

Do LLM Routers Pay for Themselves?

An audit of 17 router configurations on long, open-ended questions at billed prices

Autobench.org team · 24 October 2026 · autobench.org/reports/do-llm-routers-pay-for-themselves

Summary

The number of large language models (LLMs) available to a deployment keeps growing, and their prices differ widely: output prices in the 20-model pool studied here range from \$0.15 to \$10 per million tokens. An LLM router selects, for each request, which of these models should answer it. The main case for routing is cost reduction. Most typical requests can be answered by a cheaper model at the quality of an expensive default, so sending them to that cheaper model should lower cost without lowering quality.

We tested this claim on a frozen benchmark of long, open-ended questions that span 10 subject areas and five difficulty levels. The benchmark follows the collective LLM-as-a-judge methodology that we developed and validated in earlier Autobench work [15]. We collected answers from 20 models at their billed prices and scored every answer with a 10-model collective jury. We then recorded which model seven router systems, in 17 configurations, selected for each question, and looked up the score and cost of each selection to obtain each router configuration's trade-off between quality and cost. Scoring rules, baselines and success criteria were fixed and time-stamped before any router was queried. Quality is measured by the *best-answer rate* (or Delta-Hit, DH): the share of questions on which the selected model's answer scored within a margin, set by disagreement among the judges, of the best answers available for that question.

The paper makes four contributions, each with its main result:

1. **No audited router beat always using the Luna model on both quality and cost.** We audited 17 router configurations on long, open-ended questions at billed prices: 16 configurations of shipped router systems and our own escalation cascade. Sending every question to GPT-5.6 Luna, the best low-cost fixed model, gives a best-answer rate of 85.3% at \$0.0043 per question, and no router configuration had both a higher best-answer rate and a lower cost in any of 10,000 bootstrap resamples of the questions.
2. **Some router configurations (5 of 17) collapse to routing almost all questions to a single model.** We call a router configuration collapsed when one model receives 90% or more of its decisions; each collapsed configuration scores within 0.01 points of the mean jury score of that model. Cost settings often switch between defaults rather than trade quality for cost: Not Diamond's best-answer rate falls from 80.3% to 46.6% between two adjacent settings, and OpenRouter's most expensive tier does not exceed the tier below it.
3. **A large cost reduction is available but was not realised by any router.** Reference baselines for this setting show the size of the gap: the best fixed model at each price, a fixed model chosen without hindsight, a subject-informed baseline (one fixed model per subject area) and the hindsight limit. The hindsight limit, which picks the cheapest model giving a best answer on each question, reaches 100% by construction at \$0.0018 per question. OpenRouter Auto can route to all 20 models on the benchmark, yet the four tiers we audited reach 65.7–79.1% at \$0.00055–\$0.0289 per question.
4. **An open dataset for benchmarking routers.** The question set, model answers, jury scoring runs and router decisions are released at huggingface.co/datasets/AutoBench/router-benchmark-2026-10, so that the audit can be reproduced and new routers can be compared with the same baselines without querying the audited routers again.

Figure 1 places the audited router configurations and the reference policies on one chart of quality against cost.

