How Much Quality Survives LLM Downshifting? Judge Disagreement and the Benchmark–Reality Gap in Routing Evaluation

Anonymous authors
Paper under double-blind review

Abstract

Teams cut large language model (LLM) inference cost by routing each request to the smallest model that is good enough, instead of always calling a single frontier model. To justify the switch they quote one headline number: *"the small model preserves about X% of quality."* We show that this headline cannot be trusted. We run the most aggressive policy possible, a static 70B $\rightarrow$ 8B downshift within one open-weight family, over a 10,000-prompt curated benchmark spanning nine task types, and make four findings. (1) The cost reduction is large and *judge-independent* but *price-assumption-dependent*: token counts are essentially unchanged (total-token ratio 0.99), so projected cost falls by 98.35% (\$162.79 $\rightarrow$ \$2.68) under frontier-tier list pricing and by $\sim$ 77–80% under commodity open-weight pricing; neither figure depends on any quality judge. (2) The quality "preservation" headline is *strongly judge-dependent*: re-scoring the identical response pairs with a ten-judge cross-family panel (all open-weight, 7B–67B, and out-of-family relative to the model under test) yields per-judge rates from 61% to 97% under the fixed-order protocol (median $\approx$89.9%, mean $\approx$85%; 95% bootstrap CIs in Table 3), with only fair panel agreement (Fleiss' $\kappa = 0.37$). The single open judge behind the original headline, at 97%, sits at the top of this range, 15–23 points above the larger cross-family judges. (3) The headline is also strongly *protocol-dependent*: a blinded, order-counterbalanced re-judge of the same pairs with all ten judges finds that every judge favors whichever response is shown *second*, by 60–94 points, and that the two presentation orders agree on only 5–37% of verdicts. The midpoint of the two order-specific rates is 48–65% (median 55%), which lies 18–42 points below nine of the ten fixed-order rates; because most pairs are "preserved" in only one order, we read this as a midpoint, not an estimate of preservation. Under the blinded protocol the panel spread is 17 points versus 36 fixed-order, and the fixed-order strictness ranking is not reliably reproduced (Spearman $\rho = 0.29$, $n = 10$), so both the original headline and much of the apparent judge disagreement were position artifacts. Absent a human anchor we still cannot say which level is correct. We can say that a genuine judge disagreement of about 16 points survives de-confounding, that fixed-order judge choice moves the headline by 15–23 points across the larger cross-family judges (up to 36 if the least-reliable panelist is included), and that presentation order moves it by up to 94 points on identical outputs. (4) The headline also fails on *real* chat traffic. On a 2,000-prompt sample of WildChat, two strict cross-family judges, Qwen-32B and Gemma-27B, rate the small model "equivalent-or-better" only 21–32% of the time, although it stays "usable" $\sim$92–95% of the time (91.8–95.2%), a fixed-order rate whose order-stability we did not measure. The curated benchmark over-represents multi-step reasoning by roughly 20$\times$ relative to real traffic, and the same judges scored the same downshift far lower on real traffic than on the benchmark. Aggregate single-number preservation claims for downshifting and routing should not be trusted, including an earlier single-judge claim of our own that this paper corrects. We do not claim any protocol here recovers the true rate; without a human anchor, none is validated. What practitioners should do instead is report a judge *distribution*, with intervals, under a blinded, order-counterbalanced protocol, and validate on real traffic. These quantities are *diagnostic*. Judge spread and cross-order agreement measure whether

the evaluation is reliable enough to support an adoption decision at all, and in our data they show it is not. The data point to confidence-aware *escalation*, rather than blanket downshifting, as the policy to test next, though our single-judge frontier cannot yet show how much quality it recovers. We release the evaluation harness, prompt set, judge, cross-judge panel scripts, the blinded re-judge harness, and run logs for independent re-scoring.

1 Introduction

More and more software runs on LLM inference, and for those systems inference is usually the largest and fastest-growing variable cost. The simple thing to do is send every request to one strong model, which is wasteful. Many real requests are lookups, short extractions, summaries, or format conversions, and a model one or two orders of magnitude cheaper answers them just as well. Request-level *routing* and *cascading* take advantage of this, and a line of work has shown they cut cost substantially (Chen et al., 2023; Ong et al., 2025; Ding et al., 2024; Aggarwal et al., 2024; Hu et al., 2024).

To justify the switch, teams reach for one number: how often the cheaper model is "as good," usually scored by an LLM-as-a-judge (Zheng et al., 2023). We argue that this single number is the wrong thing to report, because it moves along two axes that a practitioner does not control: *which judge* does the scoring, and *which distribution* of prompts you test on.

We study the extreme case, so that whatever we find bounds the problem for milder policies. Our policy is a *static downshift*: replace a 70B model with an 8B model of the same family, on every request. No routing policy is more aggressive than that. The savings we measure therefore upper-bound what is on the table, and the quality loss upper-bounds what a *blanket* downshift costs; a selective router, which downshifts only some prompts, changes the scored subset and need not incur the same loss (Section 7). We then re-examine the "preservation" headline under two lenses: (a) a panel of ten judges, all out-of-family relative to the model under test, and (b) a sample of real production chat traffic. Our contributions:

