RANKSHIFT: DUAL-REFERENCE IDENTIFICATION OF ROUTING VALUE UNDER DELAYED FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Aggregate routing gains do not identify routing value. A positive gain may come from matching experts to states or simply from calling a globally stronger expert more often. It may also hide local losses against the deployed model. We therefore define operational routing value relative to two fixed references. The deployed anchor tests current-system improvement, and a same-budget state-independent policy tests value beyond a registered blind allocation. A single comparison need not determine both properties. RANKSHIFT authorizes a fixed proposal only when matured paired outcomes give positive lower bounds against both. Under explicit dependence and drift conditions, we describe the delay and support required for this decision. On a held-out language-model routing benchmark, one validation-frozen rule authorizes three of twelve proposal learners at the primary seed without retraining. Their gains are $+0.0224$, $+0.0227$, and $+0.0237$ with no observed local fallback regression. The other nine remain on the anchor. An independent routing trace plus delayed replay and three non-LLM domains test the same mechanism.

1 INTRODUCTION

Aggregate routing gains show the size of a change but leave its cause unclear. A positive average may arise because a router matches experts to states or because it calls a globally stronger expert more often. It may also hide losses against the deployed model on the states where switching occurs. We study when delayed paired evidence can separate these explanations while establishing improvement over the deployed system.

Proposal-learning methods do not answer this identification question directly (Chen et al., 2024a; Ong et al., 2025; Hu et al., 2024; Chuang et al., 2025). Our hypothesis is that the two properties require different comparisons. The anchor-relative margin tests improvement over the current system. A matched state-free margin tests whether the state map allocates the same candidate calls better than the registered blind allocation. Their joint rule is therefore a decision principle. The two margins estimate different properties.

Oracle evaluation sees every expert outcome after the fact. Deployment sees only outcomes that have matured. Delay, dependence, and drift can make a proposal look promising before enough evidence is available to justify a switch. Oracle headroom shows possible gain. Deployment requires evidence that a fixed proposal is ready to use.

Recent work separates oracle opportunity from signal-restricted and learned-router value using selection-valid intervals (Shihab et al., 2026). We use that distinction in a testable intervention. We freeze the proposal, wait for paired outcomes to mature, and compare it with both references before changing the action.

Two delayed environments can initially look identical even when they require opposite actions. Delayed online learning and policy inference describe the cost of waiting and the uncertainty created by late outcomes (Joulani et al., 2013; Shi et al., 2023; Zierahn et al., 2025). Selective prediction and safe improvement constrain risk relative to a baseline (Geifman & El-Yaniv, 2017; Cortes et al., 2018; Laroche et al., 2019; Joshi et al., 2024). Using only matured outcomes, we test whether a fixed state-conditioned proposal beats both the deployed fallback and a resource-matched blind allocation. The results below test this dual-reference hypothesis while keeping the proposal learner fixed.

RANKSHIFT turns this identification principle into an authorization mechanism between a fixed proposal router and model execution. Its record updates only after outcomes mature. The anchor

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

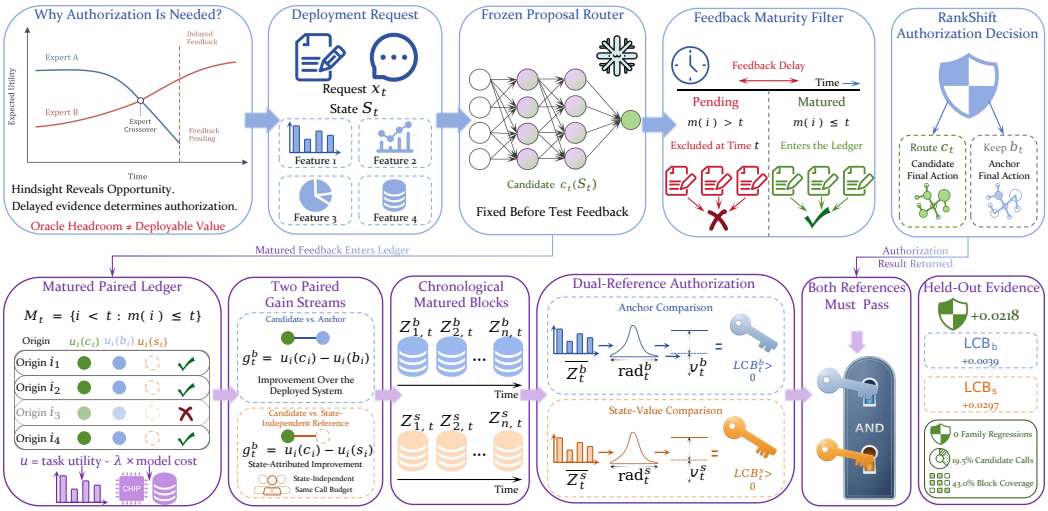

Figure 1: Why routing value requires two references. Only matured, cost-adjusted outcomes enter the paired record. Chronological blocks produce paired gain streams against the deployed anchor and a same-budget state-independent rule. The first stream tests improvement over the current system. The second tests value beyond the registered blind allocation. Positive lower bounds on both authorize the fixed unit for the next interval. Otherwise the anchor remains active.

comparison tests current-system improvement, while the resource-matched blind comparison tests state contribution relative to that comparator. The same construction applies when delayed outcomes can be linked to their original requests. Examples include code tests, interactions, conversions, and asset returns.

We first test the hypothesis on CALIBRI against modern learned routers. We then change utilities, states, and maturity clocks in finance, news, and commerce to test whether the same evidence pattern persists.

Our contributions follow the chain from ambiguity to intervention:

- **Discovery.** We identify an evaluation ambiguity. Aggregate gain is insufficient evidence of state-conditioned routing value because it mixes global expert strength with state-dependent allocation and can hide anchor regressions.
- **Theory.** We define comparator-relative routing value and construct environments with the same observed data to prove that one retained comparison need not identify both margins. We derive conditional safety and support bounds under delayed feedback.
- **Method and evidence.** We implement the principle as dual-reference authorization without changing proposal training. One frozen rule evaluates twelve proposal learners through one-reference ablations, delayed replay, support-boundary tests and cross-domain transfer. Three activate at the primary seed. Nine remain at the anchor.

2 Related Work

Routing systems. Mixtures of experts and expert tracking establish the value of specialization and switching (Jacobs et al., 1991; Freund & Schapire, 1997; Herbster & Warmuth, 1998; Dikkala et al., 2023). LLM cascades and learned routers use confidence, prompt features, or learned policies to optimize quality and cost (Ding et al., 2024; Chen et al., 2024b;a; Ong et al., 2025; Hu et al., 2024; Chuang et al., 2025; Pulishetty et al., 2025; Wang et al., 2025). These methods learn a routing proposal. They decide where to send each request. RankShift addresses the next question. It asks whether a fixed proposal adds value beyond a state-free allocation and improves the deployed system.

Evaluation ambiguity and operational identification. Classical partial identification asks what parameters are learned from incomplete information (Manski, 2003; Tamer, 2010). RankShift uses a narrower operational notion. Given fixed references (b, s) and paired outcomes linked to the original request, are both margins resolved? It makes no causal identification claim. Selection-valid tests separate oracle opportunity from realizable feature-based value (Shihab et al., 2026), while

personalization tests contextual decisions against a uniform intervention (Li & Brunskill, 2026). Both isolate contextual value. Our anchor also requires improvement over the deployed system.

Delayed feedback and safe decisions. Delayed online learning measures the price of late outcomes. Delayed policy inference handles outcomes that mature at different times (Joulani et al., 2013; Shi et al., 2023; Zierahn et al., 2025). Selective prediction, conservative exploration, and safe policy improvement constrain rejection or departure from a baseline (Geifman & El-Yaniv, 2017; Cortes et al., 2018; Mozannar & Sontag, 2020; Kazerouni et al., 2017; Wu et al., 2016; Takemura, 2023; Laroche et al., 2019; Metelli et al., 2021; Joshi et al., 2024). Conformal risk control, LEC, and TRACER provide related risk or parity controls (Angelopoulos et al., 2024; Wang et al., 2026; Rida, 2026). RankShift tests both production improvement and allocation value beyond a resource-matched blind reference.

Proposal frontiers and applications. Recent systems study cascades, dynamic model pools, and delayed agentic rewards (Dekoninck et al., 2025; Jitkrittum et al., 2026; Raj et al., 2026). Our transfer studies draw on established proposal families in finance, news, and session recommendation (Li et al., 2024; Wu et al., 2019a; Li et al., 2022; Wu et al., 2019b). Across these settings, RankShift asks whether delayed evidence separates state-conditioned allocation from global expert strength. Section 6 specifies the twelve direct router adaptations. Appendix E.5 maps their implementations and adjacent risk controls.

3 Problem Setup for State-Conditioned Value under Delayed Feedback

The empirical ambiguity above leads to a statistical question. Can delayed paired outcomes show both improvement over the current system and value beyond a matched state-free policy? We answer this question precisely before introducing the gate.

Observation and feedback setting. Before decision t, the router observes $\mathcal{M}_t = \{i < t : m(i) \leq t\}$, which contains the outcomes that have already matured. It forms $\mathcal{F}_t^-$ from these outcomes, the earlier history, and the covariates available for the current request. The utility $u_t(e) \in \mathbb{R}$ becomes observable at the outcome-independent time $m(t) \geq t$. If $m(i) = i$, the utility becomes eligible only after decision i. The state, candidate, references, confidence queries, and decisions are $\mathcal{F}_t^-$-measurable. Let $S_t = S(X_t)$ be the state used by the router. For the finite-cell theorem, we fix the state map, assignments, retained blocks, and evaluation times without using gains from the evaluation sequence. Predictable extensions require the conditions in Section 4. The candidate c_t, deployed anchor b_t, and resource-matched state-independent selector s_t induce $g_t^r = u_t(c_t) - u_t(r_t)$ for $r \in \{s, b\}$. The theorem and its analytic radius use a fixed rescaling to $[-1, 1]$. Real-domain tests retain their native units and use paired empirical intervals.

Each decision concerns a fixed *authorization unit* A. In theory $A = A_x = \{t : S_t = x\}$. Experiments may use a fixed task or family. They may also use a cell defined by router, task, and seed. Maturity, gains, references, and the decision share the same unit. Cross-unit aggregates are descriptive.

We align expert traces by request and score them only after the common outcome matures. Online collection of paired evidence is a separately budgeted shadow call, charged in delayed replay even on fallback. A deployment obtains these pairs from a calibration trace or a pre-registered probing schedule. Until the first paired block matures, its LCB is $-\infty$ and the anchor remains active. The static CALIBRI test charges only the selected model (Section 6). Domain clocks and accounting are in Appendix E.

The state-independent selector s_t uses the same expert pool and call budget as c_t, but allocates calls without observing S_t. Thus g_t^s measures improvement over this fixed state-blind reference, while g_t^b measures improvement over the deployed fallback. The same controller can test additional learned state-free rules when they are fixed before the evaluation outcomes are observed.

The meaning of Gate 2 depends on how s is defined. Composition-matched and optimized state-free references support allocation claims at different strengths. An ordinary blind rule supports only an advantage over that comparator. When $s = b$, Gate 2 collapses into anchor-risk control. If s is known to dominate b on the same unit, a positive margin against s also implies anchor improvement. Appendix D and Proposition 3 formalize this taxonomy.