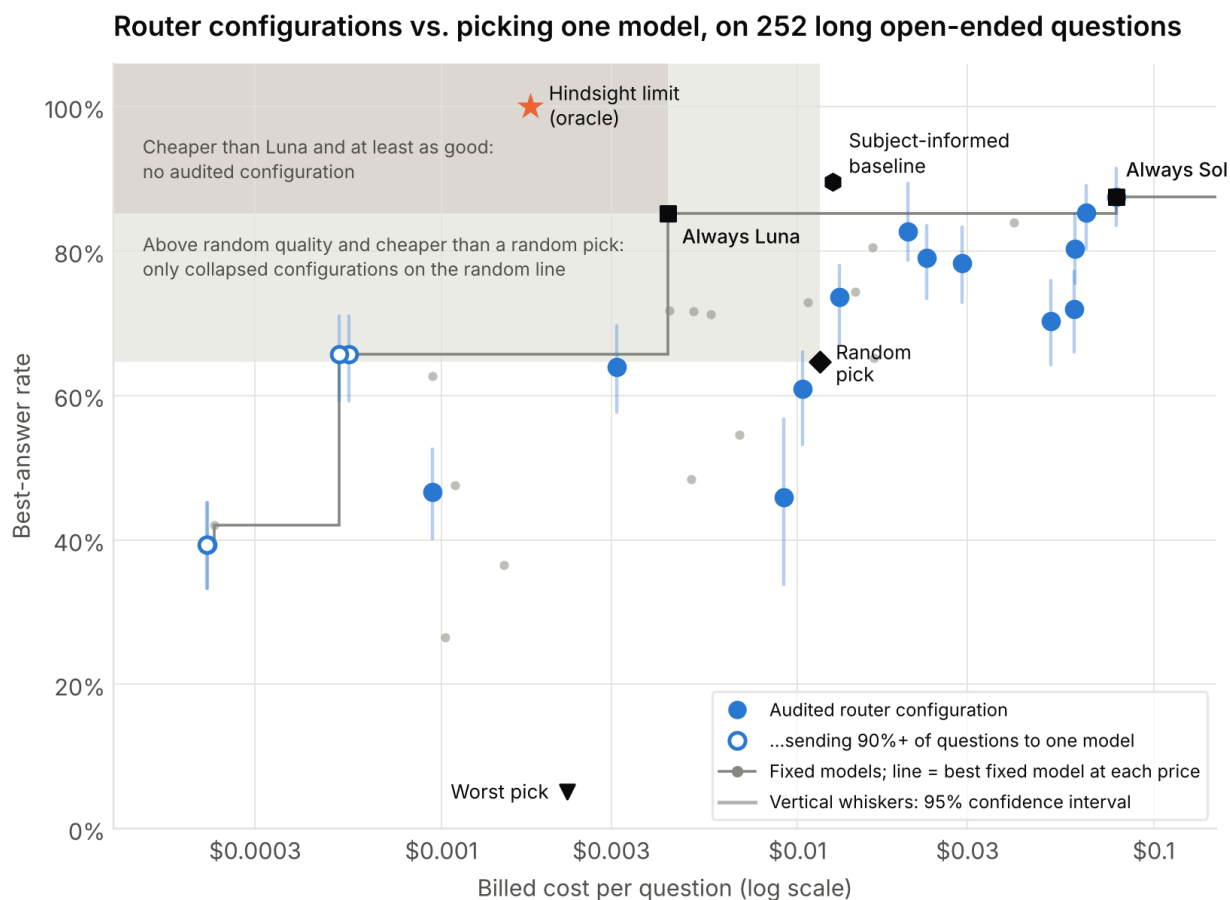


Figure 1. Best-answer rate against billed cost per question for the 17 audited router configurations (blue) and reference policies (black). Vertical whiskers are 95% confidence intervals from 10,000 bootstrap resamples of the questions. Hollow blue markers denote configurations that send 90% or more of questions to one model. Grey dots are the 20 models used as fixed policies; the grey line traces the best fixed model available at each price. The darker shaded region contains policies that are cheaper than always-Luna and at least as good; no audited configuration falls in it. The lighter region contains policies above random quality and cheaper than a random pick; the two configurations inside it, OpenRouter low and Orca quality (65.7%), are collapsed onto DeepSeek Flash and sit on the random line, 1.0 percentage point above the random pick and within its confidence interval (58.7–70.6%). The black hexagon is a subject-informed baseline, which requires each question’s subject in advance (Section 3.6). The orange star is the hindsight limit (oracle). Not Diamond’s k0 setting coincides with always-Sol.

1. Motivation and prior evidence

Routing has moved from research into products, but the savings reported for it depend strongly on how it is evaluated. Large cloud providers offer routers as platform features (the Azure model router, Amazon Bedrock prompt routing [16]), and multi-model gateways have become acquisition targets: Stripe agreed to

acquire OpenRouter in August 2026 in a transaction reported at over \$7 billion [13], and Palo Alto Networks completed its acquisition of the Portkey gateway in May 2026 [14]. Published evaluations of routing report the following results:

Source	Setting	Reported result
RouteLLM [1]	A strong and a weak model, chat and reasoning benchmarks (2024)	Cost reduced by more than half in some cases without loss of answer quality
FrugalGPT [2]	Cascade over 2023 models	Up to 98% lower cost at GPT-4 quality
LLMRouterBench [3]	33 models, 21 datasets, independent re-evaluation	Up to 31.7% cost reduction at the best single model's accuracy; several routers, including commercial ones, do not exceed the best single model
RouterArena [4]	8,400 queries, commercial and open routers	Most routers operate near the strongest model's cost; commercial routers reach higher accuracy than open ones, but at much higher cost
The Routing Plateau [5]	21 routing methods, five benchmarks	Methods of very different design reach similar accuracy, well below the hindsight limit

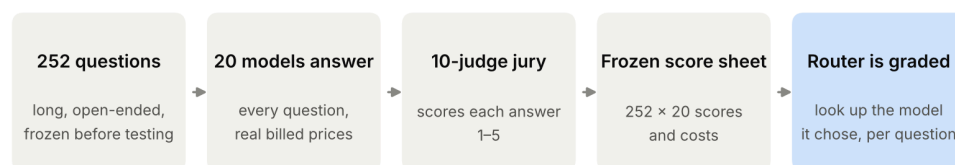
The largest savings come from evaluations by the methods' own authors: RouteLLM and FrugalGPT each evaluate the system they propose. The three 2026 studies (LLMRouterBench, RouterArena and The Routing Plateau) re-evaluate many routers on common data and report smaller gains. All three use tasks with reference answers (multiple choice, mathematics, code), and they mostly evaluate routing methods rather than the router configurations that customers can select. How these configurations behave on open-ended work, where no reference answer exists, is therefore not covered. We ask:

On a workload it was not tuned for, does a shipped router outperform the best fixed model at its price on both quality and billed cost?

The comparison is with a fixed model because a fixed model requires no additional infrastructure and is the alternative available to any deployment. The workload is long, open-ended questions (reports, analyses, plans and single-step agent plans) for two reasons: answer length makes per-token price differences large, and benchmarks with reference answers do not cover this setting.

2. Method

To answer this question without reference answers, we built a benchmark in which every model answers every question before any router is queried, and graded each router by looking up the score and cost of the model it selected (Figure 2).



Rules, baselines and pass criteria were fixed and time-stamped before any router was called.

Figure 2. Evaluation pipeline. All answers are generated and scored before any router is queried; each router is graded by looking up the frozen score and cost of the model it selected.

Questions. 252 questions spanning 10 subject areas and five difficulty levels, frozen and hashed before evaluation.

Models. 20 models frozen on 23 August 2026, with output prices from \$0.15 to \$10 per million tokens (a 67-fold range). Appendix B lists all 20 with their prices. The models referred to most often are GPT-5.6 Luna ("Luna"), GPT-5.6 Sol ("Sol") and DeepSeek V4 Flash ("DeepSeek Flash"). Every model received every question; the billed cost of each answer, including input and reasoning tokens, was recorded.

Jury. Answers were scored on a 1–5 scale by a jury of 10 models drawn from the 14 eligible judges. The procedure builds on the collective LLM-as-a-judge method of Autobench [15], extended here with per-answer billed costs and a quality threshold derived from inter-judge disagreement.

Routers. Seven router systems in 17 router configurations: four commercial services (Not Diamond, OpenRouter Auto, Orca, Azure model router), two open-source self-hosted routers (NadirClaw, Brick) and our own escalation cascade. The cascade follows the FrugalGPT design [2]: a cheap model answers first, a judge checks the answer, and the question is escalated to a stronger model when the check fails. It is included so that the cascade design is represented, and its result describes our implementation rather than the design in general. We call these 17 the audited router configurations. A router configuration is one customer-selectable setting, such as OpenRouter's "high" tier or Azure's "balanced" mode. Router selections were collected between 30 August and 1 September 2026.

Router catalogues. The routers differ in how many of the 20 pool models they can select, and the results for each router configuration describe its behaviour within that overlap:

System	Configurations	Pool models it can select
OpenRouter Auto	low, medium, high, max	20 of 20
Orca	cheapest, balanced, quality	15 of 20
Not Diamond	k0, k4, k5, native	7 of 20
NadirClaw	one	4 of 20
Azure model router	cost, balanced, quality	3 of 20 (gpt-oss-120b, Luna, Sol)
Brick	one	3 of 20
Escalation cascade (ours)	one	2 (a cheap model, then a stronger one)

Where a router accepts a candidate list, the list was restricted to the pool models it supports. Some routers also selected models outside the pool; those answers were generated once and scored by the same jury. A router evaluated with a different candidate set, or on a deployment with a different catalogue, can therefore behave differently from the configuration measured here. The comparison with fixed models is unaffected, because every fixed model is drawn from the same 20.