1. **Cost reduction is real and judge-independent.** The 70B $\rightarrow$ 8B downshift reduces projected token cost by 98.35%, a figure that depends only on token counts and prices, not on any judge. Re-pricing the same token counts under commodity open-weight rates yields $\sim$77–80% (Section 4.1).
2. **Preservation is judge-dependent; the choice of judge is decisive.** On identical response pairs, judge-reported preservation ranges from 61.3% to 97.0% across a ten-model panel (median $\approx$ 89.9%, mean $\approx$ 85%; panel-wide Fleiss' $\kappa = 0.37$ on the thresholded verdict). The single open judge behind the original headline sits at the top (97.0%), 15–23 points above the larger cross-family judges (up to 36 across the full panel). We do not claim the lower scores are the true ones, and we flag that this 15–23-point swing is a fixed-order figure, itself partly a position artifact: the blinded re-judge shrinks the panel's spread from 35.7 to 16.6 points and does not reliably reproduce its ranking. The defensible, de-confounded judge-choice effect is the residual counterbalanced spread of about 16 points (Sections 4.3, 4.4).
3. **Position bias dominates the absolute levels, and much of the apparent judge disagreement.** Re-judging the same pairs blinded and order-counterbalanced, with all ten panel judges, shows that every judge favors whichever response is shown second, by 60–94 points, and that the two orders agree on only 5–37% of verdicts. Because most pairs are "preserved" in only one order, the two-order midpoint (which we do not read as an estimate of preservation) spans 48–65% (median 55%) and sits 18–42 points below the fixed-order rates for nine of the ten judges. Under the blinded protocol the panel's fixed-order strictness ranking is not reliably reproduced (Spearman $\rho = 0.29$, $n = 10$); the fixed-order judge spread was itself largely a position artifact (Section 4.4).
4. **Benchmarks overstate real-traffic preservation.** On a WildChat sample, two strict cross-family judges (Qwen-32B, Gemma-27B) rate the small model equivalent-or-better only 21–32% of the time, even though its outputs remain usable $\sim$ 92–95% of the time. The prompt distributions also differ sharply: real traffic is dominated by short general/chat prompts and contains $\sim$ 1.3% multi-step reasoning, versus $\sim$ 27% in the curated benchmark. Benchmark preservation did not transfer in our setting, and we cannot attribute the gap to task mix alone (Section 4.5).

A secondary contribution is methodological, and cautionary. An earlier version of this evaluation reported the single-judge 96.5% figure as the headline result, and we treat correcting our own claim as part of the contribution; we release the artifacts needed to re-score the outputs with any judge (Section 9). We are also clear about scope: this measures static downshifting, not an adaptive router. Comparing routing *algorithms*, and validating the judges against humans, are future work (Sections 7, 8).

2 Related Work

Cost-efficient inference via cascades and routing. FrugalGPT (Chen et al., 2023) introduced LLM cascades that query cheaper models first and escalate on low confidence, and a family of routers followed: RouteLLM (Ong et al., 2025) learns one from preference data, Hybrid-LLM (Ding et al., 2024) trains a quality-aware router, AutoMix (Aggarwal et al., 2024) self-verifies intermediate outputs to decide whether to escalate, and RouterBench (Hu et al., 2024) benchmarks multi-LLM routing. Model selection has also been framed as prediction from benchmark data (Shnitzer et al., 2023) and as cost–performance meta-modeling (Šakota et al., 2023). These works propose and compare *routers*, and the usual way of scoring what a router buys is a single judge-based preservation number. We, by contrast, ask the prior question of how much quality a downshift actually costs and whether that number can be trusted; we show that it is fragile. Our findings are therefore complementary: they bound the headroom and show that it has to be reported as a distribution over judges and checked on real traffic.

LLM-as-a-judge and its biases. Model-based pairwise judging is now standard (Zheng et al., 2023; Gu et al., 2024) but carries documented biases, including position (Wang et al., 2023; Koo et al., 2024), verbosity and length (Zheng et al., 2023; Dubois et al., 2024), and self-preference (Panickssery et al., 2024), alongside imperfect agreement with humans (Zheng et al., 2023; Wang et al., 2023). Prior work documents these biases in pairwise *ranking* on leaderboards; Wang et al. (2023) in particular show pairwise verdicts can flip with presentation order, and Koo et al. (2024) benchmark order effects among other evaluator biases across many judge models. We measure that effect in the downshift-adoption setting, and it is the *dominant* error term, larger than the judge-choice effect itself (Section 4.4). Others have proposed a *panel* of diverse-family judges to reduce single-judge bias (Verga et al., 2024); that line aims to improve aggregate ranking quality and human alignment. We use a cross-family panel for a different purpose: to show that a single thresholded ($\geq 3/5$) preservation rate, used as a yes/no downshift-adoption gate, is itself unstable. The scale is *bimodal*, the entire decision hinges on one threshold on it, and on identical outputs the choice of judge moves that rate by 15–23 points across the larger cross-family judges, rising to 36 across the full panel, while panel agreement is only fair, at Fleiss' $\kappa = 0.37$. RouterBench (Hu et al., 2024) evaluates routers under a fixed scoring protocol on a curated benchmark mix. We show two things about that setup. When the quality axis is single-judge preservation, that axis is itself the unstable quantity; and the curated mix diverges sharply from real traffic. Either way, single-judge preservation cannot safely gate an adoption decision.

Real-world traffic. Curated benchmarks differ from the prompts production systems actually receive. We use WildChat (Zhao et al., 2024), a corpus of real user–chatbot interactions, to show that our curated benchmark mix over-represents hard reasoning relative to real traffic and thereby overstates how well a downshift will fare in deployment.

3 Experimental Setup

Policy under test. We evaluate a static downshift: the baseline answers every prompt with `llama3.1:70b`; the routed condition answers every prompt with `llama3.1:8b`. No adaptive routing is used, isolating the quality/cost effect of model size and yielding worst-case bounds for routing.

Curated prompt set. 10,000 prompts spanning nine task types (question answering, summarization, extraction, coding, reasoning, planning, creative, structured output, long context) and three complexity tiers. The set and per-prompt task/tier labels are released (Section 9).

Preservation. Each prompt's two responses are scored pairwise on a 1–5 scale. A response is *preserved* when its score is ≥ 3, meaning the small model was judged at least adequate relative to the large model.

Judges are from *different* model families than the models under test, which reduces but does not eliminate self-preference (Panickssery et al., 2024): out-of-family status controls developer-identity self-preference, not stylistic or distributional familiarity, so a familiarity component may remain in the de-confounded spread. We did *not* blind model identity or randomize presentation order. The baseline (70B) response is always shown first and labeled as the baseline, the routed (8B) response second. A *fixed* order does not control position bias. It confounds an unknown position/identity effect with the downshift we are trying to measure. Our expectation was that an effect *common to all judges* would shift every judge's absolute level together without creating *spread* between judges, since each judge re-scores the byte-identical pairs under the same protocol. The blinded, order-counterbalanced re-judge (Section 4.4) shows we were half right. The direction holds: every judge favors the second slot. The magnitude does not. The position effect is far from constant across judges (Δ spans +60.1 to +93.7 points), so judges differ in how susceptible they are to position, and that difference generates much of the fixed-order spread by itself. Across the ten judges, the fixed-order spread of 35.7 points shrinks to 16.6 after counterbalancing (a cross-protocol comparison that also absorbs the label and rubric changes; Section 4.4), and the fixed-order strictness ranking does not carry over (Spearman $\rho = 0.29$ between fixed-order and counterbalanced rates). The panel's 36-point swing (Section 4.3) is therefore dispersion *under the fixed-order protocol*: an upper bound on judge-only dispersion, not a stable ordering of judges.

