Ricardo: Outcome-Grounded Routing for Multimodal LLM Agents

Theory, a Flywheel Simulation, and Evidence from Owned Traffic

Kacper Wikieł[1]

[1]Fabryka her.fabryka.ai

June 9, 2026

Abstract

The price–quality spread in the language-model market is, by several public measures, the widest in the history of commodity software: equivalent capability has fallen $\sim 700\times$ in cost since 2023 while the *cost* of frontier-level capability has risen $\sim 300\times$. Choosing *which* model runs a task is therefore the dominant lever on the economics of an AI system—yet production routers sort by price and availability, not by *outcome*. We argue that a router is a commodity; what is defensible is the data that feeds it. We present **Ricardo**, an outcome-grounded routing layer with three ingredients: (i) a price-theoretic formulation in which model selection is Ricardian comparative advantage under one dial λ, for which we prove a specialization-gain result, a switching-threshold/frontier characterization, and a market-clearing (Walrasian/Hayekian) interpretation of λ; (ii) a reward hierarchy topped by *execution outcome*—did the work succeed—rather than an LLM judge; and (iii) a closed loop on *owned* agent traffic. A reproducible flywheel simulation shows shadow-A/B with non-inferiority promotion converging to within $\sim 4\%$ of oracle value (at $\approx 66\%$ bucket accuracy—residual errors land on near-tied buckets) and tracing the achievable cost–quality frontier as λ varies. Small-n case studies on real production traffic *illustrate* the mechanism—an LLM judge scoring agentic web tasks at 0.50 against a parameterised execution outcome of 0.09, and a frontier model winning 5/8 web escalations (illustrative, not significant at n=8). A difficulty-tiered, test-graded coding benchmark then supplies powered evidence that capability separates only above a difficulty threshold, where the cheapest *correct* model is neither the free local one nor the frontier. We are explicit throughout about what is demonstrated (mechanism) versus measured. Code is open source.

1 Introduction

A modern AI application is an orchestra of model calls—code, web, documents, speech, vision—each servable by any of several hundred models spanning four orders of magnitude in price. The *routing* question dominates both quality and spend. Over 577 days of OpenRouter statistics—the largest *public* cross-vendor signal, OpenRouter's reported ~ 25 trillion tokens/week across 750 models—cheap open-weight models hold $\approx 64\%$ of *volume* while Western frontier models keep most *revenue*, and like-for-like open tokens price $\sim 9.6\times$ cheaper.[1] The market has split into a cheap-volume tier and an expensive-frontier tier; routing is the act of making that split *deliberately, per task*. We make three departures from prior routers [1, 2, 3]: comparative advantage over absolute quality (§4); measured outcome over predicted quality (§6); and a flywheel on owned traffic over a cold start.

[1] ort.fabryka.ai, the authors' market tracker.

2 Related Work

Routing and cascades. RouteLLM [1] trains classifiers on Chatbot Arena [6] preferences; FrugalGPT [2] cascades cheap-to-dear with a learned acceptance test; commercial gateways auto-route. These optimise cost given *predicted* quality. We optimise a value functional over *measured* outcomes and read per-task escalation as market clearing [17, 16]. **Evaluation.** LLM-as-judge [4] is scalable but biased [5]; execution benchmarks—SWE-bench [7], τ-bench [8], WebArena [9], GAIA [10]—score functional correctness, which we adopt as the reward and extend with a record–replay construction. **Theory.** We borrow Pareto efficiency [12], Hayek's price-as-information [13], Arrow aggregation [14], and the Shapley value [15] for trajectory credit.

3 Preliminaries

Definition 1 (Routing instance)**.** A routing instance is a tuple (T, M, q, c) with task buckets $t \in T$, models $m \in M$, quality $q : M \times T \to [0, 1]$, and per-call cost $c : M \to \mathbb{R}_{\geq 0}$. For a price $\lambda \geq 0$, the *value* of model m on task t, and the route the router selects, are

$$\mathrm{val}_\lambda(m, t) \;=\; q(m, t) \;-\; \lambda\, c(m), \qquad\qquad \pi_\lambda(t) \;=\; \arg\max_{m \in M} \mathrm{val}_\lambda(m, t). \tag{1}$$

We say m is *Pareto-dominated* on t if $\exists\, m'$ with $q(m', t) \geq q(m, t)$ and $c(m') \leq c(m)$, one strict.