Rules fixed in advance. Scoring rules, baselines, threshold widths and success criteria were fixed and time-stamped on 29 August 2026, before any router was queried.

The benchmark is this frozen set of questions, answers, scores and costs. Grading by lookup evaluates all routers on identical answers and makes counterfactual policies (what a router could have selected) directly computable. It is valid for single-turn routing in which one model answers each question; it does not apply to routers that switch models within a multi-turn session.

2.1 Best-answer rate

The quality measure has to account for a property of open-ended questions: on a typical question several models produce answers that the jury cannot reliably separate. For each question we therefore take the mean score of the three highest-scoring answers and treat any answer within a margin of that mean as a *best answer*. The margin is set at about two standard errors of the ten-judge mean: two-thirds of the median

inter-judge standard deviation (0.20 points on the 1–5 scale). The *best-answer rate* (or Delta-Hit, DH) of a policy is the share of questions on which its selected model gave a best answer; differences between rates are given in percentage points. It is a win-or-tie rate against the strongest available answers, with the tie margin set by inter-judge disagreement. The criterion admits several models per question: the median number of models giving a best answer is 12 of 20 (interquartile range 9–17). On 90% of questions at least one model that is cheaper than Luna on that question gives a best answer. The routing problem on this benchmark is therefore to identify, before answering, which inexpensive model will suffice. LLMRouterBench reports a related pattern: when only one to three models answer a question correctly, routers often fail to select one of them [3].

2.2 Why average scores are not the primary measure

A simpler alternative would be to compare mean jury scores. They are a poor primary measure on this benchmark, however, because they are concentrated near the top of the scale: 75% of the 4,933 scored answers received 4.5 or more (per-configuration means are listed in Appendix A). Mean scores per policy range from 3.49 (worst choice on every question) through 4.55 (random choice), 4.74 (always-Luna) and 4.78 (always-Sol) to 4.91 (best choice on every question), so the differences of interest are a few hundredths of a point. The best-answer rate expresses the same information as a share of questions.

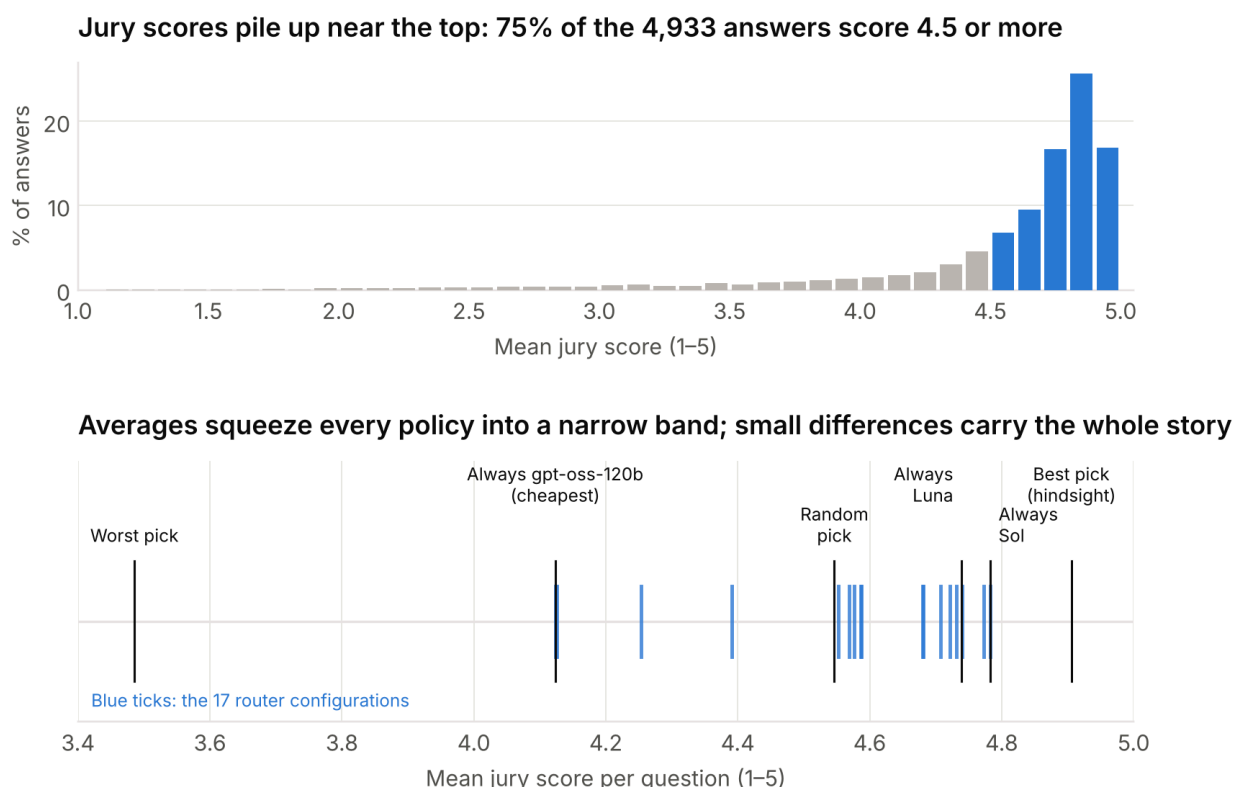


Figure 3. Top: distribution of all 4,933 jury scores. Bottom: mean jury score for reference policies (black) and the 17 audited router configurations (blue).

2.3 Reference policies

Six reference policies frame the results, from the worst to the best possible choice on each question:

Policy	Definition	Best-answer rate	Cost per question
Worst pick	lowest-scoring model on each question (hindsight)	5.2%	\$0.0023
Random pick	random in-pool pick: one of the 20 models drawn per question (frozen random policy)	64.7%	\$0.0115
Always DeepSeek Flash	single fixed model	65.7%	\$0.0005
Always GPT-5.6 Luna	single fixed model, low-cost reference	85.3%	\$0.0043
Always GPT-5.6 Sol	single fixed model, highest best-answer rate	87.5%	\$0.079
Hindsight limit (oracle)	cheapest best answer on each question	100%	\$0.0018

The hindsight limit requires knowledge of all scores before selection and is not achievable by a router; it indicates the size of the available saving. The random pick is the frozen random policy: one in-pool model drawn per question with a seed fixed in advance, so that, like the audited routers, it makes a single selection per question. Its 95% confidence interval is 58.7–70.6%; its expected value over all possible draws is 62.6% at \$0.011 per question.