Four judging conditions. (i) *Single judge*: the original run scored the full set with one open judge, `mistral:7b`; 9,999 of 10,000 pairs returned a parseable verdict. (ii) *Cross-family panel*: we re-scored a proportionally stratified draw of the identical response pairs by task type (seed 42; target 2,000, with per-category quota rounding yielding 1,999 pairs) with ten judges, all out-of-family relative to the model under test (Mistral-7B, InternLM2-20B, Phi-3-14B, Yi-34B, Mixtral-8x7B, Aya-35B, Qwen-32B, DeepSeek-67B, Gemma-27B, Command-R-35B; two pairs share a developer, Mistral-7B/Mixtral-8x7B and Aya-35B/Command-R-35B, so the panel spans eight organizations, none of them the Llama family under test); per-judge parseable n ranges 1,971–1,999, and 1,968 prompts were scored by all ten. Two judges were also run on the full set, covering 9,996 of the 9,999 single-judge pairs; their full-set rates (80.3% and 73.8%) match their subset rates within $\sim$1.5 points. Representativeness is primarily by construction (the subset is a proportional task-type stratification of the full set, seed 42); the $\sim$1.5-point match corroborates it rather than establishing it. (iii) *Real traffic*: we sampled 2,000 prompts from WildChat (Zhao et al., 2024), ran the same 70B $\rightarrow$ 8B downshift, and judged with three judges: the lenient Mistral-7B reference plus two strict cross-family panel judges, Qwen-32B and Gemma-27B (n=1,997 joined). The mid-range panel judges were not run on real traffic, so the real-traffic gap (Section 4.5) cannot be fully separated from a strict-judge selection effect. (iv) *Blinded, order-counterbalanced re-judge*: we re-scored the same 1,999-pair benchmark subset with all ten panel judges under a protocol that removes the identity and order cues of conditions (i)–(iii): the two responses are shown as anonymous "Response 1"/"Response 2" with no baseline/routed role and no model name, under a symmetric rubric (score Response 2 relative to Response 1), and *every pair is judged twice*, once with the routed response in each slot. Averaging the two orders removes a first-order position effect; the difference between them measures it. Note this protocol changes identity labels, rubric, and order together, so only the *within-protocol* order contrast is a pure position effect; comparisons to conditions (i)–(iii) also reflect the label and rubric changes.

Cost. Token usage is metered per response. Dollar cost is a projection: we apply list-price-equivalent rates to the measured token counts, namely \$15/\$75 per million input/output tokens for the 70B class and \$0.25/\$1.25 for the 8B class. Both conditions run on self-hosted GPUs, so the dollar figures are a projection, not billed API spend. The larger caveat is the rate card itself: it is *frontier-tier*, with a $\sim 60\times$ cross-tier price ratio, which makes the 98.35% reduction a *price-ratio* effect, not a token-savings one. Input tokens are identical at 1.34M, and the routed condition actually emits slightly *fewer* output tokens, 1.879M vs. 1.903M for a total-token ratio of 0.99, so under equal per-token pricing the cost would be essentially unchanged. The headline scales with the cross-tier price gap. Re-price the same token counts under commodity open-weight hosts, a $\sim 5\times$ rather than $60\times$ 70B/8B ratio, and the reduction is $\sim$77–80%. We report 98.35% as judge-independent but price-assumption-dependent: an upper bound under frontier-tier pricing.

Hardware. All generation and judging used open-weight models served via Ollama on cloud GPU instances: a V100 instance for the 10,000-prompt downshift and single-judge run, and H100 instances for the cross-judge

Table 1: Single-judge (`mistral:7b`) preservation by complexity tier (static 70B→8B, N=9,999 judged). These are the *single-judge* numbers; Section 4.3 shows the absolute levels are judge-dependent (a cross-family panel scores the same outputs 15–23 points lower; up to 36 if the least-reliable judge is included, Section 4.3).

Tier	n	Preservation (single judge)
Low	747	99.1%
Medium	5,306	98.4%
High	3,946	93.5%
All	9,999	**96.5**%

Table 2: Single-judge (`mistral:7b`) preservation by task type. The ordering (reasoning lowest) is corroborated by the panel, but the absolute levels are judge-dependent (Section 4.3). †Long context has only n=10 and is reported for completeness, not inference.

Task type	n	Preservation (single judge)
Structured output	732	98.9%
Summarization	1,006	98.8%
Planning	988	98.8%
Creative	696	98.7%
Extraction	978	98.6%
Coding	179	98.3%
Question answering	2,695	97.6%
Reasoning	2,715	91.8%
Long context†	10	100%

panel, the real-traffic study, and the blinded order-counterbalanced re-judge (Section 4.4). No proprietary API is required to reproduce the study.

4 Results

4.1 The cost reduction is real and judge-independent

The cost side of the ledger requires no judge at all. The static 70B→8B downshift reduces projected token cost from \$162.79 to \$2.68 over the run, a 98.35% reduction. Because the figure is a deterministic function of measured token counts and fixed prices, it does *not* depend on any quality judge, and it is the one headline that survives every *judge*-robustness check below, though it remains price-assumption-dependent (Section 3). Of course, a 98% cost cut is only attractive if quality holds, and the rest of this section tests whether it does.

4.2 The single-judge "preservation" headline

Scored by the single open judge, the downshift preserves acceptable quality on 9,653 of 9,999 judged prompts (96.5%). Under this judge, preservation degrades with complexity (Table 1) and varies by task type (Table 2): near ceiling on extraction, summarization, structured output, and planning, and lowest on multi-step reasoning (91.8%). Taken at face value, the number says "the small model is basically as good." The next two subsections show otherwise.