For candidate-call budget B, let $\mathcal{G} = \sigma(S)$ denote the fixed, time-independent state information, $\Pi_{\mathcal{G}}$ the policies using it, and $\Pi_{\varnothing}$ state-free policies with the same expert pool and budget. Let $U(\pi)$ be expected cost-adjusted utility and $C(\pi)$ the candidate-call, token, latency, or compute cost. Define

$$V_B(\mathcal{G}) = \sup_{\pi \in \Pi_{\mathcal{G}} : C(\pi) \le B} U(\pi) - \sup_{\pi \in \Pi_{\varnothing} : C(\pi) \le B} U(\pi). \tag{1}$$

$V_B(\mathcal{G})$ is an idealized (oracle) policy-class benchmark. The gate compares fixed proposals $c \in \Pi_{\mathcal{G}}$ with state-free policies registered before evaluation. Its claim is comparator-relative unless s attains the population frontier. For unit A_x at look ℓ, define the conservative margin

$$I_{x,\ell} = \min_{r \in \{s,b\}} \left\{ \mu_{\mathrm{cal}}^{x,r,\ell} - v_{x,\ell}^{r} \right\}, \tag{2}$$

where $\mu_{\mathrm{cal}}^{x,r,\ell}$ is the conditional calibration-block mean (Appendix B.4) and $v_{x,\ell}^{r} \ge 0$ bounds deterioration until the next look. Positive $I_{x,\ell}$ defines *dual-reference routing value* under this drift condition. It is relative to (b, s) and does not identify causal effects or the oracle value $V_B(\mathcal{G})$.

Lower-bound decision rule. At each fixed look, for cell x and reference r, let $Z_{j,t}^{x,r} \in [-1, 1]$ be the retained paired-gain mean in block j, $n_t^{x,r}$ its effective matured-block count, and $\overline{Z}_t^{x,r} = (n_t^{x,r})^{-1} \sum_j Z_{j,t}^{x,r}$. A zero count gives $-\infty$ and forces the fallback. Otherwise compute

$$\mathrm{LCB}_t^{x,r} = \overline{Z}_t^{x,r} - \sqrt{\frac{2\log(1/\alpha)}{n_t^{x,r}}} - v_t^{x,r}, \qquad \alpha = \frac{\delta_{\mathrm{stat}}}{2KM}, \tag{3}$$

where $x_t = S_t$ is the current state-cell index, $\ell(t)$ is the fixed confidence-look index, and $v_t^{x,r} = v_{x,\ell(t)}^{r} \ge 0$ is the debit for cell x and look. Here K is the number of state cells, M is the number of looks, and δ_{stat} is the statistical error budget. The method authorizes the candidate only when both reference-specific bounds are strictly positive.

4 THEORY

Theorem 1 is the central identification result. With one reference unobserved, both margins need not be identifiable. The remaining results give conditional safety, delay limits, and sufficient support for authorization.

4.1 WHY ROUTING VALUE REQUIRES TWO REFERENCES

Theorem 1 (Single-reference impossibility for routing-value identification)**.** Fix $\eta \in (0, 1/3]$ and distinct equal-cost experts b, e_0, e_1. Let $S \sim \mathrm{Bernoulli}(1/2)$, $c(S) = e_S$, and s draw uniformly from $\{e_0, e_1\}$ independently of S. Thus c, s share a marginal mixture and one-call budget. For either $r \in \{b, s\}$, consider rules observing only independent samples of $(S, c, r, u(c), u(r))$. Two zero-drift environments have the same observed law and positive observed mean gain. The omitted mean gain is positive in one and nonpositive in the other. False-authorization control at level α in the latter limits power to α in the former.

Proof. For the anchor-only pair, set $u(c, S) = 2\eta$ and $u(b, S) = \eta$ in both environments, but set the mismatched-expert utility to 0 and 2η, respectively. The observed stream is identical, while the omitted state-free margin is η and 0. For the state-free-only pair, set $u(e_S, S) = 2\eta$ and $u(e_{1-S}, S) = 0$ in both environments, but set the unobserved anchor utility to η and 3η. The retained joint law is identical, while the omitted anchor margin is η and $-\eta$. All utilities lie in $[0, 1]$. Therefore no statistic or sequential rule based only on the retained stream distinguishes either pair.

The pairs show that anchor improvement and allocation value can fail separately. The result concerns missing reference information. One dataset may support both comparisons when it observes all relevant experts.

The CSI500 ablation (Yang et al., 2020) in Appendix D.2 exhibits the anchor-only failure mode empirically. Gate 1 authorizes five seeds, three without positive dual-reference limiting mean, whereas the dual mean gate retains two with none.

4.2 Delayed feedback creates an information constraint

Theorem 2 (Delayed indistinguishability)**.** For every causal policy, a two-action construction with utilities in $[0, 1]$ and delay $D \geq 1$ gives two environments with identical first-D histories and opposite per-round optima. One incurs expected dynamic regret at least $D/2$.

The proof uses two actions. If p_t is the common probability of choosing the first action at time t, the two prefix regrets are $\sum_{t=1}^{D}(1 - p_t)$ and $\sum_{t=1}^{D} p_t$. Their sum is D.

Corollary 1 (Safety and oracle regret incompatibility)**.** Consider two observationally indistinguishable environments over T decisions. A policy chooses between a fixed fallback and a candidate with gains $+\gamma$ and $-\gamma$, where $0 < \gamma \leq 1$. If harm in the latter is at most ε, the policy has at most ε/γ expected activations. Its beneficial-environment regret is at least $\gamma T - \varepsilon$. Thus, a uniform fallback guarantee without identifying information forces fallback deferral and incurs linear regret against the beneficial candidate.

4.3 Authorization guarantees under delayed feedback

Theorem 3 (Dual-reference selective safety)**.** Over a finite horizon T, let $x \in [K]$, $r \in \{s, b\}$, and let $\tau_1 < \cdots < \tau_M$ be gain-free evaluation times. Set $\tau_{M+1} = T + 1$ and $\mathcal{T}_\ell = [\tau_\ell, \tau_{\ell+1})$. At each tested (x, r, ℓ), let $n^r_{x,\ell} \geq 1$ be the matured block count, $\bar{Z}^r_{x,\ell}$ its mean bounded gain, $\mu^{x,r,\ell}_{\mathrm{cal}}$ its calibration mean, and N a deterministic upper bound on all such counts. Zero blocks are assigned lower bound $-\infty$ by the operational fallback rule. Choose cells, assignments, retained blocks, and looks without evaluation-sequence gains. Assume conditional coupling with absolute regularity envelope $\bar{\beta}(q)$ at block gap q after this gain-free selection. With probability at least $1 - \delta_{\mathrm{drift}}$, assume, simultaneously for $t \in \mathcal{T}_\ell$ with $S_t = x$, $\mathbb{E}[g^r_t \mid \mathcal{F}^-_t] \geq \mu^{x,r,\ell}_{\mathrm{cal}} - v^r_{x,\ell}$, where $v^r_{x,\ell} \geq 0$ is a pre-registered, gain-free interval-uniform debit. If $2KM(N-1)\bar{\beta}(q) \leq \delta_{\mathrm{mix}}$ and $\delta_{\mathrm{stat}} > 0$, $\delta_{\mathrm{mix}}, \delta_{\mathrm{drift}} \geq 0$ satisfy $\delta_{\mathrm{stat}} + \delta_{\mathrm{mix}} + \delta_{\mathrm{drift}} = \delta \in (0, 1)$, then, with probability at least $1 - \delta$, every authorization set at a look and executed in its governed interval has nonnegative current conditional expected gain against both references under the gate in Equation 3.

The square-root term in Equation 3 measures uncertainty from limited matured support and a finite set of evaluation times. A larger block gap q reduces dependence, but it also reduces the effective sample size. The debit $v^r_{x,\ell}$ accounts for possible deterioration between calibration and deployment. The proof couples separated blocks and applies Hoeffding concentration with a finite union bound (Berbee, 1979; Yu, 1994; Hoeffding, 1963). Theorem 3 is a conditional design result. When a dependence envelope and drift debit are registered, the gate controls current conditional gain. Natural-data studies use the same two comparisons with paired empirical intervals. They test the mechanism without asserting analytic coverage. Appendix B.5.2 tests bounded pool-mean concentration and a known contamination debit using CALIBRI utility reservoirs. This limited audit does not establish per-request conditional safety. The tail statistic remains a separate empirical check, and Appendix B.4 gives the complete proof.

Proposition 1 (Activation boundary)**.** For K fixed state cells, suppose every tested cell/reference/look has at least one matured block, and suppose $\mu^{x,r,\ell}_{\mathrm{cal}} - v^r_{x,\ell} \geq \gamma > 0$ for both references. Assume the conditional coupling condition of Theorem 3 and a deterministic interval-uniform drift envelope. Then, with probability at least $1 - (\delta_{\mathrm{stat}} + \delta_{\mathrm{mix}})$, every fixed test has a strictly positive LCB once its effective matured support per cell, reference, and look obeys

$$n_{\mathrm{eff}} := \min_{x,r,\ell} n^r_{x,\ell} > \frac{8 \log(2MK/\delta_{\mathrm{stat}})}{\gamma^2}. \tag{4}$$

Conversely, for $\gamma \in (0, 1/4]$ and error probability $\delta \in (0, 1/8]$, even without state complexity or delay, uniformly distinguishing bounded margins $+\gamma$ and $-\gamma$ requires $\Omega(\log(1/\delta)/\gamma^2)$ observations.

The factor 8 accounts for two confidence radii. The first controls estimation error, and the second is subtracted in the LCB. This sufficient support grows as $\log K/\gamma^2$, so fine state partitions delay authorization most when the identifiable margin is small.

For persistent margin $\gamma_x > 0$, define

$$N_x = \left\lfloor \frac{8 \log(2MK/\delta_{\mathrm{stat}})}{\gamma_x^2} \right\rfloor + 1. \tag{5}$$

Algorithm 1 RANKSHIFT decision rule under delayed feedback

Require: Fixed $(c, s, b, S, m, v, q, K, M, \delta_{\mathrm{stat}})$, origin covariates X_t, gain-free blocks separated by q reveal/order indices, looks $\tau_1{:}\tau_M$, and $A_0(x) = 0$

At look $t = \tau_\ell$, set $\mathcal{M}_t \leftarrow \{i < t : m(i) \le t\}$ from matured records only
for $x \in \{1, \dots, K\}$ **do**
  for $r \in \{s, b\}$ **do**
    Restrict $i \in \mathcal{M}_t$ to $S_i = x$ and align $g_i^r = u_i(c_i) - u_i(r_i)$. Retain at gap q and aggregate into $Z_{j,t}^{x,r}$
    If $n_t^{x,r} = 0$, set $\mathrm{LCB}_t^{x,r} \leftarrow -\infty$. Otherwise set $\mathrm{LCB}_t^{x,r} \leftarrow \overline{Z}_t^{x,r} - \sqrt{\frac{2\log(2KM/\delta_{\mathrm{stat}})}{n_t^{x,r}}} - v_t^{x,r}$
  $A_\ell(x) \leftarrow \mathbf{1}\{\min(\mathrm{LCB}_t^{x,s}, \mathrm{LCB}_t^{x,b}) > 0\}$
For $t' \in [\tau_\ell, \tau_{\ell+1})$, set $x_{t'} \leftarrow S(X_{t'})$ and obtain $(c_{t'}, b_{t'})$
if $A_\ell(x_{t'}) = 1$ **then**
  return ROUTE$(c_{t'})$ ▷ persists to the next look
else if the fallback anchor is available **then**
  return ANCHOR$(b_{t'})$ ▷ fallback deferral
else
  return SYSTEM-ABSTAIN ▷ no output, retained in all denominators

Append an outcome only at maturity for the next look and never revise an earlier decision.