3. Results of the audit

3.1 No audited router configuration outperformed the best fixed model at its price

The central result of the audit is that no router configuration did better than a fixed model at its price. All but one of the 17 audited configurations lie on or below the line tracing the best fixed model at each price (Figure 1): for their mean cost, at least one fixed model matched or exceeded their best-answer rate. The exception, Orca's cheapest setting, exceeds the line by 0.4 percentage points while sending 99% of questions to the model that defines the line at that price. Against the random pick the picture is similar. Only two configurations are above random quality and cheaper than a random pick: OpenRouter low and Orca quality, both at 65.7% and both collapsed onto DeepSeek Flash, so they sit on the random line rather than inside the region. Their margin over the random pick, 1.0 percentage point, is well inside its confidence interval.

Against always-Luna, the fixed model closest to the strongest at low cost, we counted the bootstrap resamples (10,000 resamples of the 252 questions) in which a router configuration had both a higher best-answer rate and a lower cost. The count was zero for all 17 configurations, which bounds the probability of such an outcome below 3 in 10,000 at 95% confidence.

Two sensitivity checks leave this result unchanged:

- **Margin width.** Repeating the analysis at five margin widths, from 0.5 to 2 times the median judge standard deviation, the region cheaper than Luna and at least as good remains empty, and the best router configuration cheaper than Luna stays within 2.1 percentage points of the random pick (Figure 4).
- **Missing answers.** About 2% of question-model pairs have no score (time-outs or unavailable models). Counting each as a miss changes 14 of the 17 router configurations by less than 4 points; the three larger changes lower those configurations, and the region remains empty.

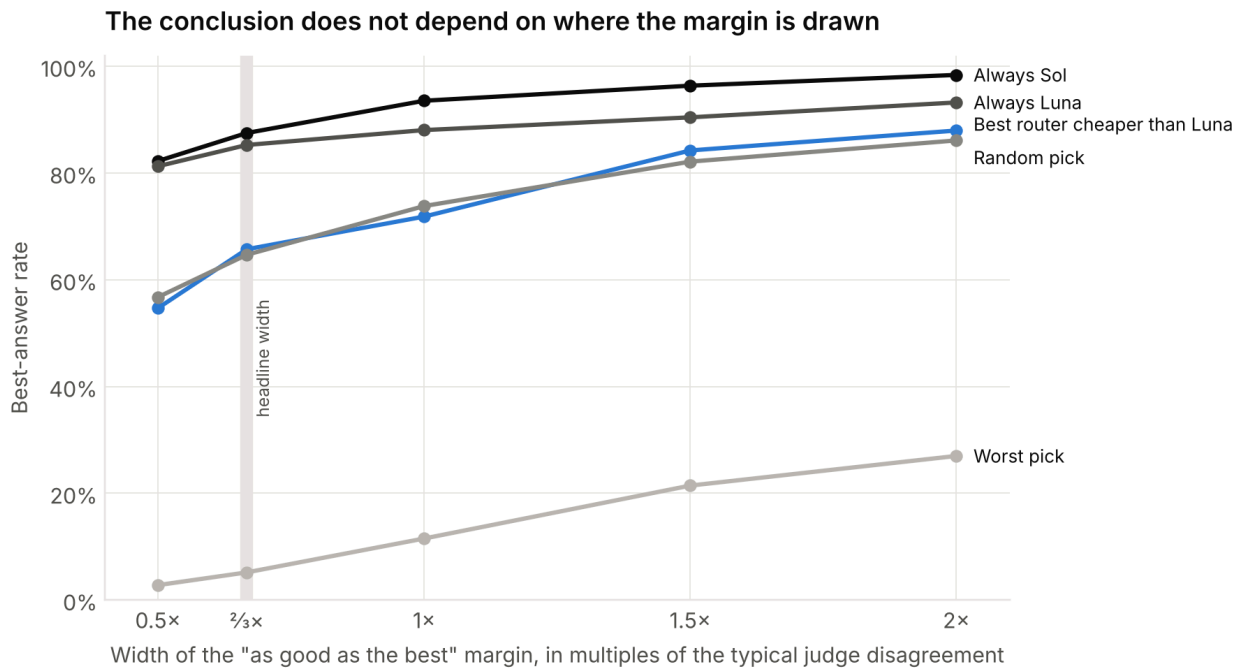


Figure 4. Best-answer rate as a function of margin width, in multiples of the median judge standard deviation. The best router configuration cheaper than Luna remains below Luna at every width.

Luna was identified as the reference after the scores were known, and selecting a best single model in-sample can bias the comparison towards the baseline [7]. We therefore also selected a fixed model without hindsight: the cheapest model within three points of the best on four-fifths of the questions, evaluated on the remaining fifth, averaged over 20 random splits. This policy reaches 83.5% at \$0.025 per question, because small training splits sometimes favour more expensive models. The best router configurations are comparable to it (for example Azure balanced, 82.7% at \$0.020) rather than better. The finding is therefore that the configurations tested did not outperform a fixed model chosen on a trial set.

3.2 Five router configurations collapse onto a single model

Several router configurations track the fixed-model line because they barely route. We say that a configuration has collapsed when one model receives 90% or more of its decisions. Five of the 17 configurations sent 90% or more of the 252 questions to one model:

- all three Orca settings: cheapest and balanced to gpt-oss-120b, quality to DeepSeek Flash;
- OpenRouter’s low tier (DeepSeek Flash);
- Not Diamond’s k0 setting (Sol on every question, equivalent to always-Sol).