λ is the only knob: it is the number of quality points one is willing to pay per dollar of inference. Large λ is cost-thrifty; small λ is quality-hungry.

4 Theory

4.1 Comparative advantage: the dominated model is still used

Proposition 1 (Specialization gain)**.** *Let A absolutely dominate B in quality, $q(A, t) \geq q(B, t)\ \forall t$, but be dearer, $c(B) < c(A)$. Then for every bucket t with $q(A, t) - q(B, t) < \lambda\,(c(A) - c(B))$ the value-optimal route selects the* dominated *model B; and if at least one such bucket exists, per-bucket routing yields strictly greater total value than committing to A everywhere.*

Proof. $\mathrm{val}_\lambda(B, t) - \mathrm{val}_\lambda(A, t) = \lambda(c(A) - c(B)) - (q(A, t) - q(B, t))$, which is > 0 exactly when $q(A, t) - q(B, t) < \lambda(c(A) - c(B))$. Total value of routing is $\sum_t \max_m \mathrm{val}_\lambda(m, t) \geq \sum_t \mathrm{val}_\lambda(A, t)$ with strict inequality at any bucket where B wins. □ □

This is Ricardo's comparative advantage: an *absolutely worse* model is optimally used where its cost edge outweighs its quality deficit. Our coding experiment (§8) is a literal instance—a \$0.00003 model is chosen over a $\sim$127$\times$ dearer one (per output token) at equal correctness.

4.2 The switching threshold and why dominated models vanish

Proposition 2 (Threshold and frontier)**.** *For two models A, B on t with $c(A) \neq c(B)$, π_λ switches at $\lambda^\star = \dfrac{q(A, t) - q(B, t)}{c(A) - c(B)}$. Over all $\lambda \geq 0$, the models selected for t are exactly the* upper-left convex hull *of $\{(c(m), q(m, t))\}_m$; any dominated model—and any Pareto-efficient model in a non-convex dent of the frontier—is selected for* no $\lambda \geq 0$ *(the weighted-sum blind spot).*

Proof. $\mathrm{val}_\lambda(m,t)$ is affine in λ with intercept $q(m,t)$ and slope $-c(m)$. The pointwise maximum of affine functions is piecewise linear and convex in λ; its active pieces are the vertices of the upper-left hull of the points $(c(m), q(m,t))$, i.e. the convex (upper-left) hull of the achievable set. A dominated point lies strictly below this upper envelope for every $\lambda \geq 0$. □ □

4.3 λ is a market-clearing price (Walras, Hayek)

Proposition 3 (Budget allocation is market clearing)**.** *Consider* $\max_\pi \sum_t q(\pi(t),t)$ *subject to* $\sum_t c(\pi(t)) \leq \mathcal{B}$*. Its Lagrangian relaxation decouples per bucket: for multiplier* $\mu \geq 0$*,* $\pi_\mu(t) = \arg\max_m q(m,t) - \mu\, c(m)$*. The least* $\mu^\star$ *with* $C(\mu^\star) \leq \mathcal{B}$ *solves the* fractional *relaxation, and* $\mu^\star$ *is the shadow price of an inference dollar. Hence the dial* λ *is an* endogenous *clearing price: each bucket is coordinated independently through it, exactly Hayek's price signal, and the allocation is Walrasian.*

Proof idea. Separability of the objective and constraint over buckets gives the decoupled argmax under any fixed μ; total cost $C(\mu) = \sum_t c(\pi_\mu(t))$ is nonincreasing in μ, so a 1-D search recovers $\mu^\star$. Since each bucket picks one model, $C(\mu)$ is a step function and generically no μ binds the budget *exactly*; the integer optimum is then attained by *randomised* (time-shared) routing between the two models bracketing $\mathcal{B}$, with an $O(1/T)$ duality gap. For the fractional relaxation, complementary slackness gives exactness at $\mu^\star$. □

The price signal is observable, not assumed. Hayek's point is that a price aggregates dispersed knowledge no planner holds. Here that signal is *empirical*: across OpenRouter we observe both *price* and *volume*—OpenRouter's ∼25 trillion tokens per week across 750 models—a broad record of the price-sensitive segment's revealed choices. Price and volume *together* are the information: volume reveals where demand actually flows, price reveals what it pays, and their ratio is the market's own estimate of value. The clearing price $\mu^\star$ of Prop. 3 is therefore *read from live data*, not posited; `orint` is exactly that observation. OpenRouter is an aggregator in front of a *slice* of inference—the bulk goes direct to the labs or through cloud platforms—so this is the largest *public* cross-vendor price signal, biased toward indie and router-friendly usage, not the whole market. The bias helps rather than hurts the argument: it is precisely the price-sensitive, multi-model segment where routing pays.

