AI Model Routing: Cost and Quality Optimization Guide

IntuitionLabs.ai · Adrien Laurent · 2026-09-05 · ai model routing · llm routing · cost optimization · model routing cost · llm inference cost · ai model comparison · routellm · openrouter · enterprise ai cost · llm evaluation

Executive Summary

Enterprises choosing between deploying a single flagship large language model (LLM) and routing queries dynamically across models of different size face a decision with large, quantifiable stakes. Official API pricing observed on September 5, 2026 shows a consistent multiple between a provider's cheap and flagship tiers: Anthropic's Opus 5 costs 5 times its Haiku 4.5 model on both input and output tokens (www.anthropic.com) (www.anthropic.com) (www.anthropic.com) (www.anthropic.com), while Google's Gemini 3.1 Pro Preview costs roughly 2.7 to 3.2 times its Gemini 3.8 Flash tier (ai.google.dev) (ai.google.dev) (ai.google.dev) (ai.google.dev). That price gap is the raw material every routing policy is trying to capture, and independent academic results show it is capturable: RouteLLM, from researchers associated with LMSYS, reports cutting costs by over 2 times relative to an always-GPT-4 baseline without compromising response quality, and needs only 26% of queries routed to GPT-4 to reach 95% of GPT-4's own MT-Bench performance, falling to 14% with additional training data (lmsys.org) (lmsys.org). FrugalGPT separately reports matching a flagship model's performance with up to 98% cost reduction, while a difficulty-aware Hybrid LLM router and the confidence-based AutoMix system report 40% fewer large-model

calls and over 50% lower computational cost respectively, each at comparable measured quality. Cloud vendors have operationalized similar logic: Amazon Web Services (AWS) states its Bedrock Intelligent Prompt Routing can cut costs by up to 30% without compromising accuracy (aws.amazon.com), and Microsoft's Azure AI Foundry Model Router offers selectable modes that trade a defined 1% to 2% quality gap for the cheapest qualifying model (learn.microsoft.com).

None of this saving is free to realize. Trustworthy routing depends on measuring output quality at the scale of production traffic, and the three dominant methods, crowdsourced pairwise voting (Chatbot Arena has logged over 240,000 votes from about 90,000 users), LLM-as-a-judge scoring (which the MT-Bench paper found can match human-level agreement rates of 81% to 85%), and fixed benchmark suites such as MMLU-Pro, HumanEval, and GPQA, each carry their own cost and bias profile. Paying human experts to grade model output cost $20 per 20 questions, about $35 an hour, in the MT-Bench study, and a separate cost analysis found a dedicated human reviewer of AI output becomes a $5,850 monthly line item at typical review volumes (www.cloudzero.com). Meanwhile the macro backdrop is a paradox: enterprise generative AI spending rose from $11.5 billion in 2024 to $37 billion in 2025 (menlovc.com), even as per-token inference prices for comparable quality fell by roughly a factor of ten every year according to Andreessen Horowitz's analysis (a16z.com) (a16z.com), because usage volume and model escalation are growing faster than unit prices are falling (menlovc.com). Gartner forecast that at least 30% of generative AI projects would be abandoned after proof of concept by the end of 2025 because of poor data quality, inadequate risk controls, escalating costs, or unclear business value (www.gartner.com), and puts full business-model-transformation deployments at $5 million to $20 million (www.gartner.com).

Commercial routing platforms, including OpenRouter, Portkey, Martian, Requesty, and Not Diamond, offer ready-made routing infrastructure spanning hundreds to over a thousand models (openrouter.ai) (portkey.ai) (docs.withmartian.com), but their published cost and quality claims are vendor-reported and not independently verified in the sources reviewed for this report; RouterBench, an academic benchmark built from over 405,000 inference outcomes, exists specifically because vendor figures are not directly comparable. The report concludes that the correct comparison for an enterprise is total cost of ownership, inference price times volume times escalation rate, plus evaluation cost, plus human-review cost for residual errors, rather than a simple comparison of list prices, and that IntuitionLabs' own governance framework of measuring "workflow penetration, time recovered, quality, risk signals, reliability, and support burden before expanding the investment" (intuitionlabs.ai) applies as directly to a routing rollout as to any other AI deployment.