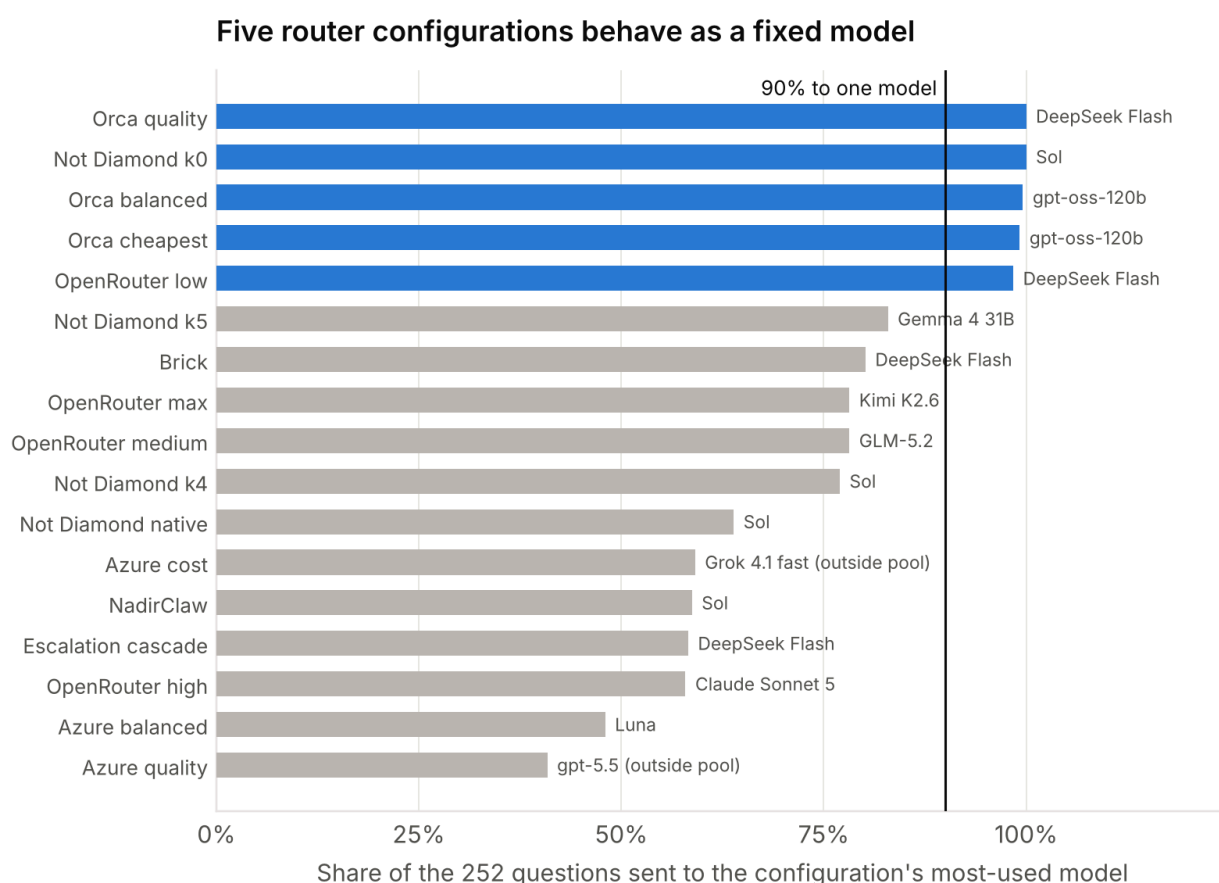


Figure 5. Share of questions each router configuration sent to its most-used model. Blue bars exceed the 90% threshold.

The mean jury scores of these five configurations are within 0.01 points of the mean score of the fixed model each one concentrates on. Concentration on one model has been reported before: RouterArena finds most routers near the strongest model's cost [4], and routers trained to predict a score per model tend to select the strongest model as the budget increases [8]. That mechanism explains concentration on the strongest model; on this benchmark concentration also occurs on the cheapest models, which it does not explain. Collapse here describes how a configuration distributes its decisions; it is unrelated to “model collapse” in training on synthetic data.

The cost settings of several router systems also switch between defaults rather than trading quality for cost. Orca's quality setting replaces gpt-oss-120b with DeepSeek Flash. Not Diamond's cost-quality setting holds quality up to k4 (80.3% at \$0.0600) and then drops at the adjacent setting, k5, to 46.6% at \$0.00094: cost falls 64-fold and the best-answer rate by 33.7 percentage points, to below the random pick. OpenRouter's max tier (78.3% at \$0.0289) does not exceed its high tier (79.1% at \$0.0230) and costs 26% more.

3.3 Routers rarely agree on the model to use

The router configurations also disagree about which model a question needs. Across the 59 pairs of configurations from different vendors, the median share of questions on which both selected the same model is zero. Because disjoint catalogues could produce this mechanically, we restricted the comparison to the one vendor pair with a large shared catalogue (Orca and OpenRouter, 15 shared models out of 20). Across the 12 configuration pairs from these two vendors, the median agreement was 0.4% of questions. The one pair with high agreement (98%) results from both configurations defaulting to DeepSeek Flash.

3.4 The available saving was within reach of a full-catalogue router

The hindsight limit shows how much a router that chose well could have saved, and OpenRouter Auto is the clearest test of whether that saving was reachable. It can route to all 20 models, so its in-catalogue hindsight limit equals the global \$0.0018 per question. The four tiers we audited lie below the best-fixed-model line throughout (Figure 6): the low tier reaches 65.7% at \$0.00055 (mostly DeepSeek Flash) and the high tier 79.1% at \$0.023, about five times Luna’s cost.