4.4 Promotion is statistically sound

Proposition 4 (Non-inferiority promotion)**.** *Let challenger outcomes be i.i.d. with mean* q_B*, and promote B over incumbent A iff the Wilson lower confidence bound* $L_\alpha(\hat{q}_B, n_B) \geq q_A - \delta$*. Then under* $H_0 \colon q_B < q_A - \delta$ *the false-promotion probability is* $\leq \alpha$ *asymptotically.*

This controls the rate at which a worse-and-not-cheaper model is wrongly promoted; δ is the tolerated quality margin and $n_{\min}$ an evidence floor. The bound treats q_A as known; with q_A also estimated and $n_{\min}$ small (we use 8), it is approximate, and in practice we report observed promotion rates rather than lean on the asymptotic guarantee.

4.5 Credit assignment in trajectories

Definition 2 (Trajectory credit)**.** For a trajectory using models at steps $\{1,\dots,k\}$ with terminal reward r, the credit of step i is its Shapley value $\phi_i = \frac{1}{k!}\sum_\sigma \big(r(P_i^\sigma \cup \{i\}) - r(P_i^\sigma)\big)$ over orderings σ, where P_i^σ are the steps preceding i. We estimate ϕ_i by *counterfactual replay*: snapshot the

sandbox at step i, substitute the candidate model's action, and re-roll forward—made cheap and reproducible by the cassette (§6).

5 The Economics of Routing

Two centuries of price theory bear on one operational question: which model should run a task, and what is that choice worth? Table 1 maps each classical idea to a concrete component of the system; the formal statements are in §4. The point is not analogy but *identity*: routing *is* a market, and the same mathematics that allocates goods allocates inference.

Table 1: The intellectual map. Each idea powers a named part of Ricardo.

Economist	Idea	Role in Ricardo	Formal
D. Ricardo (1817)	comparative advantage	the thesis	Prop. 1
V. Pareto (1906)	efficiency frontier	the proof	Prop. 2
L. Walras (1874)	general equilibrium	the clearing price λ	Prop. 3
F. Hayek (1945)	price & volume as information	the observable signal	Prop. 3
K. Arrow (1951)	preference aggregation	the routing table	§6
A. Roth (Nobel 2012)	matching markets	the router	§6
L. Shapley (1953)	value & matching	credit assignment	Def. 2

A worked example. Take the verified coding task of §8. `deepseek-v4-flash` and `claude-opus-4.8` both pass ($q = 1$) at $c = \$0.00003$ and $\$0.00426$. Then $\mathrm{val}_\lambda(\text{deepseek}) - \mathrm{val}_\lambda(\text{opus}) = \lambda(0.00426 - 0.00003) > 0$ for every $\lambda > 0$: the frontier model is Pareto-dominated and the comparative-advantage route is the cheap model *regardless of how one prices quality* (Prop. 2). When quality differs—the free local model *fails*, $q = 0$—the switch threshold against `deepseek` is $\lambda^\star = (1 - 0)/(0.00003 - 0) \approx 3.3 \times 10^4$: one keeps the free model only by valuing a quality point at under 1/33,000 of a dollar, i.e. essentially never. Escalation is thus a calculation, not a preference.

6 Method

The three mechanisms are best read as pseudocode. Score (Alg. 1) estimates q by the highest-trust signal available; Flywheel (Alg. 2) routes by value and promotes challengers under the non-inferiority test of Prop. 4; Replay (Alg. 3) serves the cassette for the non-reproducible boundary while local tools run for real. The tier-1 execution reward, read from the agent's own spans, is

$$r \;=\; \mathrm{clamp}\Big(1 - \alpha\, n_{\text{retry}} - \beta\, n_{\text{tool-err}} - \gamma\, \mathbf{1}[\text{re-prompt}] - \tfrac{1}{2}\,\mathbf{1}[\text{abandoned}],\; 0,\; 1\Big), \qquad r = 0 \text{ on a hard error.} \tag{2}$$

The weights α, β, γ are hyperparameters, so tier-1 is a *parameterised* outcome signal: objective in its inputs (an agent did or did not retry; a tool did or did not error) but subjective in their aggregation. Only tier-2—a hidden test suite, pass/fail—is parameter-free ground truth. We report the sensitivity of the web outcome to α, β, γ in §8 and treat the test-graded coding benchmark as the paper's only fully objective evidence.