Introduction and Background

Enterprise buyers evaluating generative artificial intelligence (AI) deployments increasingly face a choice that did not exist in the earliest wave of large language model (LLM) adoption: whether to standardize on a single flagship model for every request, or to route requests dynamically across multiple models of different size, cost, and capability. The commercial logic is straightforward. Frontier models command a premium price per token, while smaller models are dramatically cheaper but answer a narrower share of queries at acceptable quality. **AI model routing** is the practice of directing each incoming query, programmatically, to the cheapest model expected to answer it correctly, and escalating to a larger or more capable model only when the smaller model's output is predicted to fall short.

This report examines the economics of that choice: what routing systems exist today, what academic evidence says about the cost and quality tradeoffs they produce, and what it costs an enterprise to measure "quality" well enough to trust a routing decision at all. Menlo Ventures estimates that enterprise generative AI spending rose from $11.5 billion in 2024 to $37 billion in 2025 (menlovc.com). That spending growth occurred even as per-token inference prices for comparable output quality fell by roughly a factor of ten every year, according to venture-capital-firm analysis of public API pricing (a16z.com). Rising aggregate spend alongside falling unit prices is the core paradox this report addresses: usage volume, model proliferation, and escalation to larger models are growing faster than unit prices are falling, which is precisely the condition that makes routing policy an economic decision rather than a purely technical one.

As an adjacent advisory and AI consultancy to the life-sciences sector, IntuitionLabs frames this problem in terms of governed, measured adoption rather than unmanaged access: its AI Acceleration practice is built around **"governed information, specialist implementation, role-based adoption, and measured results in your environment"** (intuitionlabs.ai), a framing consistent with the routing-as-cost-control discipline this report documents, whether the underlying workflow calls one model or several. The remainder of this report is organized as a reproducible method: how routing quality and cost are measured, how routing architectures make the escalation decision, what commercial and cloud-native platforms currently offer, what the peer-reviewed literature has quantified, and what the resulting total cost of ownership looks like once human review and error correction are included alongside inference spend. All prices, benchmark figures, and platform claims below are dated to their observation date because this is a fast-moving market; a routing decision made on today's price sheet may not hold in six months.

Methodology: How Routing Quality and Cost Are Measured

A routing policy is only as good as the yardstick used to judge whether the cheaper model's answer was "good enough." Three families of measurement dominate current practice, each with a different cost profile and a different failure mode.

Pairwise human preference at scale. LMSYS's Chatbot Arena collects anonymous, crowdsourced pairwise comparisons between two models' answers to the same prompt, then converts the win/loss/tie record into a ranking. As of the arena's original evaluation paper, the system had accumulated **over 240,000 votes from about 90,000 users in over 100 different languages** (arxiv.org), averaging roughly **8,000 votes collected for each model** (arxiv.org). By December 2023 the platform had passed **130,000 total votes** ranking more than 40 deployed models (lmsys.org), and it has since switched its ranking math from online Elo scoring to a Bradley-Terry maximum-likelihood model, in part because Elo's outcome depends on the order battles are played, which produces unstable rankings at scale (lmsys.org); ties are handled by counting each tied battle as **half a win and half a loss** for both models (lmsys.org). To validate that anonymous crowd votes are trustworthy, the Arena team had graduate-student experts manually re-judge a sample of battles, a process that took **on average 3 to 5 minutes per data point** (arxiv.org) and found crowd-versus-expert agreement of **72% to 83%** (arxiv.org), comparable to the agreement rate between two experts on the same battles.

LLM-as-a-judge. Because human pairwise voting does not scale to every routing decision an enterprise makes in production, a second family of methods uses a strong LLM itself to grade a weaker model's output. The foundational MT-Bench and Chatbot Arena paper found that **strong LLM judges like GPT-4 can match both controlled and crowdsourced human preferences** (arxiv.org), with GPT-4-to-human agreement (85% in the

paper's non-tie setup) actually exceeding **agreement among humans (81%)** on the same 80-question, multi-turn benchmark set, which drew on **3,000 expert votes and 30,000 conversations with human preferences** made publicly available (arxiv.org). The same paper is explicit that LLM judges are not a free lunch: it documents **position, verbosity, and self-enhancement biases, as well as limited reasoning ability** as systematic failure modes that must be corrected for (arxiv.org). Human expert grading itself is not free either; the paper's authors paid expert human evaluators **$20 for judging 20 questions, which corresponds to an hourly rate of around $35** (arxiv.org), a useful anchor for the labor cost of any routing policy that falls back to human review.

