity, model, MCP and agent traffic governance.

Buyer's test: Test actual fail-closed enforcement and every relevant traffic protocol; separate SaaS and hybrid responsibilities. Portkey and Prisma AIRS are not independent shortlist entries. [25] [7]

Website: https://www.paloaltonetworks.com/prisma/airs

LiteLLM — Documented offer. Broad provider adapters, self-operated gateway options, virtual credentials, budget controls and configurable fallback.

Buyer's test: Own deployment and supply-chain integrity. The October 2026 release compromise makes verified artifacts, credential rotation and restricted release permissions a procurement issue. [26] [52]

Website: https://www.litellm.ai

Bifrost / Maxim AI — Documented offer. A performance-oriented gateway with enterprise operational controls and adaptive load distribution.

Buyer's test: Its microsecond overhead measurements exclude upstream inference; reproduce with your authentication, logging, streaming and policy plugins enabled. [27] [48]

Website: https://www.getmaxim.ai/bifrost

TrueFoundry — Documented offer. A gateway integrated with enterprise AI infrastructure and private/hybrid deployment needs.

Buyer's test: Evaluate where policy, telemetry and credentials reside in your selected deployment; confirm upgrade, failover and support ownership. [28]

Website: https://www.truefoundry.com/ai-gateway

TensorZero — Documented offer. An open inference platform linking gateway events, evaluations, experimentation and model/prompt optimization.

Buyer's test: Feedback is a strategic asset only when the enterprise owns reliable outcome labels and governs experiments. Self-operation carries an operational burden. [29] [69]

Website: https://www.tensorzero.com

Braintrust Gateway — Documented offer. Model access closely connected to tracing, evaluation datasets, scoring and release checks.

Buyer's test: Evaluate the exact gateway release/SLA and logging scope. Observability and caching expand the data boundary unless configured deliberately. [30]

Website: https://www.braintrust.dev

Helicone — Reconfirm support. Gateway and observability with an open-source/self-hosting path.

Buyer's test: Public corporate-transition material makes current stewardship and hosted support worth reconfirming. An accessible repository is not a managed-service continuity guarantee. [70] [71]

Website: https://www.helicone.ai

Respan / Keywords AI — Documented offer. Gateway functionality connected to tracing and AI application evaluation under the Respan identity.

Buyer's test: Confirm which legacy Keywords AI interfaces remain supported and how provider keys and traces are isolated between environments. [31]

Website: https://www.respan.ai

LangDB — Emerging. Routing and tracing with cloud/self-hosted configuration paths and policy-oriented controls.

Buyer's test: Check demonstrated production scale, policy versioning and credential separation rather than extrapolate from catalog breadth. [32]

Website: https://langdb.ai

Orq.ai — Documented offer. Gateway controls within a broader AI development, evaluation and operational platform.

Buyer's test: Compare the value of the whole lifecycle against its complexity; validate routing coverage by modality and enterprise deployment contract. [33]

Website: https://orq.ai

nexos.ai — Documented offer. Central model access, organizational governance and intelligent routing for enterprise adoption.

Buyer's test: Check regional processing, private connectivity, context transformation and the portability of policies and audit history. [34]

Website: https://nexos.ai

10.5 Developer platforms

Vercel AI Gateway — Documented offer. Model access and routing closely integrated with the AI SDK and application developer workflow.

Buyer's test: Useful beyond Vercel-hosted applications, but confirm provider-specific features, ZDR constraints and the distinction between one invoice and independent supply. [21] [72]

Website: https://vercel.com/ai-gateway

Cloudflare AI Gateway — Documented offer. An edge-adjacent gateway with visibility, caching, dynamic rules and an Auto Router capability.

Buyer's test: Auto Router is documented as beta. Distinguish gateway access, Workers AI inference and adaptive model selection; they are different layers. [22] [73]

Website: https://developers.cloudflare.com/ai-gateway

AssemblyAI LLM Gateway — Documented offer. A speech-centric route into language models, attractive where transcription and downstream reasoning form one pipeline.