On a real 32-call trajectory Replay reproduces all 18 external calls exactly, tolerates paraphrased arguments (fuzzy 0.99), and flags off-script calls—the fidelity statistic that gates whether a comparison is trusted.

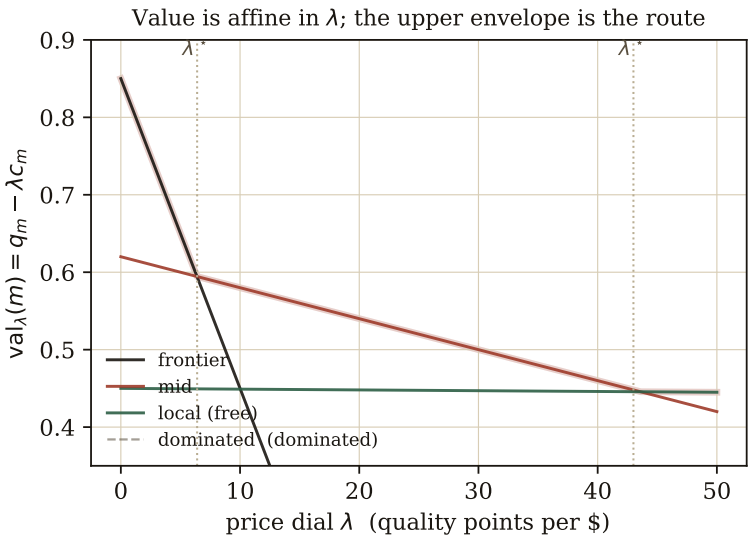

Figure 1: Value is affine in λ: each model's $\mathrm{val}_\lambda(m) = q_m - \lambda c_m$ is a line, routing follows the *upper envelope* (bold), and a dominated model never surfaces for any λ (Prop. 2). Dotted marks are the switch thresholds $\lambda^\star$ of Prop. 2.

Algorithm 1 SCORE(x, y, t) — reward by highest-trust tier

if t agentic **and** execution signals present **then** ▷ tier 1: outcome
  return $\big(r$ via Eq. (2), trust $= 1\big)$
else if reference $y^\star$ available **then** ▷ tier 2: objective
  return (VERIFY$(y, y^\star)$, 1) ▷ test pass-rate / edit distance
else ▷ tier 3: calibrated judge
  return (JUDGE(x, y), ρ_{human}) ▷ ρ: agreement w/ human votes
end if

7 A Flywheel Simulation

We instantiate a routing instance with T=12 buckets and four models—a free but uneven `local`, two mid-tier models with engineered comparative advantages, and a strong-but-dear `frontier`—and run shadow-A/B with UCB candidate selection and posterior-mean value routing. Outcomes are Bernoulli in the latent $q(m, t)$; the learner observes only outcomes, never q. Figure 3 (left) shows value regret against the oracle π_λ collapsing over rounds across 8 seeds, reaching within $\sim$3.6% of oracle value across 8 seeds (final bucket accuracy $\approx$66%; the residual mis-routings fall on buckets where two models are near-tied in value, so they barely cost anything): *the flywheel learns a near-value-optimal table from outcomes alone.* Sweeping the dial (right) traces the achievable cost–quality frontier of Prop. 2; the single-model policies (*always local*, *always frontier*) are interior points the router strictly improves upon—more quality than local, far less cost than frontier.