4.3 Preservation is judge-dependent: a ten-model panel

We re-scored the *identical* response pairs with ten judges from different families (Table 3). One caveat up front: these rates were collected under the fixed-order protocol, so the table's *absolute* levels inherit the position artifact we measure in Section 4.4. Counterbalancing moves nine of the ten judges' rates down by 18–42 points and one up. What this subsection can establish is the cross-judge *dispersion under this protocol*, which shrinks by more than half, and reorders, once the artifact is removed. Reported preservation ranges from 61.3% (Command-R-35B) to 97.0% (Mistral-7B), median ≈89.9% (95% CI [88.6, 90.9]), mean ≈85%. The same outputs, in other words, are judged "preserved" anywhere from three-in-five to nearly all of the time, depending only on who scores them.

Table 3: Cross-family judge panel on the identical 70B→8B response pairs (proportionally stratified 1,999-pair subset from a target-2,000 draw, seed 42; N=1,968 scored by all ten judges; per-judge parseable n 1,971–1,999). Preservation = fraction scored $\geq 3/5$; the 95% CI is a percentile bootstrap that resamples *items*, so it is an interval for a judge's rate on this item sample, not for a population of judges (the ten are hand-picked, not sampled). Panel-wide agreement is only fair: Fleiss' $\kappa = 0.37$ (mean pairwise Cohen $\kappa = 0.39$, range 0.08–0.63). The last column (κ vs. the lenient Mistral-7B reference) is low precisely because the larger judges diverge from it. We make no claim about which end of the range is accurate (no human anchor); Command-R-35B ($\kappa = 0.08$, near-random) anchors the low end and is the least reliable point. All rates in this table are fixed-order: Section 4.4 shows counterbalancing lowers nine of the ten by 18–42 points, raises Command-R-35B by 2.5, and reorders the panel, so neither the levels nor the ranking should be read as stable judge properties.

Judge	Preservation (≥ 3)	95% CI	κ vs. Mistral-7B
Mistral-7B (lenient)	97.0%	[96.2, 97.7]	n/a
InternLM2-20B	95.2%	[94.3, 96.2]	0.39
Phi-3-14B	93.2%	[92.2, 94.4]	0.36
Yi-34B	90.3%	[89.0, 91.6]	0.32
Mixtral-8x7B	89.9%	[88.6, 91.2]	0.31
Aya-35B	89.8%	[88.5, 91.2]	0.37
Qwen-32B	81.5%	[79.9, 83.3]	0.22
DeepSeek-67B	81.5%	[79.7, 83.2]	0.22
Gemma-27B	73.8%	[72.0, 75.9]	0.12
Command-R-35B	61.3%	[59.1, 63.4]	0.08
Panel	median 89.9% [88.6, 90.9], mean 85%, range 61–97%; Fleiss' $\kappa = 0.37$		

The single open judge behind the original headline turns out to sit at the top of the distribution: Mistral-7B, at 97.0%, is the most lenient of the panel, while the larger cross-family judges score the downshift 15–23 points lower (Qwen-32B and DeepSeek-67B at 81.5%, Gemma-27B at 73.8%). We use *lenient/strict* only to describe where a judge falls in this rate distribution, *not* as a claim that the low scores are the accurate ones: without human labels we can establish dispersion, not ground truth. Close aggregate rates do not imply agreement, either. Panel-wide agreement is only fair, Fleiss' $\kappa = 0.37$ across the ten judges with mean pairwise Cohen $\kappa = 0.39$ (range 0.08–0.63), and the per-judge κ in Table 3, which measures agreement with the *lenient* Mistral reference, is generally lowest for the judges whose aggregate rates diverge most from it. Judges disagree at the per-prompt level even when aggregate rates are close, so no single judge, and no single aggregate number, faithfully represents "preservation."

Much of this judge-dependence is mechanical. For nine of the ten panel judges the underlying 1–5 pairwise scores are *bimodal*: mass concentrates at 4 (clearly preserved) and 2 (clearly worse), with the 3 bucket nearly empty (0–4%). The ≥ 3 threshold for these judges sits in a sparse region, so each rate is essentially the fraction scored ≥ 4; this is why small differences in strictness swing the headline so far. The exception that anchors the high end is the lenient Mistral-7B reference itself, which places $\sim$16% of its verdicts at exactly 3, so its 97.0% preservation exceeds its ≥ 4 fraction of 81.4% by 15.6 points.

One further caution concerns the low end. The 61.3% low point comes from Command-R-35B, whose $\kappa = 0.08$ is near-random agreement, and it is the least reliable estimate; the more defensible effect is the 15–23-point gap between the lenient judge and the larger judges. What should be reported is the distribution, with intervals. Even that is not the end of the story: as the next subsection shows, these absolute levels, and the ranking of judges behind them, are distorted by a position artifact larger than the judge-choice effect.

4.4 Position bias dominates: a blinded, order-counterbalanced re-judge

Condition (iv) re-scores the identical 1,999 benchmark pairs with all ten panel judges, blinded and in *both* presentation orders (Table 4).

First, **position bias is enormous and uniformly favors the second slot**. With identities hidden and the rubric symmetric, the same judge scores the same routed response "preserved" 83.2–99.5% of the time when it appears second, but only 4.2–33.8% of the time when it appears first. Within-judge gaps run from +60.1 points (DeepSeek-67B) to +93.7 (Phi-3-14B), and every 95% CI excludes zero by a wide margin. The per-pair picture is starker. The two orders agree on only 5.0–37.4% of verdicts, and for Phi-3-14B *the*

verdict flips with presentation order on 95% *of pairs*; under the strictest consistency reading, preserved in *both* orders, the rates are 3.6–32.5%.

Second, **the fixed-order headlines sat near the top of a wide position-induced range**. The original protocol always showed the routed response second, i.e. in the favored slot. Averaging the two orders is only a first-order correction, and we are explicit that it assumes an additive position effect the data strain: between 62.6% and 95.0% of each judge's pairs are "preserved" in exactly one of the two orders, so the average is better read as the *midpoint of two order-specific rates* than as an estimate of preservation. Under that reading the midpoint runs from 47.9% (Gemma-27B) to 64.5% (Mistral-7B), median 55.0%, and nine of the ten judges land 18–42 points below their own fixed-order rates in Table 3 (one, Command-R-35B, lands 2.5 points above). We do not present the midpoint as ground truth; without a human anchor we cannot certify any level. The assumption-free part of the finding is the pair of order-specific rates themselves (routed-first 4.2–33.8%, routed-second 83.2–99.5%), and either one puts the fixed-order headline, with the routed response always second, near the top of its own range, including for our corrected single-judge number.