Buyer's test: Measure complete speech-to-decision latency and data handling, rather than compare only a text gateway's dispatch overhead. [74]

Website: https://www.assemblyai.com/products/llm-gateway

10.6 API/security fabric

Kong AI Gateway — Documented offer. AI-native routing and policy enforcement for teams already thinking in API management and traffic controls.

Buyer's test: Separate product/deployment generations and test semantic-router side calls, model aliases and MCP/tool policy, not just the ingress rule. [35]

Website: https://developer.konghq.com/ai-gateway

Tyk AI Gateway / AI Studio — Documented offer. AI access and controls connected to API lifecycle and developer/organization management.

Buyer's test: Request the specific feature and deployment matrix for the edition being bought; gateway governance and learned selection are separate questions. [36]

Website: https://tyk.io/ai-gateway

Gravitee — Documented offer. A single governance proposition spanning LLM, MCP and A2A interactions.

Buyer's test: Validate authorization at the individual tool and action, and confirm that streaming responses cannot escape inspection policy. [37]

Website: https://www.gravitee.io/platform/ai-gateway

WSO2 AI Gateway — Documented offer. AI traffic routing and policy as part of an API platform, including multi-provider configurations.

Buyer's test: Use the current product/version documentation: older API Manager and newer gateway deployment behavior should not be treated as equivalent. [38]

Website: https://wso2.com/api-platform

F5 AI Gateway — Documented offer. Model routing, MCP control and AI guardrails combined with a security/traffic-management operating model.

Buyer's test: Vendor cost-reduction percentages are not a contractual outcome. Test enforcement, distributed failure modes and complete inspection overhead. [39]

Website: https://www.f5.com/products/ai-gateway

Agent Router (formerly Envoy AI Gateway) — Open project; commercial options. Open-source control plane for AI/model and MCP traffic, with Envoy carrying the data path. The project moved to the Agentic AI Foundation in September 2026; existing API and deployment names remain.

Buyer's test: Distinguish operating the open project from buying the separately linked Tetrate hosted service. Verify release-specific controls, credentials, supported protocols and who owns upgrades and incident response. [75] [76]

Website: https://theagentrouter.ai

agentgateway / Solo.io — Open project + commercial. An open agent-aware data plane with virtual models, priority routing and an enterprise ecosystem.

Buyer's test: Separate open-project features from Solo's commercial support; test model and tool traffic as one authorization graph. [77]

Website: https://agentgateway.dev

IBM API Connect / watsonx.ai — Documented offers. API Connect applies AI policies through DataPower; watsonx.ai Model Gateway separately provides a unified model interface, vault-backed credentials, access policies, limits and load balancing.

Buyer's test: These are distinct integration paths, not one universal auto-selector. Match the product, edition and deployment to the required governance boundary and confirm supported model endpoints. [78] [79]

Website: https://www.ibm.com/products/api-connect

Azure API Management AI gateway — Documented offer; preview components. Policy-oriented model, agent and tool traffic management: identity, token quotas, caching, backend balancing, circuit breakers and telemetry. It can govern non-Microsoft backends.

Buyer's test: Separate this gateway from Foundry's model-selection service. Capabilities vary by tier; the unified multi-provider model API is documented as preview. Evaluate translation and failover with your protocols. [80]

Website: https://learn.microsoft.com/en-us/azure/api-management/genai-gateway-capabilities

10.7 Cloud/data estates

Amazon Bedrock — Documented offer. Intelligent prompt routing inside a cloud-native model service and its identity/billing estate.

Buyer's test: Documented prompt-router eligibility is scoped to supported models and families; it is not unrestricted cross-provider or cross-cloud arbitrage. [81]

Website: https://aws.amazon.com/bedrock

Microsoft Foundry — Documented offer. A deployable model router with quality/cost/balanced policies and a multi-vendor model pool.

Buyer's test: The October 2026 documentation lists 35 supported models; some require separate deployments. Apply Azure Policy and region/version constraints to the effective pool. [82] [83]

Website: https://ai.azure.com

Google Cloud Apigee / API Gateway — Documented offer. Apigee provides enterprise AI API governance; API Gateway also documents a distinct Model Garden routing capability.