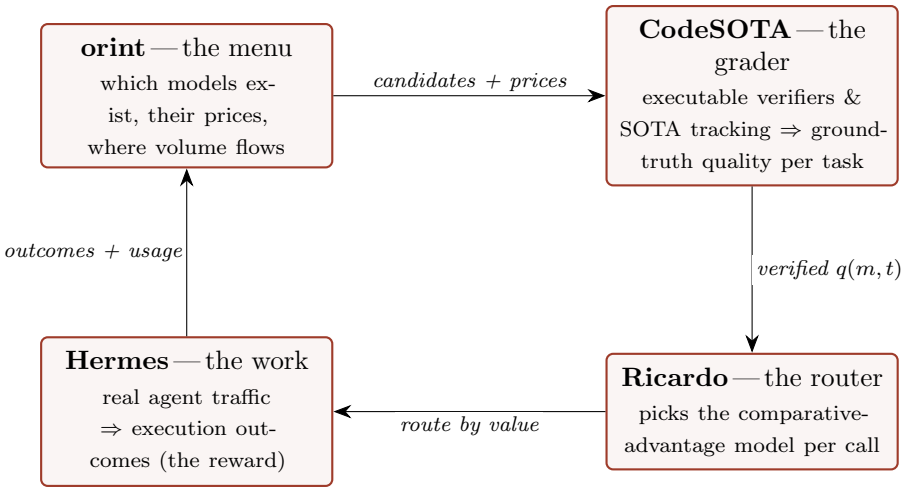

Figure 2: **The Ricardo loop.** *orint* maps the market—the menu of models and their prices. *CodeSOTA* turns that menu into *verified, executable* quality per task, $q(m,t)$: it runs the candidate models through verifier environments and SOTA benchmarks, so the router's grade is a measured fact, not an opinion. *Ricardo* routes each call to the comparative-advantage model. *Hermes* does the real work and emits the execution outcome that is the reward. Outcomes and usage flow back to refresh the menu and re-grade the models—each lap sharpens the private table.

8 Evidence from Owned Traffic

Setup. Trajectories come from a production deployment we operate of the Hermes agent (Nous Research).[2] The incumbent model it calls is a local `qwen3.6-35b-a3b` ($c \approx \$0$); challengers are on OpenRouter; the calibrated judge is `deepseek-v4-flash`.

The judge is lenient; the outcome is not. Over 77 judged trajectories (Fig. 4, left; Table 2) the judge and a tier-1 execution-outcome estimate agree on *chat* (0.98) but diverge sharply on agentic *code/web*: judge 0.61/0.50 vs. outcome 0.11/0.09. Opinion-based scoring overstates agentic success, and the failing buckets are also the most expensive—direct motivation for outcome-grounding. *Sensitivity.* The web outcome depends on the free weights of Eq. (2); sweeping $\alpha \in [0.10, 0.25]$, $\beta \in [0.15, 0.35]$ and the abandoned-task weight $\in [0.3, 0.7]$ (36 settings) moves it only over $[0.076, 0.144]$ (median 0.091), and *every* setting stays below 0.25—far under the judge's 0.50. The exact number is parameter-dependent; the divergence from the judge is not.

Escalation pays only at the frontier. Cassette-style replay on 8 web tasks (Table 4, spend \$0.37): only `claude-opus-4.8` earns escalation (+0.175, 5/8, \$0.043); `minimax-m2.5` is within noise; `glm-4.6` is dominated. At n=8, 5/8 is not significant ($p \approx 0.36$ under a fair coin), so we read this as suggestive, not established. Note the incumbent's q=0.43 here is the replay-*synthesis* judge quality on these 8 tasks given pre-gathered evidence—a different metric and task set from the tier-1 execution outcome of 0.09 in Table 2 (full agentic runs, n=27); the two are not directly comparable, and the gap between them is itself the judge-vs-outcome divergence. The price–quality frontier (Fig. 4, right) is the active set of Prop. 2.

[2] Hermes is an open agent by Nous Research; the authors own the hosted deployment and its task traffic—the source of the feedback data—not the agent or any model.

Algorithm 2 FLYWHEEL — outcome-grounded routing (per bucket t)

$\mathcal{T}[t] \leftarrow$ priors from public benchmarks
for each live task (x, t) **do**
  $m \leftarrow \arg\max_{m'} \ \hat{q}(m', t) - \lambda\, c(m')$; serve $m(x)$ ▷ route by value
  for all candidate $B \in \mathcal{C}$ **do** ▷ off hot path, asynchronous
    $y_B \leftarrow B(x)$; $(q, \_) \leftarrow \textsc{Score}(x, y_B, t)$ ▷ matched pair vs. x
    update posterior $(\hat{q}_B,\ n_B,\ L_\alpha(\hat{q}_B))$
  end for
  for all $B \in \mathcal{C}$ with incumbent A **do** ▷ promotion, Prop. 4
    if $L_\alpha(\hat{q}_B) \geq \hat{q}_A - \delta$ **and** $c_B < c_A$ **and** $n_B \geq n_{\min}$ **then**
      promote B in $\mathcal{T}[t]$
    end if
  end for
end for