Standardized benchmark suites. A third measurement layer sits below both of the above: fixed-answer academic benchmarks that can be scored automatically and cheaply, useful for coarse routing decisions rather than fine-grained production judging. **MMLU** (Massive Multitask Language Understanding) tests accuracy across **57 tasks including elementary mathematics, US history, computer science, law, and more** (arxiv.org); its harder successor, **MMLU-Pro**, adds more distractor answers and reasoning-heavy questions, **causing a significant drop in accuracy of 16% to 33% compared to MMLU** for the same models (arxiv.org), which matters for routing because a policy tuned against the easier benchmark will overestimate a cheap model's real-world reliability. **HumanEval** measures functional code correctness; in its introducing paper, the underlying Codex model **solved 28.8% of the problems, while GPT-3 solved 0% and GPT-J solved 11.4%** in a single sample, rising to roughly 70% when 100 completions were sampled per problem (arxiv.org). **GPQA**, a graduate-level "Google-proof" multiple-choice benchmark, found that PhD-level subject experts reach **65% accuracy** (arxiv.org) while skilled non-expert validators with unrestricted web access reach far lower scores despite **spending on average over 30 minutes with unrestricted access to the web** per question (arxiv.org), illustrating how expensive rigorous manual grading becomes as question difficulty rises, and why routing systems increasingly substitute automated judges for that manual effort wherever the stakes allow it.

Routing Architectures: From Static Deployment to Dynamic Escalation

A **single-model deployment** sends every request, regardless of complexity, to one model, typically the flagship the team trusts most. It is simple to operate and audit, but it means paying flagship prices for queries a far cheaper model could have answered correctly. A **routing (or cascade) deployment** instead scores each incoming request, either before or after a first pass, and decides whether the cheap path is sufficient or whether the request must escalate to a larger model.

Two production implementations illustrate how cloud vendors have operationalized this decision. **Amazon Web Services' (AWS) Bedrock Intelligent Prompt Routing** states that it **"can reduce costs by up to 30% without compromising on accuracy"** (aws.amazon.com) by predicting, for each prompt, which model within a single model family will give the desired response at the lowest cost; it currently supports routing only **"any two models from the same family with Anthropic"**, Meta Llama, or Amazon Nova model families (aws.amazon.com), meaning cross-vendor routing is out of scope for this particular service. AWS documentation defines the escalation trigger as a **routing criteria threshold**: it explains that **"a response quality difference of 10% means that, say the response quality of the fallback model"** is used as the baseline the router must beat by that margin before it will switch away from the designated fallback model (docs.aws.amazon.com). **Microsoft's Model Router for Azure AI Foundry** takes a similar approach but is deployed like a single model from the caller's perspective; Microsoft states it **"delivers high performance while saving on costs, reducing latencies, and increasing responsiveness, while maintaining comparable quality"** (learn.microsoft.com). Its three selectable modes make

the cost-quality tradeoff explicit: a "Balanced" mode **"considers all underlying models within a small quality range (for example, 1% to 2% compared with the highest-quality model"** and picks the cheapest one in that band, while a looser "Cost" mode widens the acceptable quality gap further before defaulting to the most expensive option (learn.microsoft.com); automatic failover to a different model is enabled by default in case the first-choice model has a transient outage (learn.microsoft.com). Google has pursued the same idea at the API layer: its April 2025 announcement of a Vertex AI Model Optimizer describes a feature designed **"to automatically generate the highest quality response for each prompt based on your desired balance of quality and cost"** (cloud.google.com), an explicit, customer-tunable dial between the two variables this report is about.

Academic work has formalized the escalation decision beyond these vendor implementations. **RouteLLM**, from researchers associated with LMSYS, reports that its trained routers **reduce costs by over 2 times in certain cases, without compromising the quality of responses** relative to always calling the frontier model (arxiv.org). Its most literature-cited result is architectural: on the MT-Bench evaluation, a matrix-factorization router can **achieve 95% of GPT-4 performance using only 26% of queries routed to GPT-4 calls** (lmsys.org), and when the router is trained on additional LLM-judge-labeled data, that figure falls further, to a router that **needs the number of GPT-4 calls required to achieve 95% GPT-4 performance further halved, to 14% of total calls** (lmsys.org). **FrugalGPT**, an earlier cascade-based approach, demonstrated that a learned cascade **can match the performance of the best individual LLM (e.g. GPT-4) with up to 98% cost reduction** (arxiv.org), by first trying cheap models and only escalating when a lightweight scoring function flags the cheap answer as unreliable. A separate difficulty-aware router described in a 2024 paper found it could **"make up to 40% fewer calls to the large model, with no drop in response quality"** (arxiv.org), by classifying queries as easy enough for a small model before dispatch rather than after the fact. **AutoMix** takes yet another approach, using a small model's own self-verified confidence plus a decision-process router to decide when to escalate, **reducing computational cost by over 50% for comparable performance** (arxiv.org). Read together, these four independent academic results converge on the same order of magnitude: a well-tuned router can retain 95 percent or more of flagship-model quality while cutting between roughly 40% and 98% of the cost a single-model deployment would have paid, with the size of the saving depending heavily on how tolerant the underlying task is of occasional escalation misses.