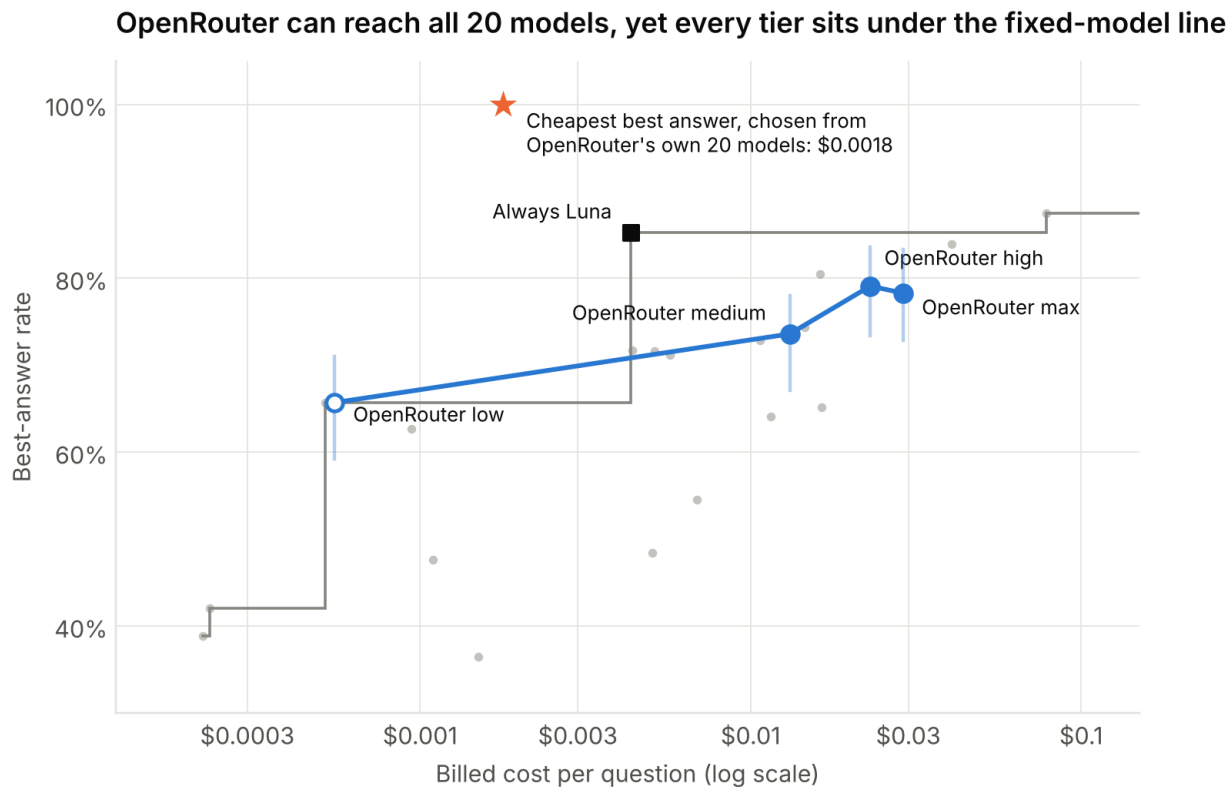


Figure 6. OpenRouter Auto tiers relative to the best fixed model at each price. The orange star is the hindsight limit within OpenRouter’s catalogue. Whiskers are 95% confidence intervals.

OpenRouter’s documentation at the time of data collection (30 August–1 September 2026) describes the Auto Router as classifying each request into one of about 30 task types and ranking models by OpenRouter customers’ spend on that task type over the preceding seven days, within the selected cost tier [11]. Such a ranking reflects popularity rather than per-question adequacy, which is consistent with tiers that behave as spend levels and with the low tier’s concentration on DeepSeek Flash. LLMRouterBench also reports OpenRouter’s router below the best single model on tasks with reference answers [3]. For routers with smaller catalogues the comparison is less direct; Azure’s router, for example, does not include DeepSeek Flash or the other low-cost models that perform well on this benchmark.

3.5 Rankings on short multiple-choice questions do not transfer

Routers and models are often compared on short questions with reference answers, and such comparisons do not predict behaviour on this benchmark. On a slice of RouterArena's short multiple-choice questions, DeepSeek Flash, Luna and Sol are within 1.3 percentage points of each other and Sol costs 5.7 times as much as DeepSeek Flash. On this benchmark the same three fixed policies separate (65.7%, 85.3% and 87.5%), and Sol's cost is 132 times DeepSeek Flash's on a paired 242-question comparison (Figure 7).

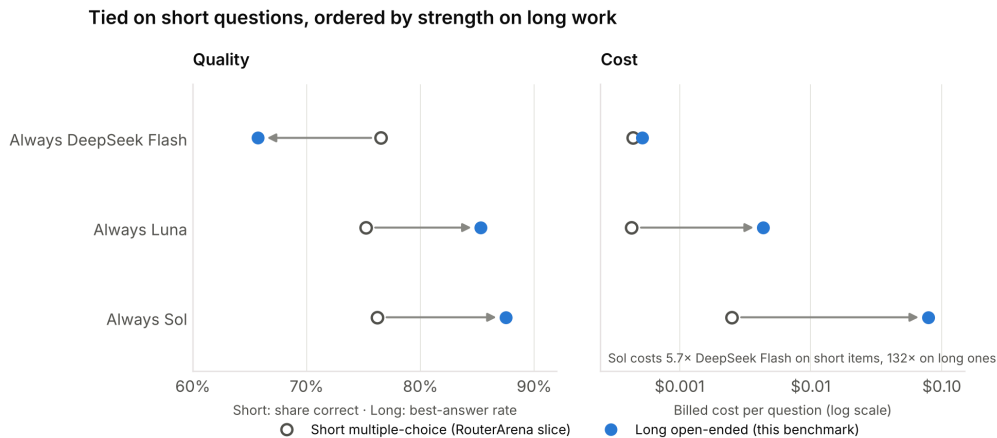


Figure 7. The same three fixed policies on short multiple-choice questions (hollow) and on this benchmark (filled). Quality measures differ: share correct on the short questions; best-answer rate on this benchmark.

The cost ratio grows because long answers contain thousands of output and reasoning tokens, so per-token price differences are multiplied. Performance on short-question leaderboards is therefore weak evidence about performance on long, open-ended work.

3.6 Subject-informed baseline

The fixed-model baselines above ignore what a question is about. Recent work reports that a fixed mapping from task type to model captures much of the benefit of routing [6], so we built such a mapping from the 10 subject labels of our question set: for each subject, the cheapest model within three points of the best on four-fifths of the questions, evaluated on the held-out fifth. It reaches 89.6% at \$0.0125 per question (the black hexagon in Figure 1). This baseline requires each question's subject to be known at routing time, was fitted on about 20 questions per subject, and assigns the same model to all questions of a subject, so it does not adapt to individual requests or users. It is reported as an informed reference point rather than as a deployable router.

4. Implications for practitioners

For a team choosing a routing policy, the audit identifies the comparison a router has to win: a fixed model, or a fixed model per subject area or task type, chosen on a trial set. The results describe one question set, one evaluation period and routers that were not tuned to this workload, and they are not a vendor ranking.

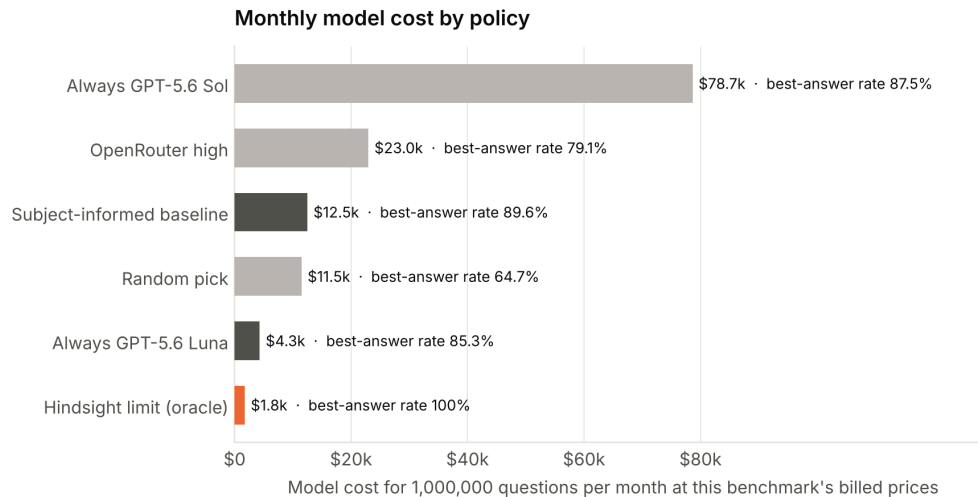


Figure 8. Monthly model cost at one million questions, at this benchmark's billed prices, with each policy's best-answer rate.