Algorithm 3 REPLAY(cassette, tool, args) — mock only the boundary

if tool $\notin$ EXTERNAL **then**
  return RUNINSANDBOX(tool,args) ▷ real local effect
end if
$E \leftarrow \{e \in \text{cassette} : e.\text{tool} = \text{tool}\}$
for all $e \in E$ **do**
  if e.args = args **then**
    return (e.result, EXACT)
  end if
end for
$e^\star \leftarrow \arg\max_{e \in E}\ \text{sim}(\text{args}, e.\text{args})$
if sim $\geq \theta$ **then**
  return ($e^\star$.result, FUZZY)
else
  return MISS ▷ lowers per-run replay fidelity
end if

Hard-truth verification: correctness at a fraction of the cost. An `is_valid_ipv4` task graded by executing 18 hidden tests (Table 3): `deepseek-v4-flash` passes at \$0.00003 while `claude-opus-4.8` passes the *same* tests at \$0.00426—identical correctness, $\sim$127$\times$ less per output token (\$25 vs \$0.197 per M)—and the local incumbent *fails*. This is Prop. 1 realised: the comparative-advantage choice is the cheap-correct challenger, neither the free-wrong nor the dear-correct model.

9 Discussion and Limitations

The real experiments are deliberately small (n=8 web; single coding task) and demonstrate the *mechanism* and its failure modes, not a production table; the simulation supplies the controlled convergence evidence the field samples cannot. Replay measures synthesis on fixed evidence, not tool use; the high-fidelity path is a full agent re-run under the cassette. Credit assignment (Def. 2) is the hardest open problem. The calibrated-judge tier is only as good as its human anchoring. None of these undercut the central claim. A router is a commodity; the defensible asset is the *feedback data*

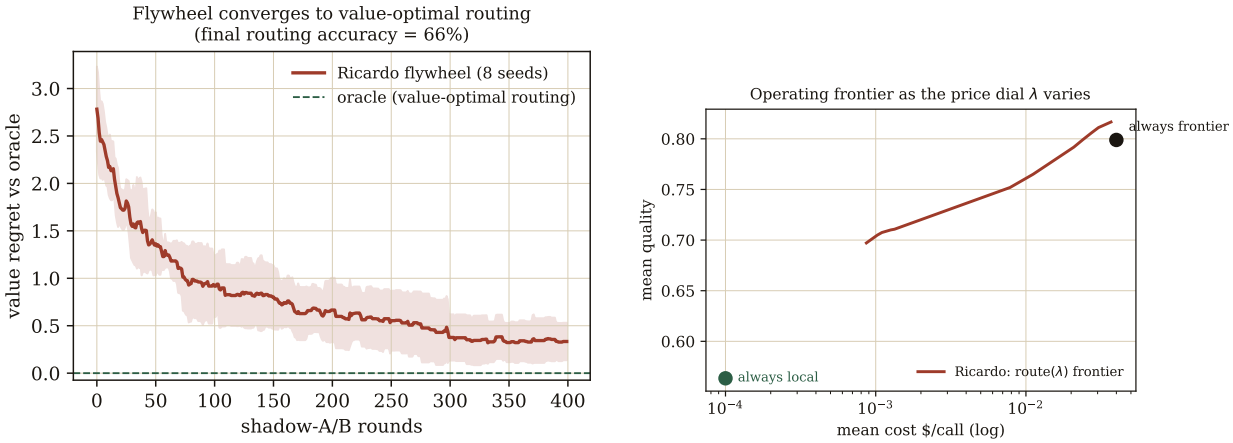

Figure 3: **Left:** value regret vs. the oracle over shadow-A/B rounds (8 seeds, band = ±1 s.d.). **Right:** the operating frontier as λ varies, against single-model baselines.