Under the activation assumptions with zero debit, suppose calibration margins stay above γ_x. One effective block arrives every q_x visits with delay at most D_x. If a registered look occurs within G_x visits after support reaches N_x and within the horizon, activation occurs by $q_x N_x + D_x + G_x$ on the joint event. See Appendix B.7. Appendix B.6 gives a linear-controller extension under a fixed linear model and a conditional batch-MGF assumption (Abbasi-Yadkori et al., 2011). Appendix B.8 gives the regret decomposition.

Designing the state map. The state map should use the coarsest features available at deployment that still preserve stable differences in gain. States that are too fine split the matured evidence into small groups. States that are too coarse average away useful crossover. The main evidence comes from wrapping frozen learned proposals without changing their features, scores, or training. RankShift-Code provides an interpretable metadata-aware illustration of the principle. Benchmark and language labels select LiveCodeBench (Jain et al., 2025) and four fixed McEval (Chai et al., 2025) languages, while MultiPL-E and all other states use the anchor. Continuous states may instead use a fixed low-dimensional feature map. Any refinement must satisfy the activation boundary using gain-free support information. Outcome-adaptive state construction requires a separate simultaneous confidence argument.

5 FROM DUAL-REFERENCE IDENTIFICATION TO AUTHORIZATION

The ambiguity arises in evaluation. RankShift therefore leaves the router and expert pool unchanged and adds an identification layer before deployment. Two tests reject the distinct failure modes before a proposed switch is executed.

Test the dual-reference hypothesis. The candidate c and state map S are fixed on a development prefix. The deployed system supplies anchor b, while selector s spends the same candidate budget without observing S. Thus g^b tests improvement over the current system and g^s tests whether state allocates calls better than a blind rule. Theorem 1 shows why both are needed.

Turn the comparisons into a selective policy. RankShift aligns candidate and reference utilities by origin and charges costs. It adds each paired gain to the record only after its outcome matures. At each evaluation time fixed in advance, it computes one lower bound per reference. Both must be strictly positive before the candidate is used. Otherwise the anchor remains active. Mature cells with a clear margin therefore activate. Unresolved cells retain the existing system. A future outcome can change a future decision, but never an earlier one.

At a checkpoint τ_ℓ, a cell passing both strict tests remains authorized on $[\tau_\ell, \tau_{\ell+1})$ until the next look. Theorem 3's interval-uniform drift envelope justifies this carry-forward. The analytic finite-cell rule in Algorithm 1 and the paired block-bootstrap rule in the 35-unit study implement the same two comparisons. The empirical version replaces the Hoeffding radius with paired bootstrap bounds. We call this formal two-mean rule *RankShift-Mean*. The headline empirical wrapper uses *RankShift-Full*, which also requires nonnegative worst-family gain fixed on validation. This tail screen is an

extra empirical safety condition outside Theorem 3. Table 3 reports the mean-only and full rules separately.

Algorithm 1 supplies the common deployment interface used in theory and replay. Across domains, only the expert pool, state, utility, and maturity clock change. Every uncovered origin executes the anchor and remains in the reported policy value. This common interface is what allows the same authorization principle to operate across model routing, ranking, and recommendation.

6 Experimental Setup

The evaluation follows the hypothesis in three steps. We first diagnose whether full-budget proposal gains pass both references. We then hold proposal scores fixed and test whether selective authorization removes the failures. Finally, we vary support, utility, and maturity clocks to test the predicted activation boundary and transfer of the authorization mechanism.

CALIBRI (Campos et al., 2025) is the head-to-head router test. LLMRouterBench (Li et al., 2026) is public multi-model transfer. Finance uses Qlib (Yang et al., 2020) and a fixed Binance market-data snapshot. News uses MIND and EB-NeRD (Wu et al., 2020; Kruse et al., 2024). Commerce uses the Coveo Data Challenge release and Amazon-M2 (Tagliabue et al., 2021; Jin et al., 2023). CALIBRI combines LiveCodeBench, McEval, and MultiPL-E (Jain et al., 2025; Chai et al., 2025; Cassano et al., 2023). All decisions use matured outcomes, aligned blocks, and domain-specific costs.

Random and state-free controls test selection. The twelve learned proposals are RouteLLM, RouterDC, Avengers-Pro, EmbedLLM, GraphRouter, Meta-Router, HybridLLM, CSCR, FrugalGPT, LENS, DiSRouter, and StageRoute (Ong et al., 2025; Chen et al., 2024b; Zhang et al., 2025; Zhuang et al., 2025; Feng et al., 2025; Zhang et al., 2026; Ding et al., 2024; Shirkavand et al., 2025; Chen et al., 2024a; Liu et al., 2026; Zheng et al., 2026; Li & Li, 2026) share one two-model interface, split, frozen BGE-M3 cache, 315-call budget, cost proxy, and paired test. Every primary row uses seed 2027. Five-seed sensitivity is stored for eleven adapters. RouterDC is primary-seed only. RankShift-Code is separately disclosed as a metadata heuristic. It lies outside the feature-matched learned-baseline comparison. For every unwrapped score vector, the offline held-out check applies a stable $(-\text{score}, \text{row index})$ ordering and selects exactly 315 rows. Held-out outcomes are read only after that mask is fixed. The resulting quantity is an offline proposal check. The wrapper experiment fixes proposal scores and freezes authorization from matured validation evidence before the untouched test split is evaluated. A live service can implement the same contract with a threshold or quota fixed on a deployment validation stream, while the wrapper may retain fewer calls. Appendix E.5 maps the implementations and transfer controls.

Uncertainty and hyperparameters. The primary CALIBRI seed uses 50,000 paired problem-block resamples, whereas the 35-unit study uses 10,000. LLMRouterBench sensitivity uses seeds $0{:}4$, while other domains retain chronological hierarchical or moving blocks. We report paired intervals, tail gain, positive-block fraction, value capture, and runtime. Pools, rules, and hyperparameters use development data, and blocks match feedback maturity.

Shadow-call accounting. Online collection of paired candidate outcomes requires a shadow-probing budget separate from deployed calls. Our traces are paired. Source-order CALIBRI replay remains net positive after charging all 1,300 shadow calls (Appendix E.4).

LLM protocol. RankShift-Code fixes Qwen3-Coder-30B-A3B-Instruct as anchor and GPT-OSS-20B as candidate. Python reads benchmark and language labels directly. LiveCodeBench and McEval $\{$Perl, Python, Rust, Scala$\}$ use the candidate. MultiPL-E and all other cells fall back. It therefore does not use BGE-M3. Utility is correctness minus 10^{-5} times selected tokens. With two actions, matching 315 candidate calls fixes the full marginal action mix and removes only the request-to-action association. Ten thousand uniform subsets measure allocation variability, and 50,000 stratified block bootstraps use native paired gains (Campos et al., 2025). The gate hot path has median $0.240\,\mu$s batch-amortized latency. The generator varies $D \in \{0, 64, 128, 256\}$. Replay tests permutation, always-on, and support controls. Appendix E gives the full protocol.

7 Results

Table 1 diagnoses the ambiguity. Table 2 isolates authorization with frozen proposals. Table 3 tests both references. Table 4 tests transfer across utilities and feedback clocks.

Table 1: Positive aggregate gain does not establish routing value. Eleven of twelve proposal means are positive and every state-free lower bound is positive, yet every LCB_b is negative (bold). Here b denotes the deployed anchor and s the registered state-free comparator.

Proposal	Mean	$\mathbf{LCB}_b$	LCB_s	Proposal	Mean	$\mathbf{LCB}_b$	LCB_s
CSCR	+.0165	**−.0006**	+.0249	GraphRouter	+.0153	**−.0014**	+.0240
LENS	+.0167	**−.0003**	+.0252	RouteLLM-MF	+.0094	**−.0090**	+.0173
StageRoute	+.0140	**−.0029**	+.0232	Avengers-Pro	+.0115	**−.0067**	+.0189
EmbedLLM-KNN	+.0142	**−.0037**	+.0224	FrugalGPT	−.0067	**−.0168**	+.0067
HybridLLM	+.0065	**−.0094**	+.0173	DiSRouter	+.0170	**−.0003**	+.0255
Meta-Router	+.0171	**−.0003**	+.0254	RouterDC	+.0161	**−.0010**	+.0246

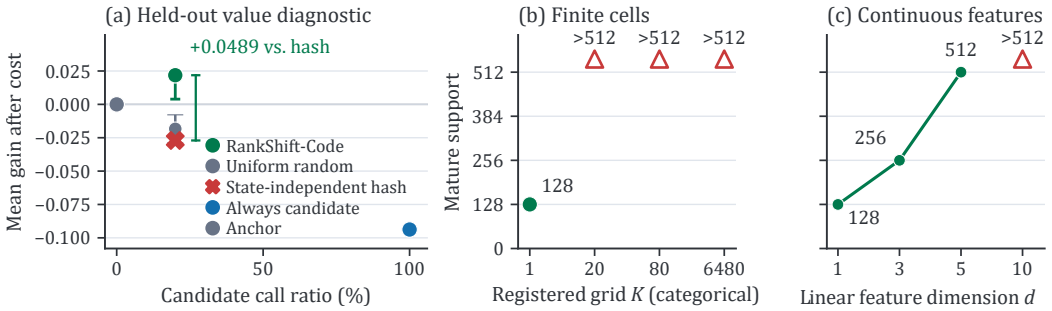

Figure 2: CALIBRI policy value and the authorization boundary. (a) Held-out mean gain after cost is plotted against the candidate-call ratio. The whisker shows the paired uncertainty interval for RankShift-Code. (b,c) Minimum tested mature support needed to reach 80% activation power. > 512 means that no tested support through 512 met the criterion. K is categorical in (b), so equal spacing is intentional. RankShift-Code is an interpretable illustration. Table 2 tests the main authorization claim on frozen learned proposals.

7.1 Positive mean gains can coexist with local regressions

We first test this prediction on fixed 315-call proposals. Eleven of twelve learned rows have positive means, and every row clears the state-free bound, yet every row fails the anchor bound. Mean gain therefore does not establish dual-reference routing value relative to the registered comparators. Each learned row is reused as a frozen proposal by the wrapper below.
Figure 2 summarizes the support boundary predicted by Proposition 1. Richer state representations require more mature evidence before RankShift allows a switch.

The anchor bound tests current-system improvement. The blind bound tests whether gain persists beyond the registered state-free comparator. The worst-family check catches local regression.

The independent routing-trace transfer and task-level decisions appear in Appendix E.6.

Wrapper test of the dual-reference hypothesis. Table 2 holds proposal scores and training fixed and changes only which states are allowed to execute. One pass/fail rule frozen on validation tests both bounds and the family condition for all twelve. The untouched test split is evaluated once after those decisions are fixed. At the primary seed, three rows pass without retraining, and the other nine remain explicit zero-call fallbacks (Appendix Table 13).