Buyer's test: Do not conflate the two products or assume their scope and lifecycle match. Verify the supported deployed endpoints and protocols. [84] [85]

Website: https://cloud.google.com/apigee

Databricks Unity AI Gateway — Documented offer. Model services, routing and AI governance close to the data and evaluation estate.

Buyer's test: Distinguish weighted endpoint routing from any adaptive selection feature; confirm model-service and workspace entitlement, availability and boundaries. [86]

Website: https://www.databricks.com

Snowflake Cortex AI Gateway — Documented offer. Governed model access, attribution and quotas alongside Snowflake data; dynamic selection is an explicit product direction.

Buyer's test: Check release/region eligibility and default privileges. Internal token-efficiency examples do not establish external workload success or lower all-in cost. [87] [51]

Website: https://www.snowflake.com

Alibaba Cloud Model Studio — Documented offer. Intelligent routing with quality-first or cost-first selection within a configured model set.

Buyer's test: The reviewed documentation scopes it to text input, Chat Completions and Beijing/Singapore regions. Confirm applicable contract, version and data-processing conditions. [88]

Website: https://www.alibabacloud.com/product/modelstudio

10.8 Open distribution

New API — Open project. Self-hostable model/channel aggregation, billing and distribution tooling used by individuals and internal platforms.

Buyer's test: It is software, not proof of authorized upstream supply or enterprise assurance. The operator must provide security, upgrades, policy and vendor diligence. [89]

Website: https://docs.newapi.pro

11. Adjacent infrastructure and research that belongs on the radar

The routing layer does not replace model developers or inference operators. Direct model APIs from OpenAI, Anthropic, Google, Mistral, Cohere, xAI and major open-model developers remain the supply and capability reference points. Their inclusion in another provider's catalog is not evidence of identical licensing, service levels or data terms. Buyers should use the model provider's current contract and model documentation as the source of truth, not the reseller's shorthand.

Likewise, Together AI (https://www.together.ai), Fireworks AI (https://fireworks.ai), DeepInfra (https://deepinfra.com), Novita AI (https://novita.ai), SiliconFlow (https://siliconflow.com), Groq (https://groq.com), Cerebras (https://www.cerebras.ai), Baseten (https://www.baseten.co) and SambaNova (https://sambanova.ai) belong in the supply-side diligence universe. These companies supply inference capacity or deployment specialization. Fireworks also offers Nexus/FireRouter and is profiled in the adaptive-selection directory; its inference business is not counted as an additional routing vendor. The others are not automatically classified as neutral cross-vendor intelligent routers. They can be direct routes, backends to a gateway or alternatives to a reseller chain. The buyer must compare the actual endpoint and contractual service. [90] [91] [92] [93] [94] [95] [96] [97] [98]

NVIDIA Dynamo (https://docs.nvidia.com/dynamo) and llm-d (https://llm-d.ai) address another layer: distributed inference execution, scheduling and cache-aware placement. Dynamo's documentation explicitly describes disaggregated serving and KV-aware routing. Those capabilities can improve a chosen backend, but are not substitutes for the enterprise's decision about which provider, model or data jurisdiction may execute a task. Do not mix GPU placement benchmarks with cross-model selection benchmarks. [99] [100]

The research toolkit also deserves a place on the shortlist. RouteLLM (https://github.com/lm-sys/RouteLLM) provides a learnable routing baseline. LLMRouter (https://arxiv.org/abs/2608.06867) is a broader research library, not an enterprise SLA. LLMRouterBench (https://github.com/ynulihao/LLMRouterBench) helps challenge easy benchmark narratives. TwinRouterBench (https://github.com/CommonstackAI/TwinRouterBench) emphasizes live agent completion. RouterWorld (https://arxiv.org/abs/2610.26511) focuses on latency and partial observability. SCRouter (https://arxiv.org/abs/2610.18566) explores security-conditioned model and budget choice. These are methods, datasets and projects—not interchangeable production suppliers. [4] [101] [5] [42] [43] [44]

Historical names require fresh verification. Unify (https://unify.ai) remains relevant to the history of commercial routing, but the current public product direction reviewed here emphasizes continual learning. This report does not treat historical routing descriptions as proof of a currently orderable enterprise router. Martian and Helicone remain in the directory with explicit product/support qualification questions. A company's past category membership is not its current delivery contract. [102] [62] [71]

12. The Frontier judgment: the market is selling choice. Keep the power to choose.

The argument for routing is larger than cost reduction. A well-designed access layer can let an organization discover capabilities faster, compare suppliers with less friction, concentrate security controls, allocate spend and move useful workloads as the market changes. That is why it is becoming a consequential component of the enterprise AI operating model.