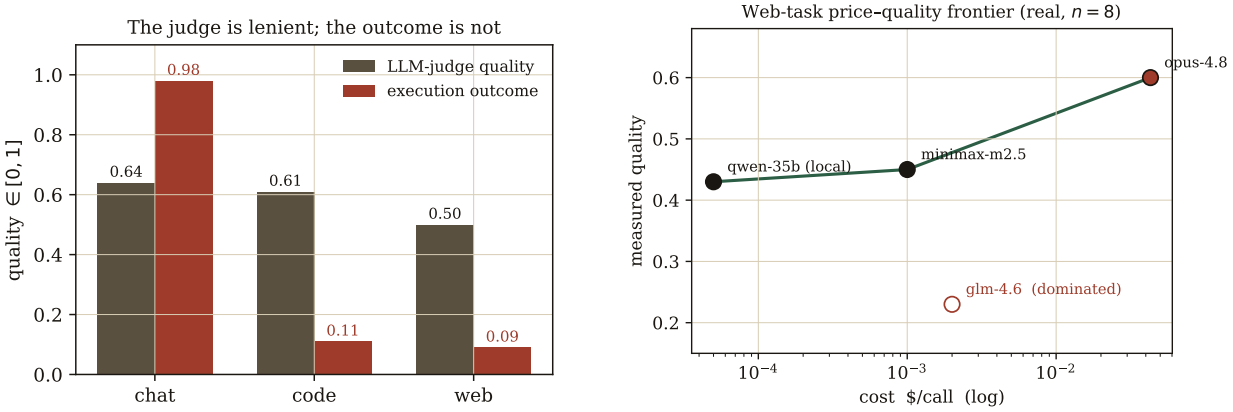

Figure 4: **Left:** judge vs. execution outcome per modality (real, 77 trajectories). **Right:** the web price–quality frontier; `glm-4.6` is dominated.

of economically valuable work—a compounding corpus of (prompt → model → verified outcome) triples produced when an outcome-grounded router does real, paid, multimodal tasks under a closed feedback loop. The data is a by-product of *doing the work*, which is precisely why it cannot be scraped, synthesized, or bought: it exists only where genuine demand meets a genuine outcome signal.

10 Conclusion

Ricardo makes the routing decision the way the market already makes it in aggregate—cheap by default, escalate where it pays—but per task, by measured outcome, with receipts. Theory says dominated models vanish and λ clears the market; simulation says the table is learnable from outcomes; owned traffic says the savings are real.

Table 2: Baseline, 77 trajectories.

Modality	n	Judge	**Outcome**
chat	29	0.64	**0.98**
code	21	0.61	**0.11**
web	27	0.50	**0.09**

Table 3: Code task, 18 executed tests.

Model	Verifier	$/call
qwen-35b (local)	**FAIL**	0.00091
deepseek-v4-flash	**PASS**	**0.00003**
gemini-3.1-flash-lite	PASS	0.00019
claude-opus-4.8	PASS	0.00426

Table 4: Web escalation replay (n=8); incumbent free/local at q=0.43.

Model	Quality	Uplift	Win rate	$/call
qwen-35b (local)	0.43	—	—	∼0
claude-opus-4.8	**0.60**	**+0.175**	**5/8**	0.043
minimax-m2.5	0.45	+0.03	2/8	0.001
glm-4.6	0.23	−0.20	1/8	0.002

References

[1] I. Ong et al. RouteLLM: Learning to Route LLMs with Preference Data. *arXiv:2406.18665*, 2024.

[2] L. Chen, M. Zaharia, J. Zou. FrugalGPT. *arXiv:2305.05176*, 2023.

[3] Not Diamond. Awesome AI model routing. `https://github.com/Not-Diamond/awesome-ai-model-routing`, 2024.

[4] L. Zheng et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *NeurIPS*, 2023. arXiv:2306.05685.

[5] Y. Dubois et al. Length-Controlled AlpacaEval. *arXiv:2404.04475*, 2024.

[6] W.-L. Chiang et al. Chatbot Arena. *arXiv:2403.04132*, 2024.

[7] C. Jimenez et al. SWE-bench. *ICLR*, 2024. arXiv:2310.06770.

[8] S. Yao et al. τ-bench. *arXiv:2406.12045*, 2024.

[9] S. Zhou et al. WebArena. *arXiv:2307.13854*, 2023.

[10] G. Mialon et al. GAIA: A Benchmark for General AI Assistants. *arXiv:2311.12983*, 2023.

[11] D. Ricardo. *On the Principles of Political Economy and Taxation*. John Murray, 1817.

[12] V. Pareto. *Manuale di economia politica*. 1906.

[13] F. A. Hayek. The Use of Knowledge in Society. *American Economic Review*, 35(4), 1945.

[14] K. J. Arrow. *Social Choice and Individual Values*. Wiley, 1951.

[15] L. S. Shapley. A Value for n-Person Games. 1953.

[16] A. E. Roth, M. Sotomayor. *Two-Sided Matching*. Cambridge Univ. Press, 1990.

[17] L. Walras. *Éléments d'économie politique pure*. 1874.