Table 2: Frozen-proposal wrapper test on held-out CALIBRI. Authorization masks are fixed on validation. Status reports held-out qualification. All numbers use untouched test data. Appendix Table 13 gives the complete ledger.

Proposal	Variant	Calls	Mean	Worst	LCB_b	LCB_s	Test status
CSCR	Proposal	315	+.0165	−.0103	−.0006	+.0249	NQ
	+ RankShift	213	**+.0224**	**.0000**	**+.0079**	**+.0241**	**Qualified**
LENS	Proposal	315	+.0167	−.0103	−.0003	+.0252	NQ
	+ RankShift	214	**+.0227**	**.0000**	**+.0081**	**+.0243**	**Qualified**
StageRoute	Proposal	315	+.0140	−.0166	−.0029	+.0232	NQ
	+ RankShift	211	**+.0237**	**.0000**	**+.0095**	**+.0257**	**Qualified**

7.2 One-reference ablations test distinct failure modes

Table 3 holds the paired-block bounds and frozen proposals fixed while removing one gate. It tests whether either comparison can replace the joint rule.

Table 3: Dual-gate ablation on 35 development units. Nonpositive counts authorized units with a nonpositive limiting mean. C_{DR} is retained positive dual-reference value.

Variant	Authorized	Nonpositive	$\mathbf{C}_{\mathrm{DR}}$
Always-on	$35/35$	$12/35$	1.000
Gate 1 only, anchor LCB	$32/35$	$9/35$	1.000
Gate 2 only, state-independent LCB	$19/35$	$0/35$	0.973
Dual Gate, mean-only	$\mathbf{19/35}$	$\mathbf{0/35}$	$\mathbf{0.973}$
Dual Gate + tail screen	$\mathbf{7/35}$	$\mathbf{0/35}$	$\mathbf{0.686}$
Always-fallback	$0/35$	$0/35$	0.000

The single-reference checks differ across these data. Gate 2 removes nine anchor-only activations with nonpositive observed limiting means in the 35-unit study. In Appendix D.3, 46 of 420 cells pass the state-free test but fail anchor certification. The dual rule rejects all 46. Eight corresponding ungated proposals lose to the anchor on untouched test data. Although Gate 2 and the dual mean rule coincide in the 35-unit study, Theorem 1 explains the need for anchor information in general. Appendix Figure 3 visualizes the resulting safety and value frontier.

7.3 Transfer test across utilities and feedback clocks

Finance, news, and commerce vary utilities and feedback clocks (Table 4). Comparator strength determines interpretation as specified in Section 3. When $s_t = b_t$, as in LLMRouterBench and Amazon-M2, the gates coincide. GSM8K and financial ablations isolate Gate 2's attribution role (Appendix D.2).

Table 4: Positive paired routing opportunities across domains. Finance gains use 10^{-4} RankIC units. News and commerce gains are percentage-point NDCG/MRR changes.

Family	Dataset	Direct comparator	Gain [Paired Interval]	Evidence
Finance	CSI300	registered anchor	$+\mathbf{19.3}$ $[13.2, 24.9]$	CI > 0
Finance	CSI500	registered anchor	$+\mathbf{37.3}$ $[30.9, 44.9]$	$\mathbf{4/4}$ cost CIs > 0
Finance	Binance	registered anchor	$+\mathbf{17.6}$ $[12.8, 22.7]$	CI > 0
News	MIND	state-free GBDT	$+\mathbf{1.281}$ $[1.015, 1.544]$	$\mathbf{5/5}$ seeds > 0
News	MIND	Adapted MINER	$+\mathbf{6.622}$ $[6.233, 6.995]$	tail gain $+5.467$
News	EB-NeRD	strongest control	$+\mathbf{0.153}$ $[0.062, 0.228]$	$\mathbf{5/5}$ seeds > 0
Commerce	Coveo	SR-GNN anchor	$+\mathbf{2.395}$ $[2.358, 2.425]$	$\mathbf{5/5}$ seeds > 0
Commerce	Amazon-M2	best fixed	$+\mathbf{5.635}$ (LCB $+\mathbf{5.082}$)	$\mathbf{6/6}$ locales ≥ 0

8 Conclusion

A positive aggregate estimate alone does not certify local anchor improvement or comparator-relative allocation value. RankShift tests both with fixed references and matured paired outcomes. The central result shows that one retained comparison can leave the other margin unidentified. In held-out CALIBRI, the resulting authorization rule qualifies $3/12$ frozen proposals and reaches $+0.0237$ without retraining. One-reference checks expose both concerns. Anchor-only control admits nonpositive observed allocation gains. Some state-free-only activations lose to the anchor on held-out data. Support-boundary tests and delayed replay further connect activation to the amount and timing of evidence.

The broader lesson is that routing should be evaluated by identifying value before optimizing deployment. RankShift applies this lesson without changing proposal training. It registers an anchor and a claim-matched state-free comparator. It preserves the link between decisions and delayed outcomes, then authorizes only after both margins resolve. Its conclusion is relative to the chosen comparator. Stronger claims require a stronger registered state-free frontier. The same principle may extend to adaptive systems whose delayed outcomes can be linked to their original decisions. In such settings, evidence about the source of gains should precede deployment changes. Future work can reduce shadow-probing cost and adapt state partitions while preserving the same identification target.

AI USE STATEMENT

We used generative AI tools for language editing, LaTeX and figure formatting, and the retrieval and discovery of related work.

REPRODUCIBILITY STATEMENT

Section 6 specifies the evaluation settings, costs, maturity clocks, baselines, and statistical procedures. The appendix gives the assumptions, complete proofs, implementation details, protocols, random seeds, and unit-level results. The anonymous supplementary package contains the RankShift implementation, fixed configurations, data-preparation instructions, archived result ledgers, and scripts that verify Tables 1–3 and the public RouterBench MMLU (Hendrycks et al., 2021)/GSM8K (Cobbe et al., 2021) extension. Bundled checks replay or reconstruct the stored evidence. Fresh raw-data reruns require the documented external datasets, caches, embeddings, repositories, or checkpoints, which are not redistributed under third-party licenses. Upon acceptance, we will release the submitted first-party code, configurations, and documentation publicly.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2312–2320, 2011.

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*, volume 37, pp. 131000–131034, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

H. C. P. Berbee. *Random Walks with Stationary Increments and Renewal Theory*, volume 112 of *Mathematical Centre Tracts*. Mathematisch Centrum, Amsterdam, 1979.

Viola Campos, Robin Kuschnereit, and Adrian Ulges. Multicalibration for LLM-based code generation. *arXiv preprint arXiv:2512.08810*, 2025.

Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q. Feldman, Arjun Guha, Michael Greenberg, and Abhinav Jangda. MultiPL-E: A scalable and polyglot approach to benchmarking neural code generation. *IEEE Transactions on Software Engineering*, 49(7):3675–3691, 2023.

Linzheng Chai, Shukai Liu, Jian Yang, Yuwei Yin, Ke Jin, Jiaheng Liu, Tao Sun, Ge Zhang, Changyu Ren, Hongcheng Guo, Zekun Wang, Boyang Wang, Xianjie Wu, Bing Wang, Tongliang Li, Liqun Yang, Sufeng Duan, and Zhoujun Li. McEval: Massively multilingual code evaluation. In *International Conference on Learning Representations*, 2025.

Lingjiao Chen, Matei Zaharia, and James Zou. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024a.

Shuhao Chen, Weisen Jiang, Baijiong Lin, James T. Kwok, and Yu Zhang. RouterDC: Query-based router by dual contrastive learning for assembling large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 66305–66328, 2024b.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8359–8388, 2024.

Yu-Neng Chuang, Prathusha Kameswara Sarma, Parikshit Gopalan, John Boccio, Sara Bolouki, Xia Hu, and Helen Zhou. Learning to route LLMs with confidence tokens. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10859–10878, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Scott Yang. Online learning with abstention. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1059–1067, 2018.

Jasper Dekoninck, Maximilian Baader, and Martin Vechev. A unified approach to routing and cascading for LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 12987–13010, 2025.

Nishanth Dikkala, Nikhil Ghosh, Raghu Meka, Rina Panigrahy, Nikhil Vyas, and Xin Wang. On the benefits of learning to route in mixture-of-experts models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9376–9396, 2023.

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

Tao Feng, Yanzhen Shen, and Jiaxuan You. GraphRouter: A graph-based router for LLM selections. In *International Conference on Learning Representations*, 2025.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Mark Herbster and Manfred K Warmuth. Tracking the best expert. *Machine Learning*, 32(2):151–178, 1998.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. In *Proceedings of the 4th International Conference on Learning Representations*, 2016.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

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.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida I. Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, 2025.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 14165–14178, 2023.

Wei Jin, Haitao Mao, Zheng Li, Haoming Jiang, Chen Luo, Hongzhi Wen, Haoyu Han, Hanqing Lu, Zhengyang Wang, Ruirui Li, Zhen Li, Monica Xiao Cheng, Rahul Goutam, Haiyang Zhang, Karthik Subbian, Suhang Wang, Yizhou Sun, Jiliang Tang, Bing Yin, and Xianfeng Tang. Amazon-M2: A multilingual multi-locale shopping session dataset for recommendation and text generation. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 36, pp. 8006–8026, 2023.

Wittawat Jitkrittum, Harikrishna Narasimhan, Ankit Singh Rawat, Jeevesh Juneja, Congchao Wang, Zifeng Wang, Alec Go, Chen-Yu Lee, Pradeep Shenoy, Rina Panigrahy, Aditya Krishna Menon, and Sanjiv Kumar. Universal model routing for efficient LLM inference. In *International Conference on Learning Representations*, 2026.

Shalmali Joshi, Junzhe Zhang, and Elias Bareinboim. Towards safe policy learning under partial identifiability: A causal approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13004–13012, 2024.

Pooria Joulani, András György, and Csaba Szepesvári. Online learning under delayed feedback. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1453–1461, 2013.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi-Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3910–3919, 2017.

Johannes Kruse, Kasper Lindskow, Saikishore Kalloori, Marco Polignano, Claudio Pomo, Abhishek Srivastava, Anshuk Uppal, Michael Riis Andersen, and Jes Frellsen. RecSys challenge 2024: Balancing accuracy and editorial values in news recommendations. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, pp. 1195–1199, 2024.

Romain Laroche, Paul Trichelair, and Rémi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3652–3661, 2019.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Hao Li, Yiqun Zhang, Zhaoyan Guo, Chenxu Wang, Shengji Tang, Qiaosheng Zhang, Yang Chen, Biqing Qi, Peng Ye, Lei Bai, Zhen Wang, and Shuyue Hu. LLMRouterBench: A massive benchmark and unified framework for LLM routing. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 37733–37754, 2026.

Jian Li, Jieming Zhu, Qiwei Bi, Guohao Cai, Lifeng Shang, Zhenhua Dong, Xin Jiang, and Qun Liu. MINER: Multi-interest matching network for news recommendation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 343–352, 2022.

Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. Neural attentive session-based recommendation. In *Proceedings of the 2017 ACM International Conference on Information and Knowledge Management*, pp. 1419–1428, 2017.

Shaoang Li and Jian Li. Near-optimal online deployment and routing for streaming LLMs. In *International Conference on Learning Representations*, 2026.