A third result concerns the panel itself. The position effect, while uniformly favoring the second slot, is far from constant across judges (Δ spans 33.6 points), which means **much of the fixed-order judge disagreement was also artifact**: judges differ in position susceptibility as much as in strictness. Under the blinded symmetric-rubric protocol the panel spread is 16.6 points, against 35.7 fixed-order. That comparison changes the labels and the rubric as well as the order, so we attribute the reduction to the protocol as a whole, not to order alone; the clean order-only claim is the within-judge Δ above. A real judge disagreement of 16.6 points survives de-confounding, and 15.9 points among the larger cross-family judges alone (excluding the lenient Mistral-7B and the near-random Command-R-35B), so judge choice still matters, just far less than the fixed-order spread implied. The fixed-order strictness *ranking*, by contrast, is not reliably reproduced: the rank correlation is $\rho = 0.29$, which at $n = 10$ ($p = 0.41$) is too weak to separate from either zero or a strong positive correlation, so we claim only that the ranking does not survive intact, with Command-R-35B the clearest single inversion (apparent strictest at 61.3% fixed-order, second-most-lenient at 63.8% counterbalanced). Read the Section 4.3 panel range, then, as dispersion *under the fixed-order protocol*: some genuine judge disagreement remains after de-confounding, but neither the fixed-order levels nor the fixed-order ordering of judges are stable judge properties.

Fourth, **no absolute single-number claim survives**. On the same pairs under an identical rubric, a single judge's "preservation" rate shifts by as much as 94 points with presentation order alone: 4.2% routed-first versus 97.9% routed-second for Phi-3-14B, with the two orders spanning 4.2% to 99.5% across the panel. For the downshift-adoption decision, the practical lesson is that the quoted quantity depends on the protocol before it depends on the model.

4.5 Benchmarks overstate real-traffic preservation

Curated prompts are also not the distribution a deployed router sees. On a 2,000-prompt WildChat sample (n=1,997 judged), the same downshift looks markedly worse under the larger judges (Table 5). The lenient judge again reports high preservation (93.3%). The larger judges do not: they rate the small model "equivalent-or-better" (≥ 3) only 32.4% of the time (Qwen-32B; 95% CI $[30.3, 34.5]$) and 21.2% (Gemma-27B; $[19.4, 23.0]$). Yet the small model rarely *fails*. It is "usable" (≥ 2) 95.2% $[94.2, 96.1]$ and 91.8% $[90.6, 93.0]$ of the time respectively, and the modal large-judge score is 2 ("slightly worse but usable"): 62.7% of Qwen-32B verdicts and 70.7% of Gemma-27B verdicts. The deployment reality is *usable-but-worse*, not lossless. Even that reading rests on generous ground. These real-traffic verdicts use the original fixed-order protocol, with the routed response second, in the slot every judge favors (Section 4.4), so both the 21–32% equivalence rates and the 92–95% "usable" rates are, if anything, *optimistic* under that protocol, and order-stability of the ≥ 2 threshold is unmeasured. The blinded re-judge covered the benchmark subset only; extending it to WildChat is future work.

The two distributions also differ sharply in task mix (Table 6). Real traffic is dominated by short general/chat prompts (78.8%; median prompt $\approx$ 150 characters; $\approx$ 70% low complexity) and contains only $\approx$1.3% multi-step reasoning, whereas the curated benchmark, which we assembled from public sources, contains no general/chat bucket and $\approx$ 27% reasoning by construction, an over-representation of the hardest

Table 4: Blinded, order-counterbalanced re-judge of the identical 1,999 benchmark pairs with all ten panel judges (condition (iv); anonymous labels, symmetric rubric, each pair judged in both orders; per-judge parseable n 1,987–1,999). "Fixed order" is the same judge's rate under the original protocol (routed always second, labeled; Table 3). The counterbalanced column is the mean of a per-pair variable in $\{0, 0.5, 1\}$ (routed preserved in both / one / neither order), *not* a preservation proportion; its 95% percentile-bootstrap CI resamples items and is correspondingly tighter than a proportion CI at the same n. Δ is the within-judge position effect (routed-second minus routed-first); every Δ CI excludes zero, with half-widths ≤ 2.5 points. "Agree" is the fraction of pairs receiving the same verdict in both orders. Rows are ordered by fixed-order rate; the counterbalanced column does not preserve that order, though at $n = 10$ the rank correlation ($\rho = 0.29$) is too weak to call the reordering itself significant.

Judge	Fixed order	Routed 2nd	Routed 1st	Counterbalanced	Δ	Agree
Mistral-7B	97.0%	96.4%	32.6%	64.5% [63.4, 65.7]	+63.8	34.7%
InternLM2-20B	95.2%	99.5%	22.8%	61.2% [60.2, 62.2]	+76.7	23.2%
Phi-3-14B	93.2%	97.9%	4.2%	51.1% [50.6, 51.6]	+93.7	5.0%
Yi-34B	90.3%	96.6%	7.3%	52.0% [51.3, 52.6]	+89.4	8.7%
Mixtral-8x7B	89.9%	85.9%	15.6%	50.7% [49.6, 51.8]	+70.3	24.4%
Aya-35B	89.8%	99.4%	16.8%	58.1% [57.2, 59.0]	+82.7	17.3%
Qwen-32B	81.5%	89.7%	8.4%	49.0% [48.1, 49.9]	+81.4	16.5%
DeepSeek-67B	81.5%	93.8%	33.8%	63.8% [62.6, 65.1]	+60.1	37.4%
Gemma-27B	73.8%	83.2%	12.6%	47.9% [46.8, 49.0]	+70.7	25.6%
Command-R-35B	61.3%	99.4%	28.1%	63.8% [62.8, 64.8]	+71.3	28.7%
Panel	counterbalanced range 47.9–64.5%, median 55.0%; fixed-order span 35.7 → 16.6 points					

Table 5: Real-traffic (WildChat) downshift quality, n=1,997, with 95% bootstrap CIs. "Equivalent" = score ≥ 3; "usable" = score ≥ 2. The lenient and the larger judges disagree by 60+ points on equivalence.