Commercial and Cloud-Native Routing Platforms

Enterprises that do not want to build and train a custom router can buy one. The market spans multi-model API gateways, cloud-vendor-native routers, and standalone routing-as-a-service products; all of the figures below are the vendors' own claims about their own products, not independently verified by this report, and are labeled as such.

OpenRouter operates as a multi-model gateway, advertising **"300T+ Monthly Tokens 10M+ Global Users 80+ Providers 500+ Models"** on its own homepage (openrouter.ai). Its Auto Router works by having **"a fast, lightweight classifier assign each prompt one of ~30 fine-grained task types"** and then ranking eligible models by aggregate real-world spend share over a trailing window, rather than a fixed benchmark score; the company states there is **"no additional fee"** for Auto Router beyond the standard per-token rate of whichever model is ultimately selected (openrouter.ai). Its companion Pareto Router lets a caller specify a single coding-quality floor, expressed as **"a `min_coding_score` preference between 0 and 1"** (openrouter.ai), and picks the cheapest model that clears that bar. A separate Model Fallbacks feature retries a request against the next listed model automatically if the primary model errors due to downtime, rate limiting, or context-length failures (openrouter.ai).

Portkey markets a comparable AI gateway, stating it gives access to **"1,600+ LLMs via a unified API"** (portkey.ai) and describing itself as **"the world's fastest AI Gateway with advanced routing & integrated Guardrails"** serving **"3000+ GenAI teams"** (portkey.ai) (portkey.ai). Its Conditional Routing feature can, for example, **"route it to the cheapest model"** for a non-production testing environment while sending production traffic elsewhere (portkey.ai), and its Load Balancing feature is designed to **"distribute traffic across multiple LLMs to prevent any single provider from becoming a bottleneck"** (portkey.ai). **Martian**, positioned primarily as an AI interpretability research lab (docs.withmartian.com), continues to operate a commercial Gateway product that provides **"unified access to 200+ AI models through a single API"** (docs.withmartian.com) with OpenAI- and Anthropic-compatible request formats (docs.withmartian.com). **Requesty**, another gateway product, advertises **"smart routing to cheaper equivalent models, caching, automatic fallback from expensive providers"** across 600-plus models (requesty.ai), and prices itself with a flat markup: **"You pay for what your application spends, plus 5%."** on top of the underlying model's own cost (requesty.ai). **Not Diamond**, a standalone router, states that its product **"intelligently predicts which model to use for each input, reducing costs while maintaining accuracy"** (notdiamond.ai); its own published customer testimonials include a claim from one customer's Head of AI Labs that Not Diamond **"increased the average accuracy by 39%"** across their internal benchmarks (notdiamond.ai), and a claim attributed to another customer that switching **"significantly reduced our inference costs while also driving improvements in output quality"** (notdiamond.ai); these are vendor-published testimonials, not independently audited results, and this report treats them accordingly.

Table 1 below summarizes how these platforms differ in scope and mechanism.

Platform	Routing approach	Model scope (as stated)	Published cost/quality claim
OpenRouter	Task-classifier Auto Router + Pareto Router (quality-floor routing)	500+ models, 80+ providers (openrouter.ai)	No published headline savings figure; no extra fee beyond selected model's own price (openrouter.ai)
Portkey	Conditional routing, load balancing, gateway-level rules	1,600+ LLMs (portkey.ai)	Markets itself as fastest gateway; 3,000+ teams (portkey.ai)
AWS Bedrock Intelligent Prompt Routing	Same-family threshold routing (quality-difference criterion vs. fallback model)	Two models within one family (Anthropic, Llama, or Nova) (aws.amazon.com)	Up to 30% cost reduction "without compromising on accuracy" (aws.amazon.com)
Azure AI Foundry Model Router	Selectable Balanced/Cost/Quality modes with a stated quality-gap band	Underlying model set not itemized on the page cited	Balanced mode targets 1-2% quality gap vs. top model before choosing cheapest (learn.microsoft.com)
Martian Gateway	Single OpenAI/Anthropic-compatible API across many providers	200+ models (docs.withmartian.com)	No headline savings figure published on pages reviewed
Requesty	Smart routing to cheaper equivalent models plus caching and fallback	600+ models (vendor-stated)	Flat 5% markup on underlying spend, 0% with a bring-your-own-key setup (requesty.ai)
Not Diamond	Predictive per-input model selection	Not itemized on pages reviewed	Customer-reported 39% accuracy increase on internal benchmarks (notdiamond.ai)