Tong Li, Zhaoyang Liu, Yanyan Shen, Xue Wang, Haokun Chen, and Sen Huang. MASTER: Market-guided stock transformer for stock price forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 162–170, 2024.

Zhaoqi Li and Emma Brunskill. A statistical test for the benefits of personalizing interventions. *Science*, 393(6807):eaeb9506, 2026.

Juntao Liu, Lixing Yu, Kun Yue, and Zhiwen Tang. LENS: Latent precision inference in multi-LLM routing. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 3933–3956, 2026.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer, New York, 2003.

Alberto Maria Metelli, Matteo Pirotta, Daniele Calandriello, and Marcello Restelli. Safe policy iteration: A monotonically improving approximate policy iteration approach. *Journal of Machine Learning Research*, 22(97):1–83, 2021.

Mihran Miroyan, Tsung-Han Wu, Logan King, Tianle Li, Jiayi Pan, Xinyan Hu, Wei-Lin Chiang, Anastasios N. Angelopoulos, Trevor Darrell, Narges Norouzi, and Joseph E. Gonzalez. Search arena: Analyzing search-augmented LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087, 2020.

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*, 2025.

Roshini Pulishetty, Mani Kishan Ghantasala, Keerthy Kaushik Dasoju, Niti Mangwani, Vishal Garimella, Aditya Mate, Somya Chatterjee, Yue Kang, Ehi Nosakhare, Sadid Hasan, and Soundar Srinivasan. One head, many models: Cross-attention routing for cost-aware LLM selection. *arXiv preprint arXiv:2509.09782*, 2025.

Ritik Raj, Souvik Kundu, Sarbartha Banerjee, Dheemanth Joshi, Ishita Vohra, and Tushar Krishna. TRACE-ROUTER: Task-consistent and adaptive online routing for agentic AI. *arXiv preprint arXiv:2607.22465*, 2026.

Adam Rida. TRACER: Trace-based adaptive cost-efficient routing for LLM classification. *arXiv preprint arXiv:2604.14531*, 2026.

Lei Shi, Jingshen Wang, and Tianhao Wu. Statistical inference on multi-armed bandits with delayed feedback. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31328–31352, 2023.

Ibne Farabi Shihab, Abu Sa-Adat Mohamed Moon-Im Al Ahsan, and Md Najmus Swaqeeb. Opportunity is not realizability: Selection-valid diagnostics for multi-LLM routing. *arXiv preprint arXiv:2608.08265*, 2026.

Reza Shirkavand, Shangqian Gao, Peiran Yu, and Heng Huang. Cost-aware contrastive routing for LLMs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 172830–172862, 2025.

Jacopo Tagliabue, Ciro Greco, Jean-Francis Roy, Bingqing Yu, Patrick John Chia, Federico Bianchi, and Giovanni Cassani. SIGIR 2021 e-commerce workshop data challenge. *arXiv preprint arXiv:2104.09423*, 2021.

Kei Takemura. Contextual conservative interleaving bandits. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33468–33489, 2023.

Elie Tamer. Partial identification in econometrics. *Annual Review of Economics*, 2:167–195, 2010.

Wei Wang, Tiankai Yang, Hongjie Chen, Yue Zhao, Franck Dernoncourt, Ryan A. Rossi, and Hoda Eldardiry. Learning to route LLMs from bandit feedback: One policy, many trade-offs. *arXiv preprint arXiv:2510.07429*, 2025.

Zhiyuan Wang, Aniri, Tianlong Chen, Yue Zhang, Heng Tao Shen, Xiaoshuang Shi, and Kaidi Xu. LEC: Linear expectation constraints for selection-conditioned risk control in selective prediction and routing systems. In *International Conference on Machine Learning*, 2026.

Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. Neural news recommendation with multi-head self-attention. In *Proceedings of EMNLP-IJCNLP*, pp. 6389–6394, 2019a.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. MIND: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3597–3606, 2020.

Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. Session-based recommendation with graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 346–353, 2019b.

Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1254–1262, 2016.

Xiao Yang, Weiqing Liu, Dong Zhou, Jiang Bian, and Tie-Yan Liu. Qlib: An AI-oriented quantitative investment platform. *arXiv preprint arXiv:2009.11189*, 2020.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, 1994.

Yichi Zhang, Fangzheng Xie, Shu Yang, and Chong Wu. Meta-router: Bridging gold-standard and preference-based evaluations in LLM routing. In *International Conference on Learning Representations*, 2026.

Yiqun Zhang, Hao Li, Jianhao Chen, Hangfan Zhang, Peng Ye, Lei Bai, and Shuyue Hu. Beyond GPT-5: Making LLMs cheaper and better via performance–efficiency optimized routing. In *International Conference on Distributed Artificial Intelligence*, pp. 122–129, 2025.

Hang Zheng, Hongshen Xu, Yongkai Lin, Shuai Fan, Lu Chen, and Kai Yu. DiSRouter: Distributed self-routing for LLM selections. In *International Conference on Learning Representations*, 2026.

Richard Zhuang, Tianhao Wu, Zhaojin Wen, Andrew Li, Jiantao Jiao, and Kannan Ramchandran. EmbedLLM: Learning compact representations of large language models. In *International Conference on Learning Representations*, 2025.

Lukas Zierahn, Dirk van der Hoeven, Tal Lancewicki, Aviv Rosenberg, and Nicolò Cesa-Bianchi. A unified analysis of nonstochastic delayed feedback for combinatorial semi-bandits, linear bandits, and MDPs. *Journal of Machine Learning Research*, 26(104):1–60, 2025.

A NOTATION AND DECISION SETTING

Table 5 collects the objects used by the decision rule. Every decision-time object is fixed or measurable before the corresponding outcome is revealed. The finite-cell theorem also distinguishes the information available to the router from the smaller field used to design the statistical test.

Table 5: Notation used in the routing setting and confidence bounds.

symbols	meaning
t, $m(i)$, $\mathcal{M}_t$	decision origin, maturity time, and $\mathcal{M}_t = \{i < t : m(i) \le t\}$, the outcomes available before origin t
$\mathcal{F}_t^-$, $\mathcal{H}_t^-$	full information immediately before decision t, and its gain-free design subfield
S, S_t	fixed state map and its value at origin t
x_t, τ_ℓ, $\mathcal{T}_\ell$, $\ell(t)$	current state-cell index, the ℓth evaluation time, its governed interval, and the look index
c_t, b_t, s_t	candidate, conservative anchor, and resource-matched state-independent reference
$u_t(e)$, g_t^r	matured expert score and paired gain $u_t(c_t) - u_t(r_t)$
$\mu_{\mathrm{cal}}^{x,r,\ell}$, $\mathrm{LCB}_t^{x,r}$	cell/look calibration mean and its reference-specific lower confidence bound
$\mathcal{G}$, $\Pi_{\mathcal{G}}$, $\Pi_{\varnothing}$	fixed state information ($\mathcal{G} = \sigma(S)$) and the policy classes with and without that information
$U(\pi)$, $C(\pi)$, B, $V_B(\mathcal{G})$	expected cost-adjusted utility, resource cost, budget, and ideal policy-class state value
$I_{x,\ell}$, γ	conservative dual-reference margin at a registered look and its positive lower bound
$Z_{j,t}^{x,r}$, $n_t^{x,r}$, $\overline{Z}_t^{x,r}$	retained block-mean gain, effective matured block count, and mean across blocks for cell x at origin t
$v_t^{x,r}$, $v_{x,\ell}^r$	drift debit for cell x at an origin, and its cell/evaluation representation in the theorem
K, M, N	numbers of state cells, evaluation times, and retained blocks
N_x, q_x, D_x, G_x	required effective support, visits per effective block, feedback delay, and maximum wait for a registered look after support is reached
d, J, h_i, ϕ_t, V_t^h, $b_t^{h,r}$, $\theta_{h,r}$, β_t^h	feature dimension, family count and label, predictable feature, family Gram matrix, score vector, parameter, and confidence radius
R, B_θ, λ	noise sub-Gaussian scale, parameter-norm bound, and ridge parameter
q, $\bar{\beta}(q)$	retained block gap and absolute regularity coefficient
α, δ_{stat}, δ_{mix}, δ_{drift}	per-test level and error budgets for statistical bounds, dependent block coupling, and drift

B THEORY PROOFS AND CONTROLLED AUDITS

This section separates the paper's central identification result from the supporting delayed-feedback guarantees. Theorem 1 establishes why RankShift requires two references. The remaining results

describe when those two evidence streams can be collected under delay, dependence, drift, and finite support. Controlled audits test the stated conditions. They do not replace the proofs.

B.1 Central identification theorem

Theorem 1 is the basis of the paper. Its short two-state proof remains in Section 4 so that the identification argument is visible with the method it motivates. The proof shows that either single reference admits an environment with the same observed data for that reference while the omitted margin changes sign. The results below therefore study the evidence cost and safety of the required joint rule. The central theorem supplies its motivation.

B.2 Assumptions and delayed causal setting

Before decision t, the router receives exactly the outcomes in $\mathcal{M}_t = \{i < t : m(i) \le t\}$. It forms the predecision field $\mathcal{F}_t^-$ from the prior history, this matured feedback, and the covariates available at the current origin. Utility $u_t(e) \in [0,1]$ becomes observable at time $m(t) \ge t$, which is chosen independently of the outcome. State, candidates, references, confidence queries, and decisions at t are $\mathcal{F}_t^-$-measurable. Feedback outside $\mathcal{M}_t$ is unavailable. For the finite-cell result, let $\mathcal{R}$ denote the design chosen on a development prefix disjoint from the evaluated sequence, let X_t collect origin covariates, and let ξ_t be exogenous randomization. The gain-free design field

$$\mathcal{H}_t^- = \sigma(\mathcal{R}, X_{1:t}, m(1{:}t), \xi_{1:t}) \subseteq \mathcal{F}_t^-$$

contains no gain from the evaluation sequence and no nonconstant statistic of such gains. The state rule, candidate and reference assignments, cell and block assignments, retention rule, and confidence schedule are realized using $\mathcal{H}_t^-$ only. A score with $m(i) = i$ is released only after decision i and is first eligible at the next origin. Tied maturities use a fixed reveal order. For candidate c_t, state-independent reference s_t, and conservative anchor b_t, write $g_t^r = u_t(c_t) - u_t(r_t)$, $r \in \{s, b\}$. Each gain lies in $[-1, 1]$.

B.3 Proof of delayed indistinguishability

Theorem 4. For every causal policy, in a two-action construction with utilities in $[0,1]$ and positive integer delay D, there exist two environments with D consecutive decision origins before any reward matures. Their observations are identical throughout this prefix, yet the policy has expected dynamic regret at least $D/2$ in one of them.

Proof. There are two actions. In environment E_+, action one has utility one and action zero has utility zero. Environment E_- reverses these utilities. No reward from the first D origins is observable before those decisions are made. Let p_t be the common probability of choosing action one at $t \in \{1, \dots, D\}$. Indistinguishability implies that the same p_t applies in both environments. Consequently,

$$R_D(E_+) = \sum_{t=1}^{D} (1 - p_t), \qquad R_D(E_-) = \sum_{t=1}^{D} p_t.$$

Their sum is D, so their maximum is at least $D/2$. This includes randomized policies and arbitrary state representations. □

Corollary 2 (Safety and oracle regret incompatibility)**.** Let two environments be observationally indistinguishable for $T \in \mathbb{N}$ decisions. The policy chooses between a fixed fallback and a candidate with gain $+\gamma$ in E_+ and $-\gamma$ in E_-, with $0 < \gamma \le 1$. If expected cumulative harm relative to the fallback in E_- is at most ε, expected regret to the candidate in E_+ is at least $\gamma T - \varepsilon$, where $\varepsilon \ge 0$.