Judge	"Equivalent-or-better" (≥ 3)	"Usable" (≥ 2)
Mistral-7B (lenient)	93.3% [92.2, 94.4]	n/a
Qwen-32B	32.4% [30.3, 34.5]	95.2% [94.2, 96.1]
Gemma-27B	21.2% [19.4, 23.0]	91.8% [90.6, 93.0]

category by roughly 20×. This reasoning weight is a property of our evaluation mix, not of benchmarks in general. We can document this shift but not decompose the gap into mix versus within-category effects: we did not compute per-category preservation on WildChat, and a reweighting check (applying the WildChat mix to the benchmark's per-task rates) is not well defined on our data because the dominant real-traffic category, general/chat, has no benchmark counterpart. Nor does the mix shift have an unambiguous sign a priori: the benchmark over-weights multi-step reasoning, the category on which the small model scores *worst* under the single judge (Table 2), which taken alone would make the benchmark conservative. One hypothesis consistent with the verdicts is that the larger judges penalize the small model most on open-ended conversational prompts, which the benchmark lacks entirely; we have not measured per-category strict-judge rates on either distribution, and per Section 3 the gap also cannot be fully separated from strict-judge selection, fixed-order position bias, or response-length effects. What the data establish is the weaker, decision-relevant fact: the same judges score the same downshift very differently on the two distributions, so a curated-benchmark preservation rate cannot be assumed to transfer to production traffic.

4.6 A policy frontier under the single judge, and why it hides escalation's value

To connect the finding to routing decisions, we evaluated several policies on the same 10,000-prompt run (Table 7). Cost and escalation rate are measured and judge-independent. The preservation column is not: it comes from the single lenient judge and inherits its position at the top of the panel (Section 4.3), so it is the most favorable of the panel's estimates. Read with that in mind, the pattern is suggestive. Under the lenient judge, confidence-aware escalation barely moves preservation (96.5%→96.8% across thresholds) while adding 30–60% cost, which is consistent with the lenient judge being insensitive to the quality escalation recovers. We have not shown that causally; the frontier is single-judge, and the value of escalation should be measured under the stronger judges and on real traffic (Sections 4.3, 4.5). Re-scoring this frontier under the panel is the natural follow-up (Section 8). An oracle that escalates only the prompts the small model is

Table 6: Task-mix distribution shift: curated benchmark vs. real traffic (WildChat). Real-traffic categories are from the WildChat sample; benchmark fractions are computed from Table 2 counts. The benchmark has no general/chat category and $\sim$20$\times$ the reasoning share. Selected categories shown; columns omit minor categories and do not sum to 100%.

Category	Curated benchmark	Real traffic (WildChat)
General / chat	0%	78.8%
Question answering	27.0%	7.2%
Coding	1.8%	5.1%
Creative	7.0%	5.0%
Multi-step reasoning	27.2%	1.3%

Table 7: Routing policies on the 10,000-prompt run. Cost (USD) and escalation rate are measured and judge-independent; preservation is single-judge (`mistral:7b`) and is an upper bound per Section 4.3. "Oracle" escalates only the prompts the small model is judged to fail.

Policy	Preserv. (single judge)	Cost (USD)	Escalation rate
Always strong (70B)	100.0%	162.79	100%
Always cheap (8B)	96.5%	2.68	0%
Escalation (t=0.55)	96.8%	3.64	0.7%
Escalation (t=0.85)	96.8%	4.27	1.1%
Tier router (high$\rightarrow$70B)	99.1%	71.62	39.5%
Oracle (ceiling)	100.0%	7.87	3.5%

judged to fail reaches full preservation at \$7.87 (a 3.5% escalation rate), bounding the headroom a perfect router could capture.

5 Discussion

We draw three conclusions for anyone evaluating or deploying LLM downshifting and routing.

Report cost and quality on different footings. The cost reduction is measured and judge-independent, so we can state it plainly. "Quality preservation" is not like that. It is an estimate, and its value depends on the judge, the prompt distribution, *and the judging protocol*. It should be reported as a *range over judges*, on a stated distribution, under a blinded and order-counterbalanced protocol, not as one number. A single-judge, fixed-order headline is not a safe basis for an adoption decision, especially when the judge is a small lenient open model that always sees the routed response in the favored slot. In our data that exact combination overstated counterbalanced preservation by 32 points (Section 4.4). We are careful about what this protocol is: not a validated replacement instrument. With cross-order agreement as low as 5.0%, the counterbalanced mean is a coarse summary of order-unstable verdicts (Section 7), and without a human anchor we cannot show that any protocol here estimates true preservation. Its value lies elsewhere. The quantities it adds (judge spread, interval width, the order gap, cross-order agreement) measure the evaluation's *own reliability*, and a reporting standard has to be able to return the verdict "this evidence cannot support the decision." A single fixed-order number hides exactly that evidence by construction, whereas the distributional protocol surfaces it. On our data the diagnostic comes back negative, and the practical consequence, not the counterbalanced mean, is what matters: withhold the blanket-downshift decision, or obtain human labels (Section 8). Whether any LLM-judge protocol produces a decision-grade preservation *level* is exactly what the human-anchor study must settle.

What we claim generalizes, and what we do not. The magnitudes here (the judge spread, the 60–94-point position effect, the $\sim$20$\times$ reasoning over-representation) are measurements of one model pair, one prompt set, and one traffic corpus. We do not claim they transfer (Section 7). The claim we do make beyond this setting is structural. For nine of the ten cross-family judges the pairwise scores are bimodal, with the 3 bucket nearly empty (Section 4.3). A gate defined as a threshold on that scale therefore sits in a sparse part of the distribution, which makes it hypersensitive to judge strictness and presentation order. This instability is a property of the quantity practitioners actually quote, a thresholded preservation *rate* used as a binary go/no-go, not of Llama-3.1 or of our prompts. Prior work documents these biases for judge *rankings* (Zheng et al., 2023; Wang et al., 2023). Counterbalancing order across many comparisons can leave a ranking intact

(Zheng et al., 2023), though Wang et al. (2023) show that without it even an aggregate ranking can flip. A per-pair thresholded gate has no such safety margin. That is why the same biases that move a leaderboard by a rank or two move this metric by tens of points. Any downshift evaluation that uses a single judge, in a fixed order, thresholding a bimodal scale, is exposed to the failure mode we measure, whatever the models. The protocol prescription (report a judge distribution with intervals, counterbalance order, validate on real traffic) therefore does not depend on our magnitudes recurring. What *would* need new evidence, and we state it as future work rather than claim it, is which judge level is correct (human anchoring) and whether these magnitudes recur across families and size gaps (Section 8).