None of the figures in the right-hand column of Table 1 have been independently benchmarked by a third party in the sources reviewed for this report; they are the vendors' own disclosures, and an enterprise evaluating any of these products should expect to validate the claim against its own workload before committing budget to it. **RouterBench**, an academic benchmark built specifically to standardize evaluation of multi-model routing systems, was constructed from **"a comprehensive dataset comprising over 405,000 inference outcomes from representative LLMs"** (arxiv.org) precisely because vendor-reported figures are not directly comparable to one another without a common test set.

Academic Evidence on Cost-Quality Tradeoffs

Where the previous section reported vendor claims, this section reports figures published in peer-reviewed or preprint academic work, where the authors publish enough of their method that a third party could in principle reproduce the result. **RouteLLM**'s authors report that, evaluated against commercial routing products from Martian and Unify AI on the MT-Bench benchmark, their open router achieved **"the same performance as these commercial routers while being over 40% cheaper"** (lmsys.org), and separately reported per-benchmark cost reductions of **"over 85% on MT Bench, 45% on MMLU, and 35% on GSM8K"** while holding to 95% of GPT-4's own performance level on each (lmsys.org). The spread across those three benchmarks, from 35% to 85% in achievable savings, is itself an important data point: a routing policy's savings are benchmark- and task-dependent, and a single blended "expected savings" number should be treated skeptically without knowing the query mix it was measured on.

Table 2 compares the four independently published academic methods discussed in this report.

Method	Mechanism	Reported cost reduction	Reported quality retention
RouteLLM (lmsys.org)	Learned router trained on human and LLM-judge preference data, calls GPT-4 only when needed	Over 2x vs. always-GPT-4 baseline (arxiv.org); as little as 14% GPT-4 calls with augmented training data (lmsys.org)	95% of GPT-4 performance on MT-Bench
FrugalGPT (arxiv.org)	Learned cascade: tries cheap models first, escalates on a scoring function's signal	Up to 98% cost reduction	Matches best individual LLM's performance
Hybrid LLM Routing (arxiv.org)	Difficulty-aware router classifies queries as easy/hard before dispatch	Up to 40% fewer large-model calls	No drop in response quality (as reported)
AutoMix (arxiv.org)	Small model self-verifies confidence; POMDP-based router decides escalation	Over 50% cost reduction	Comparable performance (as reported)

The consistent pattern across all four rows is that escalation is selective rather than binary: none of these methods send the majority of traffic to the expensive model, yet all four report retaining most or all of the flagship model's measured quality on their respective test sets. The methods differ mainly in when the escalation decision is made (before generation, as in the difficulty-aware router, versus after a cheap draft answer exists, as in FrugalGPT and AutoMix) and in what evidence triggers escalation (a learned preference classifier in RouteLLM, a lightweight

scoring function in FrugalGPT, self-verified confidence in AutoMix). Because these are academic results measured on specific published benchmarks (MT-Bench, MMLU, GSM8K, and the authors' own held-out test sets), the percentage figures should not be read as universal constants; an enterprise's own query distribution, especially one skewed toward domain-specific tasks not represented in these benchmarks, will produce a different achievable savings-versus-quality curve, which is precisely why RouterBench's 405,000-outcome dataset exists as a more standardized, though still benchmark-bound, comparison point (arxiv.org).

Data Analysis and Evidence

The routing decision is ultimately a comparison of per-token prices between a cheap model and a flagship model, multiplied by the fraction of queries a routing policy is willing to escalate, plus whatever it costs to catch and correct the queries the router gets wrong. The resulting blended cost per query, not the sticker price of either model alone, is the figure that should drive a build-or-buy routing decision. This section lays out each of those three inputs with current, dated figures.