Proof. Let N_T be the number of activations. Observational indistinguishability makes its distribution identical in both environments. Harm in E_- is γN_T, so $\mathbb{E}[N_T] \le \varepsilon/\gamma$. Candidate-oracle regret in E_+ is $\gamma(T - N_T)$, whose expectation is at least $\gamma T - \varepsilon$. For $\varepsilon = 0$, requiring zero expected harm relative to the fallback forces fallback deferral and linear oracle regret. Thus no policy can obtain both sublinear harm and sublinear oracle regret in every environment without more identifying information. □

B.4 PROOF OF DUAL-REFERENCE SELECTIVE SAFETY

Let $\tau_1 < \cdots < \tau_M$ be the fixed confidence-evaluation times. An evaluation may be fixed in $\mathcal{R}$ or triggered by covariate, maturity, or support information, but $\{\tau_\ell = t\} \in \mathcal{H}_t^-$ and it cannot be triggered by an evaluated gain. For each cell, reference, and evaluation triple (x, r, ℓ), define the pre-gain selection field $\mathcal{P}_{x,r,\ell} \subseteq \mathcal{H}_{\tau_\ell}^-$. It contains the disjoint development prefix, request covariates, fixed candidate and reference assignments, cell and block assignments, maturity times, retention indices, and outcome-independent randomization. It contains neither a retained gain nor a nonconstant statistic computed from retained gains. The retained index set $\mathcal{I}_{x,r,\ell}$ and its count $n \geq 1$ are $\mathcal{P}_{x,r,\ell}$-measurable. The theorem uses uniform weights $1/n$. Fixing the rule in advance is necessary but not sufficient. For example, a rule may assign the next cell from the sign of the last matured gain. This rule is predictable, but it is not gain-free and therefore falls outside the finite-cell theorem.

Given $\mathcal{P}_{x,r,\ell}$, assume a regular conditional coupling kernel for the retained vector Z and an auxiliary vector Z^*. The coordinates of Z^* are conditionally independent, and each one has the matching conditional law. Also assume that $\Pr(Z \neq Z^* \mid \mathcal{P}_{x,r,\ell}) \leq (n-1)\bar{\beta}(q)$, where $\bar{\beta}(q)$ is the mixing bound at separation q. Each coordinate of Z is a retained block mean in $[-1, 1]$. This condition must hold after the method selects the retained blocks. Unconditional mixing of the raw gain stream does not establish it after conditioning on $\mathcal{P}_{x,r,\ell}$. The theorem excludes any cell, block, retention, or evaluation-time rule that depends on evaluated outcomes. Define the $\mathcal{P}_{x,r,\ell}$-measurable calibration target $\mu_{\mathrm{cal}}^{x,r,\ell} = n^{-1} \sum_j \mathbb{E}[Z_j^* \mid \mathcal{P}_{x,r,\ell}]$. There are K cells, M evaluation times, and a deterministic upper bound N on the number of retained blocks in any comparison. Set $\tau_{M+1} = T + 1$ and $\mathcal{T}_\ell = [\tau_\ell, \tau_{\ell+1})$. The error budgets satisfy $\delta_{\mathrm{stat}} > 0$, $\delta_{\mathrm{mix}}, \delta_{\mathrm{drift}} \geq 0$, and $\delta_{\mathrm{stat}} + \delta_{\mathrm{mix}} + \delta_{\mathrm{drift}} = \delta \in (0, 1)$ with $2KM(N-1)\bar{\beta}(q) \leq \delta_{\mathrm{mix}}$. On a simultaneous drift event of probability at least $1 - \delta_{\mathrm{drift}}$, for every $t \in \mathcal{T}_\ell$ with $S_t = x$,

$$\mathbb{E}[g_t^r \mid \mathcal{F}_t^-] \geq \mu_{\mathrm{cal}}^{x,r,\ell} - v_{x,\ell}^r.$$

All drift inequalities are interpreted almost surely on the displayed cell event, and $v_{x,\ell}^r \geq 0$ is a gain-free, registered (or deterministic) interval-uniform drift debit. Set $\alpha = \delta_{\mathrm{stat}}/(2KM)$. This full-history condition is stronger than a pooled mean bound. If the realized reference satisfies $s_t = c_t$, then $g_t^s = 0$ and a positive conditional lower bound cannot hold at that origin. The condition must be verified for the deployed reference and cannot be inferred from pooled audits.

Theorem 5. Let $n_{x,\ell}^r$ be the effective number of retained matured blocks and let $\overline{Z}_{x,\ell}^r$ be their empirical mean. Define

$$\mathrm{rad}(n, \alpha) = \sqrt{\frac{2\log(1/\alpha)}{n}}, \qquad \mathrm{LCB}_{x,\ell}^r = \overline{Z}_{x,\ell}^r - \mathrm{rad}(n_{x,\ell}^r, \alpha) - v_{x,\ell}^r.$$

Suppose the controller authorizes a cell only when both reference lower bounds are strictly positive. It holds that decision fixed over $\mathcal{T}_\ell$. Then, with probability at least $1 - \delta$, every action executed under that authorization has nonnegative current conditional expected gain relative to both references.

The theorem is stated for comparisons with $n_{x,\ell}^r \geq 1$. The implementation sets the LCB to $-\infty$ when $n_{x,\ell}^r = 0$. This is an operational fallback convention and yields no authorization claim.

Proof. Consider one state cell, one evaluation time, and one reference. We condition on the pre-gain information $\mathcal{P}_{x,r,\ell}$ used to form this comparison. Work on the coupling extension and let $E_{\mathrm{coup}} = \{Z = Z^*\}$. Conditional on $\mathcal{P}_{x,r,\ell}$, $\Pr(E_{\mathrm{coup}}^c) \leq (n-1)\bar{\beta}(q)$, while the coordinates of Z^* are independent (not necessarily identically distributed). Hoeffding's inequality therefore gives

$$\Pr\Big(\overline{Z}^* - \mu_{\mathrm{cal}} > \mathrm{rad}(n, \alpha) \mid \mathcal{P}_{x,r,\ell}\Big) \leq \alpha.$$

Since $\overline{Z} = \overline{Z}^*$ on E_{coup}, a union bound gives

$$\Pr\big(\overline{Z} - \mu_{\mathrm{cal}} > \mathrm{rad}(n, \alpha) \mid \mathcal{P}_{x,r,\ell}\big) \leq \alpha + (n-1)\bar{\beta}(q).$$

Unioning this event across cells, the two references, and all evaluation times contributes at most $\delta_{\mathrm{stat}} + \delta_{\mathrm{mix}}$ (Berbee, 1979; Yu, 1994; Hoeffding, 1963). On its complement, the calibration mean is at least $\overline{Z} - \mathrm{rad}$. Subtracting the drift debit then lower-bounds every current mean in the governed interval. Adding the drift failure probability gives the total error budget δ. □

The theorem controls the mean that weights retained block means uniformly. CALIBRI gains also average block means equally. Calls and coverage count rows. Realized block gains are also reported separately to describe tail behavior. A tail-risk extension therefore requires its own statistic and error accounting.

The finite-cell guarantee does not cover a controller that revises its state map, candidate or reference assignment, block retention, or look schedule using matured gains from the same evaluation sequence. Such adaptation requires a selection-aware confidence sequence. When it is represented by predictable features under the stated linear model, the martingale result below provides one valid route. Otherwise a new confidence argument is required. Paired shadow outcomes used to form retained blocks must likewise be collected independently of prior authorization. If prior gains change their availability or retention, the finite-cell guarantee does not apply.

Choosing dependence and drift bounds. To apply the theorem, the block gap, evaluation schedule, and error split are chosen on a calibration prefix that is separate from the evaluated sequence. For each allowed gap q, the analysis needs a simultaneous upper bound on the conditional absolute-regularity coefficient after conditioning on the gain-free selection field. The chosen gap must satisfy $2KM(N-1)\bar{\beta}(q) \le \delta_{\mathrm{mix}}$. Sample autocorrelation, block-length stability, and permutation checks are useful sensitivity checks. They do not replace this bound. Likewise, $v^r_{x,\ell}$ must jointly bound worsening from calibration to the current distribution. This bound can come from a justified shift model or a separate change-point stress envelope with failure budget δ_{drift}. A rolling plug-in estimate alone does not keep the guarantee valid. When either bound is unavailable, RankShift reports a paired empirical interval for the fixed evaluation set but does not apply the finite-cell safety theorem to that domain.

B.5 Controlled audits of dependence and drift

B.5.1 Known-generator stress test

The controlled stress suite uses $K = M = 4$, two references, $\delta_{\mathrm{stat}} = .05$, $\delta_{\mathrm{mix}} = .01$, $\delta_{\mathrm{drift}} = .01$, $N_{\max} = 2048$, seed 2027, and exactly 10,000 replications per cell. It generates $G_j = \mathrm{clip}(\mu + 0.1S_j, -1, 1)$, where S_j is a stationary symmetric two-state Markov chain with $\rho = .2$. Retained blocks are separated by $q = 10$, giving $\bar{\beta}(q) = 0.5\rho^q = 5.12 \times 10^{-8}$ and coupling bound $2KM(N_{\max}-1)\bar{\beta}(q) = 0.0033538 \le .01$. Calibration margins are $\mu_{\mathrm{cal}} \in \{-.25, +.25\}$. Deployment gain is defined as the constant $\mu_{\mathrm{cal}} - .05$, so its conditional mean is independent of the observed calibration chain. The interval-uniform debit is therefore .05. The positive post-drift margin is .20 and the sufficient-support expression is $n_{\min} = 8\log(2MK/\delta_{\mathrm{stat}})/.20^2 = 1292.2936$. Supports are $\{128, 256, 512, 1024, 2048\}$ and delays are $\{0, 32, 64, 128, 256\}$. Retained indices are gain-free. Safe-positive activation at delay zero is $0, 0, 1, 1, 1$ over those supports, while unsafe-negative false authorization is zero throughout. The gain-free negative control has activation/false-authorization/fallback $0/0/1$. The intentionally invalid top-quartile gain-dependent selector has $1/1/0$. These are controlled stress results, not claims about natural CALIBRI coverage.

B.5.2 Real-data-backed pool-mean audit

This audit tests concentration and bounded mean drift on fixed CALIBRI reservoirs. Each quality-minus-cost gain is clipped to $[-1, 1]$ before resampling. Independent draws with replacement give $\bar{\beta}(q) = 0$ for $q \ge 1$, conditional on the fixed reservoir. Replacing an η fraction of deployment mass by gain -1 gives mean $(1-\eta)\mu_{\mathrm{pool}} - \eta$ and debit $v = 2\eta$, which bounds the mean deterioration. The target is this clipped pool mean. It does not establish per-request safety conditional on router covariates or coverage for the naturally ordered CALIBRI stream.