At one million questions per month, always using Sol costs about \$79,000, the subject-informed baseline about \$12,500, always using Luna about \$4,300 and the hindsight limit about \$1,800. The audited router configurations spend up to \$64,700 per month (Azure "quality" is the most expensive) without exceeding Luna's best-answer rate. These figures count model tokens only. Some routers add their own charges (Azure's router bills its own input prompt in addition to the selected model [12]), and gateways may charge a fee on spend; including them would widen the differences.

Questions that make a router evaluation comparable to this one:

1. Baseline: is the router compared with the best fixed model at its own price, chosen without hindsight on the target workload, and with one fixed model per subject area or task type?
2. Concentration: what share of traffic goes to the most-used model?
3. Operating range: do the cost settings trade quality for cost, or switch between fixed defaults?
4. Agreement: how often does the router select the same model as an independent router?
5. Workload match: was the reported result measured on questions similar to the target workload (for example long, open-ended questions rather than short multiple-choice items)?
6. Routing cost: does the router bill its own tokens or a share of spend?

On this evidence, a fixed model per subject area or task type, selected on a small trial set and reviewed as prices change, is the default a router should be shown to beat on the target workload.

5. Limitations

- Single benchmark and period. Results come from 252 questions and router behaviour between 30 August and 1 September 2026. Routers and prices change; the method is reusable but the numbers should be re-measured; the direction of any change is unknown.
- Single-turn only. Agentic work is represented by single-step plans. Multi-step tool-use sessions are not covered, and lookup grading does not apply to routers that switch models within a session.

- **Cost scope.** Costs are billed token costs and exclude router fees, gateway fees, caching discounts and negotiated rates. Router and gateway fees would raise the cost of routed policies; caching discounts and negotiated rates would lower costs for routed and fixed policies alike.
- **Close comparisons.** Differences of a few percentage points between individual router configurations are within the sampling uncertainty of 252 questions and should not be read as rankings.
- **Jury.** Model judges can favour their own answers and share errors. Recent measurements put the information in a 9-judge frontier panel at about two independent votes [9] and in a 10-judge panel at about 3.5 [10]. We use a disagreement-based margin and report results across margin widths (Figure 4). About half of all answers were scored by a jury that included the answering model; removing that model's own vote changes those scores by -0.014 points on average and the reference policies by less than half a point (always-Luna 85.3% $\rightarrow$ 85.7%, always-Sol 87.5% $\rightarrow$ 87.1%). Agreement between the jury and human raters on this question set has not yet been measured. None of the audited routers was trained on this jury's scores.

6. What next

Scoring another router on this benchmark is a lookup; adding a question is the expensive step, because every new question has to be answered by the whole pool and judged. The next versions of the benchmark will therefore extend it in the following directions:

- A larger question corpus, large enough to report results per subject area and per difficulty level with useful precision.
- Paraphrase sets, to measure whether a router selects the same model for the same request worded differently.
- Larger and more varied judging pools, and a human-rated subset, to measure jury bias and to report intervals that include jury uncertainty.
- More router systems and router configurations, including routers evaluated on their full catalogues.
- Multi-turn and agentic routing, where a router may switch models within a session; this requires live runs rather than lookup grading.

Data availability

The questions, the answers from all 20 models, the jury scoring runs, billed costs and the router decisions used in this study are released at huggingface.co/datasets/AutoBench/router-benchmark-2026-10, so that every result in this paper can be recomputed without querying the routers again. This paper is available at autobench.org/reports/do-llm-routers-pay-for-themselves.

Appendix A. All router configurations

Router configuration	Best-answer rate	95% confidence interval	Cost per question	Share to top model	Mean jury score	Questions scored
OpenRouter low	65.7%	59.2–71.0	\$0.00055	98%	4.59	245
OpenRouter medium	73.6%	67.1–78.0	\$0.0131	78%	4.68	246
OpenRouter high	79.1%	73.4–83.6	\$0.0230	58%	4.71	244
OpenRouter max	78.3%	72.9–83.3	\$0.0289	78%	4.72	240
Not Diamond k0	87.5%	83.5–91.5	\$0.0787	100%	4.78	248
Not Diamond k4	80.3%	75.5–85.2	\$0.0600	77%	4.74	249
Not Diamond k5	46.6%	40.2–52.6	\$0.00094	83%	4.39	251
Not Diamond native	72.0%	66.0–77.2	\$0.0598	64%	4.68	250
Orca cheapest	39.3%	33.3–45.2	\$0.00022	99%	4.13	252
Orca balanced	39.3%	33.3–45.2	\$0.00022	99.6%	4.12	252
Orca quality	65.7%	59.2–71.0	\$0.00052	100%	4.59	245
Azure cost	45.9%	33.8–56.8	\$0.0091	59%	4.25	74
Azure balanced	82.7%	78.8–89.4	\$0.0204	48%	4.73	179
Azure quality	85.3%	80.3–89.1	\$0.0647	41%	4.77	238
NadirClaw	70.3%	64.3–75.9	\$0.0514	59%	4.55	249
Brick	63.9%	57.7–69.7	\$0.0031	80%	4.58	241
Escalation cascade (ours)	60.9%	53.2–66.0	\$0.0104	58%	4.57	235
<i>Random pick (frozen random policy)</i>	64.7%	58.7–70.6	\$0.0115	—	4.55	252
<i>Always GPT-5.6 Luna</i>	85.3%	80.9–89.6	\$0.0043	—	4.74	251
<i>Always GPT-5.6 Sol</i>	87.5%	83.5–91.5	\$0.0787	—	4.78	248
<i>Fixed model chosen on a trial set (held-out)</i>	83.5%	—	\$0.025	—	—	252
<i>Subject-informed baseline (held-out)</i>	89.6%	—	\$0.0125	—	—	252

Confidence intervals are percentile intervals from 10,000 bootstrap resamples of the questions. “Questions scored” is below 252 where the selected model has no score for some questions; Azure’s cost setting returned a usable model on 74 questions. The escalation cascade is our implementation of a standard escalate-on-failure design and was run on 242 questions. When the cascade accepted the cheap model’s answer, that question was scored on the frozen answer rather than on the live answer its judge had checked. The two held-out rows are averages over 20 random five-way splits of the questions.