Per-token pricing. *Table 3* lists official, directly observed API prices (all as of September 5, 2026) for a "cheap" and a "flagship" tier model from four major providers.

Provider	Model (tier)	Price basis / context	Input price per 1M tokens	Output price per 1M tokens
OpenAI	gpt-5.6-sol (flagship)	Text tokens; per 1M tokens	$4.00 (developers.openai.com)	$20.00 (developers.openai.com)
OpenAI	gpt-5.6-luna (cheap)	Text tokens; per 1M tokens	$0.20 (developers.openai.com)	$1.20 (developers.openai.com)
Anthropic	Opus 5 (flagship)	API usage pricing	$5 (anthropic.com)	$25 (anthropic.com)
Anthropic	Haiku 4.5 (cheap)	API usage pricing	$1 (anthropic.com)	$5 (anthropic.com)
Google	Gemini 3.1 Pro Preview (flagship, <=200k context)	Paid tier; <=200k-token prompts	$2.00 (ai.google.dev)	$12.00 (ai.google.dev)
Google	Gemini 3.8 Flash (cheap, promotional through Dec. 2026)	Paid tier; promotional through Dec. 2026	$0.75 (ai.google.dev)	$3.75 (ai.google.dev)
Mistral AI	Mistral Large 3 (flagship)	API usage pricing	$0.5 (mistral.ai)	$1.5 (mistral.ai)
Mistral AI	Mistral Small 4 (cheap)	API usage pricing	$0.15 (mistral.ai)	$0.6 (mistral.ai)

The spread inside a single provider's lineup is one input to the economic argument for routing; Table 3 is a dated list-price snapshot, not a normalized cross-provider comparison, because the listed service tiers and context brackets differ. Within Anthropic's lineup alone, Opus 5 is priced at 5 times Haiku 4.5's input rate and 5 times its output rate (www.anthropic.com) (www.anthropic.com); within Google's lineup, Gemini 3.1 Pro Preview costs roughly 2.7 times Gemini 3.8 Flash on input and 3.2 times on output (ai.google.dev) (ai.google.dev). A workload that can safely send even a modest majority of its queries to the cheap tier, escalating only the fraction the router flags as uncertain, captures most of that multiple as savings, which is exactly the mechanism the academic routers in the previous section report exploiting.

Spend and cost-decline trends. Independent of any specific routing policy, two macro trends define the environment routing decisions are made in. First, aggregate enterprise generative AI spend has grown sharply: from **$11.5 billion in 2024** to **$37 billion in 2025** (menlovc.com), according to venture-capital firm Menlo Ventures' enterprise surveys. Gartner's independent forecast put total worldwide GenAI spending, a broader category than Menlo's enterprise-survey figure, at **$644 billion in 2025** (www.gartner.com). Second, and simultaneously, per-token inference prices for comparable quality have fallen steeply: analysis published by venture-capital firm Andreessen Horowitz (a16z) found that **the cost is decreasing by 10x every year** (a16z.com) for models of equivalent measured quality, with the cost of an MMLU-42-quality model falling by **a factor of 1,000 in 3 years** (a16z.com) and GPT-4-class quality falling **by about a factor of 62** since March 2023 (a16z.com). Menlo's own 2025 report reconciles the two trends directly: **"net spend on generative AI continues to rise despite falling costs of inference"** (menlovc.com), because usage volume, model count, and the share of queries escalated to larger models are all growing faster than the unit price is shrinking.

Implementation and human-review cost. Falling model prices do not mean falling project cost. Gartner has stated that generative-AI business-model-transformation deployments carry **"significant costs, ranging from $5 million to $20 million"** (www.gartner.com), and separately forecast that **at least 30% of generative AI projects would be abandoned after proof of concept** by the end of 2025 (www.gartner.com), citing high failure rates in proof-of-concept work among the causes (www.gartner.com). Menlo's own enterprise survey found **implementation costs, cited in 26% of failed pilots**, catch enterprise buyers off guard more than any other single factor (menlovc.com), even though **just 1%** of the enterprise leaders it surveyed named price itself as their primary evaluation concern (menlovc.com), a distinction between the price of a model call and the cost of deploying it correctly that a routing policy's savings calculation must not conflate. On the human-review side, cost analysis of AI-assisted workflows found that adding a dedicated human reviewer of AI output **"turns that reviewer into a $5,850 monthly line item"** (www.cloudzero.com) at typical review volumes, and framed the underlying decision rule plainly: **"oversight pays when the cost of the human check is less than the expected cost of the error it prevents"** (www.cloudzero.com). That rule applies equally to a routing policy's escalation threshold: routing to a larger model is worth its incremental price only when the larger model's incremental accuracy is worth more than the price difference, the same logic AWS's own routing-criteria documentation encodes as a configurable quality-difference threshold (docs.aws.amazon.com). Academic evaluation cost provides a second anchor point: paying human experts to judge model output, as in the MT-Bench study, cost **$20 for judging 20 questions, or about $35 per hour** (arxiv.org), a per-item labor cost against which any automated LLM-judge escalation policy should be benchmarked before an enterprise decides whether automated grading is cheap enough to run on every routed query or only on a sampled audit trail.