The always-candidate policy is compared separately with the production anchor and a fixed state-free hash allocation. The hash has a different candidate-call rate, so this audit does not measure resource-matched allocation value. Clipped reservoir means are $+.0507/+.0445$ (anchor/hash) for LiveCodeBench, $-.1429/-.1149$ for McEval, and $-.5146/-.4181$ for MultiPL-E. With $\delta_{\mathrm{stat}} = .04$ and $\delta_{\mathrm{drift}} = .01$, the reference-specific gates have zero false authorizations across all 18 family/reference/contamination settings. Each uses seven support values and 20,000 independent replications per support. A false authorization activates a gate whose constructed deployment pool mean is strictly negative. The references are simulated separately. At $\eta = 0$, the two positive

LiveCodeBench margins reach 80% activation by support $n = 8192$. Replacing 1% of deployment mass by the worst bounded gain gives $v = 0.02$. The anchor reaches 80% only at $n = 16384$ (maximum activation 97.59%), while the hash reference peaks at 46.10% and never reaches 80%. At $\eta = .02$, both LiveCodeBench references have zero activation on the tested grid. For McEval and MultiPL-E, every tested combination of family, reference, and contamination has a nonpositive pool mean and zero activation. These are per-setting stress-test rates. The support and contamination grid is not counted as additional registered looks. The anonymous supplementary package includes the complete row-level output and audit receipt.

Table 7: Clipped pool-mean audit on real CALIBRI reservoirs. "80% support" is the first tested support with activation probability at least 0.80. "–" means no tested support through $16{,}384$ reached that level.

Family	Reference	η	Base mean	80% support	Max false auth.
LiveCodeBench	anchor	0	$+.0507$	8192	0
LiveCodeBench	hash	0	$+.0445$	8192	0
LiveCodeBench	anchor	.01	$+.0507$	16384	0
LiveCodeBench	hash	.01	$+.0445$	–	0
LiveCodeBench	anchor	.02	$+.0507$	–	0
McEval	anchor	0	$-.1429$	–	0
MultiPL-E	hash	0	$-.4181$	–	0

B.6 Delayed linear confidence controller

The development controllers for Search Arena 24k (Miroyan et al., 2026) and the 10k WebDev Arena slice of the Arena platform (Chiang et al., 2024) are linear. We analyze them separately from the finite-cell mixing result above. This is a model-conditional guarantee and does not assert that the natural traces satisfy the stated MGF envelope a priori. Let the predictable family label be $h_t \in [J] = \{1, \dots, J\}$ and let $\phi_t \in \mathbb{R}^d$ satisfy $\|\phi_t\|_2 \le 1$. For each family h and reference $r \in \{s, b\}$, suppose the gain observed at maturity obeys

$$g_t^{h,r} = \phi_t^\top \theta_{h,r} + \eta_t^{h,r}, \qquad \|\theta_{h,r}\|_2 \le B_\theta,$$

when $h_t = h$, with a fixed (time-invariant) parameter $\theta_{h,r}$. Define $a(i) = \max\{m(i), i+1\}$ as the first decision at which outcome i is available. Let ρ_k index distinct availability times and set $\mathcal{B}_k^h = \{i : a(i) = \rho_k,\ h_i = h\}$. Let $\mathcal{Q}_k$ be the nested information field just before the full batch at ρ_k enters the predecision history. Assume these actual batches and their features are $\mathcal{Q}_k$-measurable and, for every $z \in \mathbb{R}^d$ and (h, r),

$$\mathbb{E}\left[\exp\left\{ z^\top \sum_{i \in \mathcal{B}_k^h} \phi_i \eta_i^{h,r} \right\} \middle| \mathcal{Q}_k \right] \le \exp\left\{ \frac{R^2}{2} z^\top \left(\sum_{i \in \mathcal{B}_k^h} \phi_i \phi_i^\top \right) z \right\}.$$

Also assume $\mathbb{E}[g_t^{h,r} \mid \mathcal{F}_t^-] = \phi_t^\top \theta_{h,r}$. For ridge parameter $\lambda > 0$, define the family-specific matured set and updates

$$\mathcal{M}_t^h = \{i : a(i) \le t,\ h_i = h\}, \qquad V_t^h = \lambda I_d + \sum_{i \in \mathcal{M}_t^h} \phi_i \phi_i^\top,$$

$$b_t^{h,r} = \sum_{i \in \mathcal{M}_t^h} \phi_i g_i^{h,r}, \qquad \widehat{\theta}_t^{h,r} = (V_t^h)^{-1} b_t^{h,r},$$

This matured set equals $\{i < t : m(i) \le t,\ h_i = h\}$ and is a union of complete availability batches. Here I_d is the identity matrix. The object $b_t^{h,r}$ here is a family-specific score vector, not the anchor action b_t used in the finite-cell notation.

Theorem 6 (Delayed linear dual-reference safety)**.** For $\delta \in (0, 1)$, set $\alpha = \delta/(2J)$ and

$$\beta_t^h = R\sqrt{\log \frac{\det V_t^h}{\det(\lambda I_d)} + 2\log(1/\alpha)} + \sqrt{\lambda} B_\theta, \qquad \mathrm{LCB}_t^{h,r} = \phi_t^\top \widehat{\theta}_t^{h,r} - \beta_t^h \|\phi_t\|_{(V_t^h)^{-1}}.$$

Under the assumptions above, with probability at least $1-\delta$, simultaneously over all evaluated decision times, families, and both references,

$$\left|\phi_t^\top(\widehat{\theta}_t^{h,r}-\theta_{h,r})\right| \le \beta_t^h \|\phi_t\|_{(V_t^h)^{-1}}.$$

Here $\|z\|_{(V_t^h)^{-1}} = \sqrt{z^\top (V_t^h)^{-1} z}$. Consequently, at any predictable evaluated origin, a controller that activates family h only when $\min_{r\in\{s,b\}} \mathrm{LCB}_t^{h,r} > 0$ has strictly positive conditional expected gain against both references on this event.

Proof. Treat each complete availability batch as one vector increment in the exponential-supermartingale proof of the self-normalized inequality. The displayed conditional MGF is assumed for these increments. Every evaluated $\mathcal{M}_t^h$ contains whole batches, so the same bound applies to its V_t^h and $b_t^{h,r}$. Gaussian integration gives the determinant term, and ridge bias contributes $\sqrt{\lambda}B_\theta$. A union bound over the $2J$ family/reference processes yields the stated simultaneous bound (Abbasi-Yadkori et al., 2011). □

For the fixed Search Arena/WebDev Arena development controller, $d=5$, $\lambda=1$, $R=1$, $B_\theta=2$, $J=3$, and $\alpha = .05/(3\cdot 2)$. These values define the linear controller used in each development dataset and seed family. RankShift-Code is a separate fixed family rule evaluated by the paired block bootstrap in Appendix E. It does not invoke this linear theorem.

B.7 Activation boundary from support and complexity

Proof of Proposition 1. The statement in the main text is proved as follows.

Proof. Set $\alpha = \delta_{\mathrm{stat}}/(2KM)$. For each of the $2KM$ tests, Hoeffding's lower-tail bound gives failure probability at most α for the event that the empirical mean falls below its calibration expectation by more than one radius. A union bound therefore gives this statement simultaneously for all cells, references, and looks with probability at least $1-\delta_{\mathrm{stat}}$. On that event, $\overline{Z} \ge \mu_{\mathrm{cal}} - \mathrm{rad}(n,\alpha)$, while the LCB subtracts a second radius and the drift debit. Thus a post-debit margin γ yields a strictly positive LCB whenever $2\,\mathrm{rad}(n,\alpha) < \gamma$. Rearranging gives $n > 8\log(2KM/\delta_{\mathrm{stat}})/\gamma^2$. Dependence is represented by retained effective blocks. Raw row count cannot replace n_{eff}. Delay withholds matured observations and therefore postpones possible activation. The coupling failure contributes the additional δ_{mix} term in the proposition. The controlled stress suite uses the registered nonzero coupling bound reported above. □

Converse proof. For the converse in Proposition 1, let $Y \in \{0,1\}$ have mean $1/2-\gamma$ in one experiment and $1/2+\gamma$ in the other, so the centered margin $\mathbb{E}[Y-1/2]$ is $-\gamma$ or $+\gamma$. The one-sample KL divergence is at most $16\gamma^2$ for $\gamma \le 1/4$. The Bretagnolle–Huber inequality bounds the sum of the two errors below by $\frac{1}{2}\exp(-16n\gamma^2)$. Requiring both errors to be at most $\delta \in (0, 1/8]$ gives $n \ge \log(1/(4\delta))/(16\gamma^2) \ge \log(1/\delta)/(48\gamma^2)$. This matches the margin and confidence dependence of the sufficient result up to constants (Lattimore & Szepesvári, 2020).

This two-point construction establishes the order-level necessity of the retained activation boundary without introducing a separate theorem.

Proposition 2 (Time to the first eligible authorization)**.** Assume the activation-boundary proposition's conditions for a cumulative history with zero drift debit and a fixed state x. Suppose every calibration margin $\mu_{\mathrm{cal}}^{x,r,\ell}$ is at least $\gamma_x > 0$. Define

$$N_x = \left\lfloor \frac{8\log(2MK/\delta_{\mathrm{stat}})}{\gamma_x^2} \right\rfloor + 1.$$

Assume every q_x visits to x produce at least one retained effective block before accounting for delay. Let D_x upper-bound the feedback delay in visits. The registered schedule must include a look within G_x further visits after N_x blocks mature and within the horizon. On the simultaneous lower-tail concentration event, the first eligible look occurs within

$$q_x N_x + D_x + G_x$$

visits. For finitely many persistently beneficial cells, fallback-deferral regret accrued before the first eligible look is at most $\sum_x (q_x N_x + D_x + G_x)$ because utility gaps are bounded by one.

Proof. The activation-boundary proposition makes both reference LCBs strictly positive once N_x effective matured blocks are available. The effective-support clock reaches this threshold after at most $q_x N_x + D_x$ visits. The next fixed look is no more than G_x visits later, giving the displayed bound. Before that look each fallback loses at most one relative to the candidate. If visits cease, the margin changes, or old observations are discounted, the premise fails and no finite raw-visit bound follows. □

B.8 Candidate approximation and selective-regret accounting

On the normalized theoretical scale, let $V_t(a) = \mathbb{E}[u_t(a) \mid \mathcal{F}_t^-] \in [0,1]$. Let $\mathcal{A}_t$ be the full action set and let the nonempty fixed candidate set $\mathcal{C}_t \subseteq \mathcal{A}_t$, the estimator $\widehat{V}_t$, and a deterministic tie-break be $\mathcal{F}_t^-$-measurable. On an event $\mathcal{E}_{\mathrm{app}}$, simultaneously for every $t \le T$, assume that a_t^* maximizes V_t over $\mathcal{A}_t$, that some $c_t^\circ \in \mathcal{C}_t$ satisfies $V_t(a_t^*) - V_t(c_t^\circ) \le \kappa_t$, and that $\sup_{c \in \mathcal{C}_t} |\widehat{V}_t(c) - V_t(c)| \le \epsilon_t$. Let $\widehat{a}_t \in \arg\max_{c \in \mathcal{C}_t} \widehat{V}_t(c)$. Then

$$V_t(a_t^*) - V_t(\widehat{a}_t) \le \kappa_t + 2\epsilon_t.$$

Add and subtract the two estimated candidate values, then use optimality within $\mathcal{C}_t$. The resulting bound combines candidate coverage with estimation accuracy.