Validate on real traffic as well as benchmarks. The same policy that looks near-lossless on a reasoning-heavy benchmark is usable-but-worse on real chat traffic once the larger judges score it. Benchmark task mixes are not automatically the conservative case. Here they are the opposite, once you remove the lenient judge.

Escalation is the policy to test next. The results have a consistent shape. The small model is usually *usable* ($\geq$90% under the larger judges), but it is frequently a notch worse, and worst on multi-step reasoning under the single judge (Table 2; we did not measure per-category strong-judge rates on real traffic). This pattern points to *confidence-aware escalation*. Serve the cheap model by default, and escalate only the requests where it is likely to be worse, instead of downshifting or upshifting everything. We attach two caveats, since the recommended policy is exactly what a production router would implement (Section 6). First, the "usable" ($\geq$ 2) rates come from the fixed-order, unblinded WildChat protocol with the routed response in the favored slot; the blinded re-judge (Section 4.4) covered only the benchmark subset at the ≥ 3 threshold, so the order-stability of the ≥ 2 verdicts is unmeasured, and with position effects of +60 to +94 points at ≥ 3, counterbalancing could pull "usable" below 90%. Second, the only escalation evidence here (Table 7) is single-judge, where escalation's value is invisible by construction. This is therefore a hypothesis the data *motivate*, not a demonstrated win; how much of the larger-judge gap escalation recovers, and at what latency, is the natural next experiment (Section 8).

6 A Deployed Routing System (Context)

The study above was motivated by, and feeds, a deployed inference proxy; we sketch it only to make the study's relevance concrete, not as an evaluated contribution. The proxy is OpenAI-API-compatible, so a client changes only its base URL. Per request it (1) classifies complexity and task type; (2) scores candidate models by a weighted combination of expected cost, task affinity, provider health, and latency, returning the decision and score breakdown in response headers for auditability; (3) applies confidence-aware escalation, answering with a cheap model and escalating only when low-confidence signals (hedging, refusal, truncation, repetition, finish-reason) fire; and (4) enforces governance (PII detection, region/compliance constraints, a hash-chained audit record) inside the same decision. Task affinity is refined online by an exponentially-weighted moving average over judged outcomes per (model, task type). This paper's findings apply to that loop as much as to any adoption headline: the judged signal it consumes is exactly the quantity Sections 4.3 and 4.4 show to be judge- and order-unstable. That instability changes the design in two ways. Any such online judge must itself run blinded and order-counterbalanced, scoring both presentation orders and averaging; even then, its output is at most a *relative*, per-task comparison between candidate models under one fixed judge, never an absolute quality level. Whether those relative per-task orderings are stable across judges is untested here, and is part of the head-to-head evaluation in Section 8. A second, smaller design consequence concerns seeding: the affinity prior should come from the per-task *panel distribution* of this study, not from the corrected single-judge fixed-order headline. A head-to-head evaluation of this policy against learned routers is future work (Section 8); we make no performance claim for the policy, or for the online-learning loop, here.

7 Limitations

We state these plainly because the cost headline invites over-claiming.

- **Static downshift, not adaptive routing.** This measures a fixed 70B→8B substitution. It bounds the headroom and its judge/distribution sensitivity; it does not measure any adaptive router, including ours.
- **Model-based judging; no human anchor; WildChat still fixed-order.** "Preservation" is a judge verdict, not a human one. The original protocol did not blind identity or randomize order; the blinded, order-counterbalanced re-judge (Section 4.4) measures that confound for all ten panel judges on the benchmark subset and finds it dominant (60–94 points). The WildChat verdicts and the full-set single-judge run are still fixed-order only, so their absolute levels should be read as inflated by an unmeasured (likely comparable) position effect. The counterbalanced mean assumes an additive position effect; with order agreement as low as 5.0%, even that average is a coarse summary of highly order-unstable verdicts. Fair agreement (Fleiss' $\kappa = 0.37$) means the panel median is a distributional summary, not ground truth, and no level is anchored until the human-validation study (Section 8).
- **All judges are small-to-mid open-weight models.** The panel spans 7B–67B open-weight judges served locally (Section 3); no frontier-scale proprietary judge (GPT-4-class, Claude, Gemini) was run, although such judges are the ones practitioners most often use to gate adoption decisions. Wang et al. (2023) find position bias is weaker, though still present, in stronger judges, so the magnitudes we measure (the judge-choice swing and the 60–94-point position effect) are established for open-weight judges at this scale and may be smaller under frontier judges; our quantitative effect sizes should not be assumed to transfer to that regime. The qualitative prescription (report a judge distribution, blind identities, counterbalance order, validate on real traffic) is judge-strength-agnostic and remains cheap insurance regardless; the released pairs can be re-scored with any judge, including frontier APIs (Section 9).
- **Verbosity/length uncontrolled.** A 70B and an 8B model differ systematically in output length, a documented LLM-judge confound (Zheng et al., 2023; Dubois et al., 2024); we did not residualize preservation for response length, and the blinded re-judge did not stratify by length either. Token counts are metered (for the cost analysis), so a length-stratified re-analysis remains feasible future work.
- **One model family, one size gap.** Results are for Llama-3.1 at 70B and 8B. Whether the same judge-sensitivity and task structure hold for other families and gaps is open.
- **Projected, not billed, cost.** Dollar reductions apply list-price-equivalent rates to measured token counts; they are not API invoices. The 98.35% figure is the maximum-savings ceiling of a full downshift, not a guaranteed production outcome.
- **One real-traffic source.** WildChat is one corpus of public chatbot traffic; an enterprise's distribution may differ again.
- **Authored by the system's builders.** We mitigate the conflict by releasing prompts, the judge, the panel scripts, and run logs for independent re-scoring, by foregrounding the unfavorable results (the panel range, the real-traffic strong-judge rates), and by explicitly correcting our own earlier single-judge headline.

8 Ongoing Work

Several extensions separate what we have measured from what we still intend to measure.