But the convenience can conceal a transfer of authority. The partner that sees demand, chooses supply and reports the results can become both broker and judge. The enterprise should not surrender all three functions together without independent evidence and the power to intervene.

The winning architecture is deliberately uneven: expansive discovery, rigorous qualification, constrained production and continuous outcome measurement. Intelligent selection is welcome inside that design. It is not permitted to rewrite the design when a provider fails or a cheaper model appears.

Do not optimize the price of intelligence while outsourcing the right to decide what counts as acceptable work.

Rent the models. Own the rules. And make the exit work before the dependency becomes indispensable.

Methodology, evidence and limitations

This is a primary-source market survey and analytical synthesis, current through 2 November 2026. It combines vendor product/technical documentation and corporate announcements; academic papers and benchmark descriptions; FinOps Foundation guidance; and previously published Frontier research. Competing analyst-firm research is not used as evidence. Public vendor statements are treated as claims about the documented offer, not independently audited assurance.

The 56 directory profiles were selected for material enterprise relevance, broad access or modality coverage, differentiated routing/control mechanisms, an established platform position, or a verifiable emerging technical proposition. The adjacent-provider and research lists expand the ecosystem boundary without labeling every inference host, SDK, evaluator or open project an intelligent routing service. This is a comprehensive decision-oriented survey, not a claim of exhaustive coverage of every reseller or new launch worldwide.

The differentiation maps are qualitative analyst classifications. They are not market-share maps, performance rankings, certification results or measurements of customer adoption. Selected vendor icons from Simple Icons are used solely for identification; name labels remain authoritative. Icon availability is not an assessment of vendor importance. [103] Corporate ownership and feature scope can change faster than documentation. Where public pages disagree, the report avoids an unsupported categorical status and identifies the item for diligence.

No new production load test, customer survey, vendor interview program, security penetration test or negotiated-price audit was performed for this report. Academic results retain their original study context; preprints are labeled. Vendor performance and savings claims are not normalized into a cross-vendor league table. The cost-per-outcome scenario and Routing Constitution are original Frontier analytical illustrations, not observed vendor results or legal advice.

Security, residency and contractual compliance depend on configuration, jurisdiction, suppliers and the actual agreement. Obtain appropriate legal/security review for high-consequence deployments. Model counts, list prices and release labels are deliberately not used as stable quality scores. Reconfirm them during procurement and after material service changes.

Rights and attribution

Copyright 2026 Frontier Research. Author: Dion Hinchcliffe. Original Frontier analysis, differentiation maps, architecture and illustrative economics. Third-party trademarks and logos remain the property of their owners and are used for identification, without implying sponsorship or endorsement. External sources are attributed through reader-visible citations and the Sources section. Contact inquiries@thisisfrontier.ai for research briefings and reprints@thisisfrontier.ai for reprints and licensing.

SOURCES / CITATIONS

Trace the argument.

Every source below is linked. Source numbering follows first appearance in the report.

01 **Frontier Research, The Token Crash Is Coming. Your AI Bill Is Going Up. (2 October 2026)**
https://thisisfrontier.ai/research/the-token-crash-is-coming

02 **Frontier Research, The Autonomous Enterprise**
https://thisisfrontier.ai/research/the-autonomous-enterprise

03 **Frontier Research, The Future of Work in the AI Era (16 October 2026)**
https://thisisfrontier.ai/research/the-future-of-work-in-the-ai-era

04 **Ong et al., RouteLLM: Learning to Route LLMs with Preference Data (2024; revised 2025)**
https://arxiv.org/abs/2406.18665