Case Studies and Real-World Examples

AWS Bedrock Intelligent Prompt Routing (documented product configuration). AWS's own documentation describes prompt routers as configurable only within a single model family, giving Anthropic, Meta Llama, and Amazon Nova families as the supported options at time of writing (aws.amazon.com), and lets the operator set a routing-criteria threshold that defines how much better an upgraded model's predicted response quality must be, relative to a nominated fallback model, before the router will pay the higher price (docs.aws.amazon.com). This is a documented product configuration rather than a named customer deployment, but it illustrates a full routing policy expressed as three parameters: a model family, a fallback model, and a quality-difference threshold, which is close to the minimum configuration an enterprise needs to operationalize routing without training a custom classifier.

OpenRouter's Auto Router (documented product mechanism). OpenRouter's own description of its default routing mechanism uses live aggregate spend share across roughly 30 task categories, rather than a static benchmark score, to rank eligible models for an incoming prompt (openrouter.ai), an approach that updates its notion of "the best model for this task type" continuously as usage patterns and model releases change, instead of requiring a manual re-tuning cycle every time a new model is released.

(Hypothetical Example) A support-ticket triage router. Consider an enterprise routing customer-support tickets, where a cheap model such as Mistral Small 4 ($0.15/$0.60 per 1M tokens, input/output) (mistral.ai) (mistral.ai) handles routine, templated requests and escalates only ambiguous or high-value tickets to a flagship model such as Mistral Large 3 ($0.5/$1.5 per 1M tokens) (mistral.ai) (mistral.ai). If 80% of tickets are safely resolved by the cheap model and 20% escalate, the blended input cost is $0.22 per 1M input tokens (0.8 × $0.15 + 0.2 × $0.50), or 56% lower than sending every ticket to Mistral Large 3 at $0.50 per 1M input tokens. This example is illustrative arithmetic based on the publicly listed prices above, not a measured production deployment, and actual savings depend entirely on how accurately the router classifies "routine" versus "ambiguous" tickets, which is precisely the accuracy question the Methodology section's quality-measurement tools exist to answer.

Implications and Future Directions

Three implications follow from the evidence assembled above. First, routing is not a substitute for measuring quality, it is a bet on how well an enterprise can measure quality cheaply enough to make the bet pay off. Every academic result in this report (RouteLLM, FrugalGPT, Hybrid LLM Routing, AutoMix) depends on some form of quality-prediction signal, whether a learned preference model or a self-verification score, and every commercial platform's advertised savings depends on the accuracy of its own internal classifier. An enterprise adopting any of these products inherits that classifier's error rate as its own false-acceptance risk: a query wrongly judged "safe for the cheap model" produces a wrong answer at the cheap model's low price, which is a different failure mode from simply paying more for a flagship answer, and one that a pure cost-per-query metric will not surface. Second, the falling-price trend documented by a16z's LLMflation analysis (a16z.com) (a16z.com) does not eliminate the case for routing; if anything, it raises the bar for what counts as "flagship," since today's cheap tier increasingly matches yesterday's flagship quality, which means a routing policy tuned six months ago may already be escalating too conservatively. Third, the total cost of a routing deployment is not just inference spend. Gartner's forecast that at least 30% of generative AI projects would be abandoned after proof of concept by the end of 2025 because of poor data quality, inadequate risk controls, escalating costs, or unclear business value (www.gartner.com) (www.gartner.com), suggests that the integration, evaluation-harness, and human-review costs surrounding a router matter as much as the per-token price difference the router is designed to exploit.