1. **Extend the blinded re-judge.** Section 4.4 covers all ten panel judges on the 1,999-pair benchmark subset. Extending the blinded, order-counterbalanced protocol to the WildChat real-traffic pairs and to the full 9,999-pair single-judge set would replace the paper's remaining fixed-order absolute levels with counterbalanced ones. Until those runs exist, each remaining fixed-order level carries the position-effect caveat from Section 7, and no conclusion in this paper rests on a fixed-order absolute level: the conclusions rely on the judge-independent cost result, the counterbalanced panel distribution, and the direction of the benchmark–reality gap.
2. **Routing head-to-head.** The deployed task-aware policy and confidence-aware escalation versus always-cheap, always-strong, a FrugalGPT-style cascade (Chen et al., 2023), and a reimplemented

learned router (Ong et al., 2025), on the same prompts and panel, as per-task cost–quality Pareto frontiers.

3. **Human validation.** A stratified subset scored by multiple human raters to calibrate the judges and report human-judge agreement.
4. **Escalation recovery.** How much of the strong-judge real-traffic gap confidence-aware escalation closes, and the latency it costs.
5. **Cross-family generality.** Repeating the downshift and the judge panel for additional model families and size gaps.

9 Reproducibility Statement

We release, as anonymized supplementary material (the de-anonymized repository link appears in the camera-ready): the evaluation harness and judge; the 11,364-prompt master eval set (with per-prompt task/tier labels); the downshift-generation, cross-family-panel, and real-traffic (WildChat) builder scripts; the blinded, order-counterbalanced re-judge harness of Section 4.4; a verification script that recomputes every quantitative claim in Section 4 from the raw verdicts; and the per-judge run logs. All models (`llama3.1:70b`, `llama3.1:8b`, and the open judges in Table 3) are open weights served via Ollama, so no proprietary API is needed.

The raw per-prompt verdict files are *included* in the supplementary: the full 9,999-prompt single-judge scores with both model responses, the 1,999-pair $\times$ ten-judge panel verdicts, the ten blinded re-judge verdict files (one per panel judge), the routing-frontier costs, and the score-level WildChat verdicts; running the verification script over them reproduces every reported number. We are upfront about two boundaries on the release. The 1,999-pair panel subset is released verbatim (prompt and both model responses per row), and the 10,000-prompt single-judge draw is released as verdict rows carrying both model responses; 9,996 of its 10,000 prompts reconstruct exactly by id from the released master eval set, and the remaining four (from a superseded eval-set revision) are released as responses and verdicts only. WildChat is the exception: we do *not* redistribute raw WildChat prompt or response text. Only score-level WildChat verdicts are included, and the WildChat sample itself is regenerated from the public corpus via the released builder, which streams without a pinned revision, so an independent re-run yields a statistically equivalent but not identical WildChat sample. And "preservation" remains a model-judge verdict throughout, not a human one (Section 7). We invite independent re-scoring with any judge.

10 Conclusion

A static 70B $\rightarrow$ 8B downshift cuts projected token cost by 98.35% under frontier-tier pricing. That is the one headline that is judge-independent, though, per Section 3, it scales with the cross-tier price gap. The quality story does not reduce to one number. On the same outputs, judges call the small model "preserved" anywhere from 61% to 97% of the time under the fixed-order protocol. The defensible core of that effect is the 15–23-point gap between the lenient original judge and the larger cross-family judges; the 61% endpoint comes from the least-reliable panelist (Section 4.3). Worse, a single judge's rate moves by up to 94 points (4.2% vs. 97.9% for Phi-3-14B) depending only on which response is shown second, and across the ten re-judged judges the two orders span 4.2% to 99.5% (Section 4.4). The two-order midpoint sits 18–42 points below nine of the ten fixed-order headlines, at 48–65% (median 55%), and the fixed-order ranking is not reliably reproduced; a residual judge disagreement of about 16 points survives de-confounding, so judge choice still matters, just far less than the fixed-order spread implied. And on real chat traffic the strict judges call the small model equivalent only 21–32% of the time, even in the favored slot, though usable $\sim$ 92–95%. Aggregate single-number preservation claims for downshifting and routing, including one we made ourselves, are therefore not trustworthy. The field should report judge distributions under blinded, order-counterbalanced protocols on realistic traffic. These reports do not establish the true preservation rate; what they do is reveal when no reliable rate exists, which in our data is the case. And practitioners should build routers that escalate the requests a cheap model is likely to get wrong, instead of downshifting them blindly. We release the full study for re-scoring.

Broader Impact Statement

This paper measures how evaluation-protocol choices change the reported quality of an LLM downshift; its intended impact is more honest evaluation practice for routing and downshifting decisions. One risk is misreading: the judge-independent cost figure could be quoted out of context to justify blanket downshifting, when the finding we most want carried forward is that the accompanying quality headlines are protocol artifacts, and Section 5 says outright that these results do not support blanket downshifting. On the data side, all verdicts were produced by open-weight models over public benchmark prompts and a public, consent-collected corpus (WildChat), and no new human data was collected. We see no significant risk of harm beyond these considerations.

References

Pranjal Aggarwal, Aman Madaan, Ankit Anand, Srividya Pranavi Potharaju, Swaroop Mishra, Pei Zhou, Aditya Gupta, Dheeraj Rajagopal, Karthik Kappaganthu, Yiming Yang, Shyam Upadhyay, Manaal Faruqui, and Mausam. Automix: Automatically mixing language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Ruhle, Laks V. S. Lakshmanan, and Ahmed Hassan Awadallah. Hybrid llm: Cost-efficient and quality-aware query routing. In *International Conference on Learning Representations (ICLR)*, 2024.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. In *Conference on Language Modeling (COLM)*, 2024.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. Routerbench: A benchmark for multi-llm routing system. *arXiv preprint arXiv:2403.12031*, 2024.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms with preference data. In *International Conference on Learning Representations (ICLR)*, 2025.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*, 2024.

Marija Šakota, Maxime Peyrard, and Robert West. Fly-swat or cannon? cost-effective language model choice via meta-modeling. *arXiv preprint arXiv:2308.06077*, 2023.

Tal Shnitzer, Anthony Ou, Mírian Silva, Kate Soule, Yuekai Sun, Justin Solomon, Neil Thompson, and Mikhail Yurochkin. Large language model routing with benchmark datasets. *arXiv preprint arXiv:2309.15789*, 2023.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild. In *International Conference on Learning Representations (ICLR)*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.