05 **Li et al., LLMRouterBench, Findings of ACL 2026**
https://aclanthology.org/2026.findings-acl.1881/

06 **Stripe, Agreement to acquire OpenRouter (19 August 2026)**
https://stripe.com/newsroom/news/stripe-agrees-to-acquire-openrouter

07 **Palo Alto Networks, Prisma AIRS AI Gateway GA following Portkey acquisition (16 July 2026)**
https://www.paloaltonetworks.com/blog/2026/07/announcing-general-availability-of-prisma-airs-ai-gateway/

08 **DigitalOcean, completed acquisition of Katanemo Labs (2 April 2026)**
https://investors.digitalocean.com/news/news-details/2026/DigitalOcean-Acquires-Katanemo-Labs-to-Accelerate-the-Inference-Cloud-for-the-Agentic-Era/default.aspx

09 **Chen, Zaharia and Zou, FrugalGPT (2023)**
https://arxiv.org/abs/2305.05176

10 **OpenRouter, Provider routing documentation**
https://openrouter.ai/docs/guides/routing/provider-selection

11 **Replicate, predictions and model-version HTTP API**
https://replicate.com/docs/reference/http

12 **Segmind, PixelFlow visual workflows**
https://docs.segmind.com/docs/pixelflow

13 **WaveSpeedAI, unified media API documentation**
https://wavespeed.ai/docs/

14 **LLMGateway, dynamic routes and endpoint scope**
https://docs.llmgateway.io/features/dynamic-routes

15 **OpenRouter, Enterprise capabilities**
https://openrouter.ai/enterprise

16 **Hugging Face, Inference Providers documentation**
https://huggingface.co/docs/inference-providers/index

17 **Requesty, routing and gateway documentation**
https://docs.requesty.ai/

18 **Eden AI, unified model access platform**
https://www.edenai.co/

19 **AI/ML API, multimodal access and JevRouter**
https://aimlapi.com/

20 **Commonstack, UncommonRoute smart routing**
https://docs.commonstack.ai/user-guide/smart-routing

21 **Vercel, AI Gateway**
https://vercel.com/ai-gateway

22 **Cloudflare, AI Gateway documentation**
https://developers.cloudflare.com/ai-gateway/

23 **fal, generative media platform**
https://fal.ai/

24 **Hugging Face, routed requests and provider-key billing**
https://huggingface.co/docs/inference-providers/main/pricing

25 **Palo Alto Networks, Prisma AIRS AI Gateway overview**
https://docs.paloaltonetworks.com/prisma-airs/ai-gateway/ai-gateway-overview

26 **LiteLLM, routing, fallback and deployment controls**
https://docs.litellm.ai/docs/routing

27 **Maxim AI, Bifrost gateway**
https://www.getmaxim.ai/bifrost

28 **TrueFoundry, enterprise AI gateway**
https://www.truefoundry.com/ai-gateway

29 **TensorZero, inference gateway, observability, evaluation and optimization**
https://www.tensorzero.com/docs/

30 **Braintrust, Gateway and observability**
https://www.braintrust.dev/product/observe

31 **Respan, AI Gateway (formerly Keywords AI)**
https://www.respan.ai/platform/ai-gateway

32 **LangDB, model routing documentation**
https://docs.langdb.ai/features/routing/

33 **Orq.ai, AI gateway**
https://orq.ai/platform/ai-gateway

34 **nexos.ai, AI gateway**
https://nexos.ai/ai-gateway/

35 **Kong, AI Gateway architecture and routing algorithms**
https://developer.konghq.com/ai-gateway/architecture/

36 **Tyk, AI Gateway**
https://tyk.io/ai-gateway/

37 **Gravitee, AI Gateway for LLM, MCP and A2A traffic**
https://www.gravitee.io/platform/ai-gateway

38 **WSO2, AI Gateway overview**
https://wso2.com/api-platform/docs/cloud/ai-gateway/overview/

39 **F5, AI Gateway**
https://www.f5.com/products/ai-gateway

40 **Nadir, decision API and optional verified cascade**
https://getnadir.com/

41 **Fireworks Nexus and FireRouter, product and vendor-reported evaluation**
https://fireworks.ai/nexus