For organizations approaching this as a governance question rather than a pure engineering one, the discipline of instrumenting a rollout before scaling it, tracking, in the words of one adjacent life-sciences AI advisory's own framework, **"workflow penetration, time recovered, quality, risk signals, reliability, and support burden before expanding the investment"** (intuitionlabs.ai), applies as directly to a model-routing rollout as to any other AI workflow: a routing policy should be judged not only on its projected cost savings but on the same evidence trail an enterprise would demand before expanding any other AI investment. Looking forward, the emergence of standardized cross-vendor benchmarks such as RouterBench (arxiv.org), built specifically because vendor-reported savings figures are not comparable to one another, suggests the market is moving toward independent verification

of routing claims, though as of this report's publication date most published cost-and-quality figures for both academic and commercial routers remain single-source and benchmark-specific rather than independently replicated across providers.

Conclusion

The evidence assembled in this report supports three conclusions about the economics of AI model routing. First, the price gap between a provider's cheap and flagship models can be material, but the displayed multiple varies by provider and quoted configuration; Table 3 should therefore be read as provider-specific list-price evidence rather than a universal three-to-five-times rule. Second, independent academic results converge on the finding that well-tuned routers can retain the large majority of flagship-model quality while escalating only a minority, sometimes a small minority, of queries to the expensive model, though the precise achievable ratio is benchmark- and workload-specific rather than a universal constant. Third, none of that potential saving is free to realize: building or buying a router, measuring its quality reliably enough to trust its escalation decisions, and paying for human review of the cases it gets wrong all carry their own costs, and Gartner's forecast cited poor data quality, inadequate risk controls, escalating costs, and unclear business value as the stated reasons that generative AI projects might be abandoned after proof of concept. An enterprise deciding between a single flagship model and a routed multi-model deployment should therefore treat the decision as a total-cost-of-ownership comparison, inference price times volume times escalation rate, plus evaluation cost, plus human-review cost for the residual error rate, rather than as a simple comparison of list prices. The methodology, architectures, platforms, and figures documented above are intended to give a reader the components needed to build that comparison for their own workload, reproducibly and with dated, sourced inputs, rather than to declare a single winning policy that applies universally.

Frequently Asked Questions (FAQs)

What is AI model routing?

AI model routing is the practice of directing each incoming query to whichever available model is predicted to answer it correctly at the lowest cost, escalating to a larger or more expensive model only when a cheaper model's output is judged insufficient, using either a pre-generation classifier or a post-generation confidence or scoring signal (docs.aws.amazon.com).

How much can model routing reduce inference costs?

Published academic results range from roughly 40% fewer large-model calls with no measured quality drop, to over 50% lower computational cost for comparable performance, to as much as 98% cost reduction while matching a flagship model's benchmark performance; commercial vendors separately claim up to 30% savings for same-family routing (aws.amazon.com), and an open router has been reported to match commercial routing products' performance while being over 40% cheaper (lmsys.org). None of these percentages transfer directly to an arbitrary enterprise workload; they are specific to the benchmark or product each was measured on.

How is LLM output quality measured at scale?

Three methods dominate: crowdsourced pairwise human voting (as in Chatbot Arena, with agreement rates of 72% to 83% between crowd and expert raters), LLM-as-a-judge scoring (with GPT-4-level judges reaching human-comparable agreement rates in controlled studies), and fixed-answer benchmark suites such as MMLU-Pro, HumanEval, and GPQA that trade nuance for automatic, cheap scoring; Chatbot Arena itself has moved from Elo scoring to a Bradley-Terry ranking model for statistical stability at scale (lmsys.org).

Is multi-model routing better than a single-model deployment for every use case?

Not necessarily. A single-model deployment is simpler to audit and has no classifier-error risk of its own, and Gartner forecast that at least 30% of generative AI projects would be abandoned after proof of concept by the end of 2025 because of poor data quality, inadequate risk controls, escalating costs, or unclear business value (www.gartner.com) (menlovc.com); routing is worth its added complexity primarily where query volume and query-difficulty variance are both high enough that the blended savings from Table 3's price gaps exceed the cost of building and maintaining the router itself.

What is an escalation rate in the context of AI model routing?

It is the share of queries a router forwards to the more expensive model rather than resolving with the cheaper one; RouteLLM's published results describe routers that escalate as few as 14% to 26% of queries to GPT-4 while retaining 95% of GPT-4's own benchmark performance (lmsys.org) (lmsys.org), though the achievable escalation rate for any given accuracy target is benchmark-specific, as the 35%-to-85% cost-reduction spread across MT-Bench, MMLU, and GSM8K in this report's Academic Evidence section shows (lmsys.org).