Appendix B. Model pool

The 20 models were frozen on 23 August 2026. Prices are list prices per million tokens at that date; the 14 models marked as judges formed the pool from which each 10-judge jury was drawn. Best-answer rate and cost per question are for the fixed policy that sends every question to that model.

Model	Identifier	Input price	Output price	Judge	Best-answer rate	Cost per question
Qwen3.5 9B	qwen/qwen3.5-9b	\$0.10	\$0.15	yes	26.5%	\$0.0010
DeepSeek V4 Flash	deepseek/deepseek-v4-flash-0731	\$0.08	\$0.18	yes	65.7%	\$0.00052
Nemotron 3.5 Lightning	nvidia/nemotron-3.5-lightning	\$0.08	\$0.20	yes	36.5%	\$0.0015
gpt-oss-120b	openai/gpt-oss-120b	\$0.05	\$0.25	yes	38.9%	\$0.00022
Gemma 4 31B	google/gemma-4-31b-it	\$0.14	\$0.40	yes	42.1%	\$0.00023
Hunyuan Hy3	tencent/hy3	\$0.132	\$0.528	yes	62.7%	\$0.00094
GPT-5.6 Luna	openai/gpt-5.6-luna	\$0.20	\$1.20	yes	85.3%	\$0.0043
MiniMax M3	minimax/minimax-m3	\$0.30	\$1.20	yes	71.7%	\$0.0044
Muse Glimmer 30B	meta/muse-glimmer-30b	\$0.35	\$1.50	yes	47.6%	\$0.0011
Gemini 3.7 Flash	google/gemini-3.7-flash	\$0.375	\$1.875	yes	72.9%	\$0.011
Kimi K2.6	moonshotai/kimi-k2.6	\$0.5415	\$2.28	yes	74.4%	\$0.015
Grok 4.3	x-ai/grok-4.3	\$1.25	\$2.50	yes	48.4%	\$0.0050
GLM-5.2	z-ai/glm-5.2	\$0.966	\$3.036	yes	71.7%	\$0.0051
Nemotron 3 Ultra 550B	nvidia/nemotron-3-ultra-550b-a55b	\$0.60	\$3.60	yes	54.5%	\$0.0069
Muse Spark 1.2	meta/muse-spark-1.2	\$1.25	\$4.25	—	64.1%	\$0.011
GLM-5.3	z-ai/glm-5.3	\$1.40	\$4.40	—	71.2%	\$0.0057
Qwen3.8 Max	qwen/qwen3.8-max	\$2.00	\$6.00	—	80.5%	\$0.016
Grok 4.6	x-ai/grok-4.6	\$2.00	\$6.00	—	65.2%	\$0.016
GPT-5.6 Sol	openai/gpt-5.6-sol	\$2.00	\$10.00	—	87.5%	\$0.079
Claude Sonnet 5	anthropic/claude-sonnet-5	\$2.00	\$10.00	—	84.0%	\$0.040

Appendix C. Terms

- **Best answer.** An answer whose jury score is within the margin (0.20 points) of the mean of the three best answers to the same question.
- **Best-answer rate (or Delta-Hit, DH).** The share of questions on which a policy's selected model gave a best answer.
- **Configuration.** One customer-selectable setting of a router product.
- **Held-out evaluation.** A policy fitted on part of the questions and scored only on the rest.

- **Hindsight limit (oracle).** The policy that, knowing all scores in advance, selects the cheapest model giving a best answer on each question.
- **Subject-informed baseline.** A fixed assignment of one model per subject area, requiring each question's subject at routing time.
- **Share to top model.** The fraction of questions a configuration sent to its most-used model.

References

1. Ong et al., *RouteLLM: Learning to Route LLMs from Preference Data*, ICLR 2025. [arXiv:2406.18665](#)
2. Chen, Zaharia and Zou, *FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance*, 2023. [arXiv:2305.05176](#)
3. Li et al., *LLMRouterBench: A Massive Benchmark and Unified Framework for LLM Routing*, Findings of ACL 2026. [arXiv:2601.07206](#)
4. Lu et al., *RouterArena: An Open Platform for Comprehensive Comparison of LLM Routers*, ICLR 2026. [arXiv:2510.00202](#)
5. Lu et al., *The Routing Plateau: Understanding the Accuracy Limits of LLM Routers*, 2026. [arXiv:2606.07587](#)
6. Lee, *Most of the LLM Routing Gap Is Task Type*, 2026. [arXiv:2608.23023](#)
7. Bhatti and Vaddina, *False Floors: LLM Safety Routing Evaluations Break Under Distribution Shift*, 2026. [arXiv:2610.01535](#)
8. Lai and Ye, *When Routing Collapses: On the Degenerate Convergence of LLM Routers*, 2026. [arXiv:2602.03478](#)
9. Kohli, *Nine Judges, Two Effective Votes: Correlated Errors Undermine LLM Evaluation Panels*, 2026. [arXiv:2605.29800](#)
10. Hossain, Yousefi and Lim, *Agreement Overstates Evidence: Error Dependence in LLM Judge Consensus*, 2026. [arXiv:2609.22512](#)
11. OpenRouter, *Auto Router: model routing* (documentation as of data collection, 30 August–1 September 2026). [openrouter.ai](#)
12. Wiggers, *Foundry Model Router Expands from Two Regions to 28, Refreshing Its Model Pool*, InfoQ, 31 August 2026. [infoq.com](#)
13. Bloomberg Government, *Stripe Clinches Over \$7 Billion Deal to Buy AI Firm OpenRouter*, 16 August 2026. [news.bgov.com](#)
14. Palo Alto Networks, *Palo Alto Networks Completes Acquisition of Portkey to Secure AI Agents*, 29 May 2026. [paloaltonetworks.com](#)
15. Loi et al., *Autobench: Automating LLM Evaluation through Reciprocal Peer Assessment*, 2025. [arXiv:2510.22593](#)
16. AWS, *Amazon Bedrock Intelligent Prompt Routing is now generally available*, 22 April 2025. [aws.amazon.com](#)