42 **Yang et al., TwinRouterBench (May 2026 preprint)**
https://arxiv.org/abs/2605.18859

43 **Li et al., RouterWorld (October 2026 preprint; research performed at Amazon)**
https://arxiv.org/abs/2610.26511

44 **Zhang et al., Security-Conditioned Model-and-Budget Selection / SCRouter (October 2026 preprint)**
https://arxiv.org/abs/2610.18566

45 **Sasidharan et al., Governing Enterprise AI Traffic at Scale: Cosmos LLM Gateway at PayPal (26 October 2026 preprint)**
https://arxiv.org/abs/2610.35509

46 **FinOps Foundation, FinOps for AI: Tools & Services Considerations**
https://www.finops.org/wg/finops-for-ai-tools-services-considerations/

47 **FinOps Foundation, Unit Economics capability**
https://www.finops.org/framework/capabilities/unit-economics/

48 **Maxim AI, Bifrost vendor-run gateway benchmarks**
https://www.getmaxim.ai/bifrost/resources/benchmarks

49 **Not Diamond, Code launch and vendor-reported savings (4 August 2026)**
https://www.notdiamond.ai/blog/not-diamond-code-intelligent-model-routing-for-coding-agents

50 **Not Diamond Code, early-access routing documentation**
https://code.notdiamond.ai/docs/

51 **Snowflake, dynamic routing and internal efficiency evaluations (18 August 2026)**
https://www.snowflake.com/en/news/press-releases/snowflake-unlocks-better-ai-economics-dynamic-model-routing/

52 **LiteLLM, canonical investigation of the v1.107.0 release compromise (October 2026)**
https://docs.litellm.ai/blog/security-incident-oct-21-2026-attack-chain

53 **Frontier Research, Black October: The Enterprise Survival Manual (22 October 2026)**
https://thisisfrontier.ai/research/black-october-the-enterprise-survival-manual

54 **Shafran et al., Rerouting LLM Routers (2025)**
https://arxiv.org/abs/2501.01818

55 **Requesty, approved models and routing terms**
https://www.requesty.ai/terms

56 **ZenMux, model routing documentation**
https://zenmux.ai/docs/guide/advanced/model-routing.html

57 **LLMGateway, routing features**
https://docs.llmgateway.io/features/routing

58 **Routeway, enterprise AI API**
https://routeway.ai/enterprise

59 **Abacus.AI, RouteLLM API**
https://routellm-apis.abacus.ai/

60 **Kie AI, model marketplace API quickstart**
https://docs.kie.ai/market/quickstart

61 **DigitalOcean, Inference Router architecture (2026)**
https://www.digitalocean.com/blog/inference-router-architecture

62 **Martian, gateway documentation**
https://gateway-docs.withmartian.com/

63 **Neutrino AI, multi-model engines**
https://www.neutrinoapp.com/

64 **BlockRun, ClawRouter routing, exclusions and payment documentation**
https://blockrun.ai/docs/products/routing/clawrouter

65 **vLLM Semantic Router, system overview**
https://vllm-sr.ai/docs/overview/semantic-router-overview/

66 **TensorOpera, model routing and agent platform**
https://tensoropera.ai/

67 **Arcee AI, Conductor Auto Mode (older documented model catalog; reconfirm current offer)**
https://docs.arcee.ai/arcee-conductor/features-and-functionality/auto-mode

68 **Aurelio, Semantic Router open-source decision layer**
https://www.aurelio.ai/semantic-router

69 **TensorZero, adaptive A/B experiments**
https://www.tensorzero.com/docs/experimentation/run-adaptive-ab-tests

70 **Helicone, gateway and observability documentation**
https://docs.helicone.ai/

71 **Helicone, corporate/product transition notice**
https://www.helicone.ai/blog/joining-mintlify

72 **Vercel, AI Gateway routing rules (July 2026)**
https://vercel.com/changelog/ai-gateway-routing-rules

73 **Cloudflare, Auto Router beta documentation**
https://developers.cloudflare.com/ai-gateway/features/auto-router/

74 **AssemblyAI, speech-oriented LLM Gateway**
https://www.assemblyai.com/products/llm-gateway

75 **Agent Router maintainers, Envoy AI Gateway becomes Agent Router (9 September 2026)**
https://theagentrouter.ai/blog/envoy-ai-gateway-is-now-agent-router/

76 **Agent Router, project and linked Tetrate hosted offering**
https://theagentrouter.ai/

77 **agentgateway, virtual models and routing**
https://agentgateway.dev/docs/standalone/latest/documentation/llm/virtual-models/

78 **IBM, API Connect AI Gateway on DataPower**
https://www.ibm.com/docs/en/api-connect/software/10.0.x_cd?topic=definitions-using-ai-gateway-support-apis-ai-applications

79 **IBM watsonx.ai, Model Gateway**
https://www.ibm.com/docs/en/watsonx/w-and-w/2.3.x?topic=models-model-gateway

80 **Microsoft, AI gateway in Azure API Management**
https://learn.microsoft.com/en-us/azure/api-management/genai-gateway-capabilities

81 **Amazon Bedrock, intelligent prompt routing**
https://docs.aws.amazon.com/bedrock/latest/userguide/prompt-routing.html

82 **Microsoft Foundry, model router concepts**
https://learn.microsoft.com/en-us/azure/foundry/openai/concepts/model-router

83 **Microsoft Foundry, router release history through October 2026**
https://learn.microsoft.com/en-us/azure/foundry/foundry-models/whats-new-model-router

84 **Google Cloud, build an AI gateway with Apigee**
https://docs.cloud.google.com/apigee/docs/api-platform/fundamentals/build-ai-gateway

85 **Google Cloud API Gateway, Model Garden model routing**
https://docs.cloud.google.com/api-gateway/docs/model-routing-overview

86 **Databricks, Unity AI Gateway model services**
https://docs.databricks.com/gcp/en/ai-gateway/model-services

87 **Snowflake, Cortex AI Gateway**
https://docs.snowflake.com/en/user-guide/snowflake-cortex/cortex-ai-gateway

88 **Alibaba Cloud Model Studio, Intelligent Routing**
https://docs.modelstudio.console.alibabacloud.com/en/model-studio/model-routing

89 **New API, project scope and model/channel routing**
https://docs.newapi.pro/en/docs/guide/wiki/basic-concepts/project-introduction

90 **Together AI, inference platform**
https://www.together.ai/

91 **Fireworks AI, inference and routing platform**
https://fireworks.ai/

92 **DeepInfra, inference platform**
https://deepinfra.com/

93 **Novita AI, inference platform**
https://novita.ai/

94 **SiliconFlow, inference platform**
https://www.siliconflow.com/

95 **Groq, inference platform**
https://groq.com/

96 **Cerebras, inference platform**
https://www.cerebras.ai/

97 **Baseten, inference infrastructure**
https://www.baseten.co/

98 **SambaNova, inference platform**
https://sambanova.ai/

99 **NVIDIA Dynamo, distributed inference documentation**
https://docs.nvidia.com/dynamo/

100 **llm-d, distributed inference project**
https://llm-d.ai/

101 **LLMRouter: A Comprehensive Library for LLM Routing (2026 preprint)**
https://arxiv.org/abs/2608.06867

102 **Unify, current product direction**
https://unify.ai/

103 **Simple Icons, identification icons and trademark attribution**
https://github.com/simple-icons/simple-icons

FRONTIER.

TAKE THE NEXT STEP

Push this further.

Research is useful when it changes a decision. Bring us the question, the constraint, or the argument you want tested.

RESEARCH / BRIEFINGS

inquiries@thisisfrontier.ai

Briefings, advisory work, executive discussions, and research inquiries.

REPRINTS / LICENSING

reprints@thisisfrontier.ai

Reprints, enterprise distribution, excerpts, and licensing.

FRONTIER

www.thisisfrontier.ai

Research, Data Insights, News, and the live Frontier corpus.

All rights reserved. Contact Frontier for redistribution or reprint rights.

SEE FURTHER. MOVE SOONER.