Let $A_t \in \{0,1\}$ denote the authorization carried by the latest registered look, let the executed action be $\widehat{a}_t$ when $A_t = 1$ and b_t otherwise, and define $r_t^b = V_t(a_t^*) - V_t(b_t)$. The conditional pseudo-regret decomposition is

$$\mathrm{Reg}_T^{\mathrm{sel}} = \sum_t (1 - A_t) r_t^b + \sum_t A_t [V_t(a_t^*) - V_t(\widehat{a}_t)].$$

On $\mathcal{E}_{\mathrm{app}}$, this is at most $\sum_t (1 - A_t) r_t^b + \sum_t A_t \min\{1, \kappa_t + 2\epsilon_t\}$. The first term measures fallback regret and the second measures candidate approximation.

For the deployed proposal $c_t := \widehat{a}_t$ and realized gain $g_t = u_t(c_t) - u_t(b_t)$, let $R_T^{\mathrm{base,real}}$ and $R_T^{\mathrm{sel,real}}$ denote anchor and selective regret over horizon T. The pathwise identity $R_T^{\mathrm{sel,real}} = R_T^{\mathrm{base,real}} - \sum_t A_t g_t$ motivates candidate headroom $H_T^{\mathrm{cand}} = \sum_t [g_t]_+$. When $H_T^{\mathrm{cand}} > 0$, signed capture is $C_T = \sum_t A_t g_t / H_T^{\mathrm{cand}}$. The development-unit frontier uses a distinct dual-reference limiting-value capture. For fixed unit h, let $\bar{g}_h^b$ and $\bar{g}_h^s$ be its paired mean gains, define $\ell_h = \min\{\bar{g}_h^b, \bar{g}_h^s\}$, and for activation indicator A_h set

$$C_{\mathrm{DR}}(A) = \frac{\sum_h A_h [\ell_h]_+}{\sum_h [\ell_h]_+},$$

whenever the denominator is positive. This metric gives 0.9731 and 0.6857. It differs from the anchor-based signed capture C_T above. An oracle headroom comparison requires a fixed candidate set and verifies $c_t \in \mathcal{A}_t$ and identical origin and utility universes for candidate, oracle, and fallback.

C Algorithm and implementation details

C.1 Matured-feedback update

At a registered checkpoint τ_ℓ, the implementation first adds every earlier outcome with $m(i) \le \tau_\ell$ to the two gain histories. It then computes the two lower bounds for each fixed state cell using only these matured gains. A cell is authorized only when both bounds are strictly positive. The resulting authorization map remains fixed over $[\tau_\ell, \tau_{\ell+1})$ and is replaced only at the next scheduled look. This carry-forward rule uses the interval-uniform drift condition of Theorem 3. Otherwise the implementation returns the anchor, or system-abstains only if no anchor is available. A new outcome enters the histories only when it matures, so later feedback cannot change an earlier decision. Ties in maturity time follow a fixed outcome-independent order.

C.2 From analytic bounds to paired evaluation

The controlled generator matches the finite-state theorem. Gains are bounded, outcomes mature chronologically, selection rules remain fixed, and dependence and drift are known. It lets us compare the observed activation boundary with the boundary predicted by the theorem.

The real-data studies use the same decision structure. We set each candidate rule on development data and add outcomes only after they mature. We compare experts on the same request. Paired empirical intervals measure the two margins in their native units. Finance, news, commerce, and CALIBRI contain positive routing opportunities. Search Arena, WebDev Arena, and the arena-human-preference-140k release (Chiang et al., 2024) retain the anchor when the two margins do not clear zero. These settings show how the support requirement affects the rule beyond the controlled generator.

The Binance and Arena-derived inputs are frozen external snapshots. The protocol records their snapshot roles and split years. the anonymous package does not redistribute the raw data or embed direct download URLs.

D Why two references are necessary

We compare two nested rules on the same 35 development units. The mean-only rule requires positive paired bootstrap bounds against both references and selects $19/35$ units. Adding a non-negative worst-decile requirement selects $7/35$. Both rules use the dual-reference comparison in Algorithm 1. The second adds an empirical check for block-level stability. The analytic controller gives the finite-cell guarantee when dependence and drift bounds are available.

D.1 Why neither reference can be removed

Theorem 1 gives a counterexample within its restricted observation model. The finite-sample acceptance set is $\{\widehat{\Delta}_b - r_b > 0, \widehat{\Delta}_s - r_s > 0\}$, where the two estimates and one-sided radii may share paired observations. With the same anchor-only record and fixed gain-free design, the omitted state-free bound can have either sign. A threshold κ based only on that information therefore cannot make $\widehat{\Delta}_b - \kappa > 0$ reproduce the joint rule uniformly over this class. The rule requires information about both margins. One observed stream can supply that information when its recorded actions cover both references.

Extra information can reduce the comparison. For zero drift, if $d = U(s) - U(b)$ is known, the population rule equals $\Delta_b > \max\{0, d\}$, where $\Delta_b = U(c) - U(b)$. If $U(s) \geq U(b)$ on the same unit, a positive state-free margin implies anchor improvement. This includes a population-optimal state-free reference whose feasible class contains b. These population relations do not order separately estimated lower bounds.

The symbol s denotes a protocol-specific state-free comparator, so its construction determines the claim. Composition matching fixes marginal expert use. A population-optimal frontier represents the best feasible state-free policy. An arbitrary blind rule supports only a comparison with itself. If $s = b$, Gate 2 equals Gate 1 and adds no attribution contrast. Proposition 3 states this hierarchy. None of these comparisons estimates the oracle quantity $V_B(\mathcal{G})$.

Proposition 3 (Comparator construction determines claim strength)**.** Fix a registered evaluation collection A and a proposal c. Assume the resource cost depends only on expert counts. Let s_{cmp} assign c's fixed expert multiset uniformly at random over A using randomization independent of states and outcomes. Let $U(s_{\mathrm{cmp}})$ average over it. Then c and s_{cmp} have identical marginal expert use. A positive $U(c) - U(s_{\mathrm{cmp}})$ measures the benefit of the registered assignments relative to random allocation at that composition. Define $\Pi_{\varnothing,B} = \{s \in \Pi_{\varnothing} : C(s) \leq B\}$. If $s^* \in \arg\max_{s \in \Pi_{\varnothing,B}} U(s)$, then $U(c) > U(s^*)$ implies $U(c) > U(s)$ for all $s \in \Pi_{\varnothing,B}$. Against an arbitrary fixed blind s, $U(c) > U(s)$ supports only comparator-relative advantage.

Proof. Permutation preserves expert counts and count-based cost while removing the registered link between state and assignment. The frontier claim follows from the definition of s^*. An arbitrary

comparator has no such maximizing property. These are operational comparisons within A. They do not establish causal identification.

D.2 Anchor-only failures and the need for Gate 2

We evaluate two contract-preserving one-reference adaptations on the same paired blocks, validation looks, and call quota. AnchorSafe retains a cell using only $\mathrm{LCB}_b > 0$, whereas StateAttribution uses only $\mathrm{LCB}_s > 0$. They are compatible anchor-safety and selection-conditioned reference controls, not reimplementations of LEC or TRACER under different objectives. Their intersection is the dual mean rule.

Table 8 breaks down the same development lower bounds. Gate 1 uses only the anchor lower bound, Gate 2 uses only the state-independent lower bound, and the mean-only dual gate requires both. The final column adds the worst-decile screen. "G1 nonpositive" counts Gate-1 authorizations whose limiting mean across the two references is nonpositive. These one-gate rows are contract-preserving adaptations of the adjacent anchor-safety and reference-parity controls, not reimplementations of LEC or TRACER under their different objectives.

Table 8: Per-domain gate ablation on 35 development units. Each dataset contributes five seed units.

Dataset	G1	G1 non-positive	G2	Dual mean	Tail screen
MIND	$5/5$	$0/5$	$5/5$	$5/5$	$2/5$
EB-NeRD	$5/5$	$0/5$	$3/5$	$3/5$	$0/5$
Coveo	$5/5$	$0/5$	$5/5$	$5/5$	$5/5$
CSI300	$5/5$	$0/5$	$4/5$	$4/5$	$0/5$
CSI500	$5/5$	$3/5$	$2/5$	$2/5$	$0/5$
Binance	$5/5$	$4/5$	$0/5$	$0/5$	$0/5$
LLMRouterBench (two-model)	$2/5$	$2/5$	$0/5$	$0/5$	$0/5$
All units	$32/35$	$9/35$	$19/35$	$19/35$	$7/35$

The nine Gate-1-only authorizations that lack a positive limiting margin against both references are exactly three CSI500, four Binance, and two LLMRouterBench units from the two-model development protocol. This 35-unit protocol uses a distinct state-free comparator with $s_t \neq b_t$. The public 13-model transfer is a separate protocol whose optimized mixture has $s_t \equiv b_t$. News and commerce contribute none. In units where the two references differ, improvement over the deployed anchor does not survive the matched state-independent comparison. This is the failure mode targeted by Gate 2. A strong candidate on average does not show that the state map identifies where to switch. Across these units, every Gate-2 pass also clears Gate 1, so Gate 2 and their joint rule coincide. This ablation shows the filtering role of the state-independent comparison. Theorem 1 establishes the anchor comparison's necessity within its observation model.

D.3 State-free-only failures and the need for Gate 1

We separately test the failure mode missing from the 35-unit ablation. A cell can clear the state-free comparison while failing the anchor comparison. The frozen supplementary protocol adapts twelve proposal methods to the same aligned 13-model trace and uses shared prompt-only features. It fixes five hash seeds and seven tasks. It also fixes a 50/20/15/15 fit/tune/authorization/test split, a 30% candidate-call budget, and at least 20 authorization rows per mature task cell. This yields $12 \times 5 \times 7 = 420$ cells defined by router, seed, and task.

Among the 420 cells, 46 have $\mathrm{LCB}_s > 0$ and $\mathrm{LCB}_b \leq 0$ with nonzero candidate calls. The pattern appears in all 12 router adaptations, all five seeds, five tasks, and 18 distinct clusters defined by router and task. The dual rule rejects all 46. On the untouched test split, eight of the corresponding ungated proposals have negative realized gain against the anchor. Table 9 shows that most reverse cells occur on SimpleQA, while the remaining cases span scientific, knowledge, code, and general reasoning tasks.

Table 9: Reverse one-reference cells in the frozen 420-cell check. Counts require positive state-free LCB, nonpositive anchor LCB, maturity, and a nonzero proposed call. The dual rule rejects every listed cell.

Task	Reverse cells	Distinct router adaptations
GPQA	1	1
HLE	1	1
LiveCodeBench	3	3
MMLU-Pro	2	1
SimpleQA	39	12
All	46	12

This extension checks the mechanism. It does not estimate prevalence. The cellwise lower bounds lack a familywise adjustment across 420 searches. The 46 cells share prompts and fitted proposals. The frozen protocol artifact also lacks an external preregistration timestamp. Thus, the result shows that the reverse sign pattern occurs across methods and seeds. It does not establish 46 independent discoveries or a population frequency of $46/420$. The $9/35$ development result is sign-pattern evidence that Gate 2 has a separate role. The eight negative outcomes are separate held-out evidence that removing Gate 1 can harm the anchor.

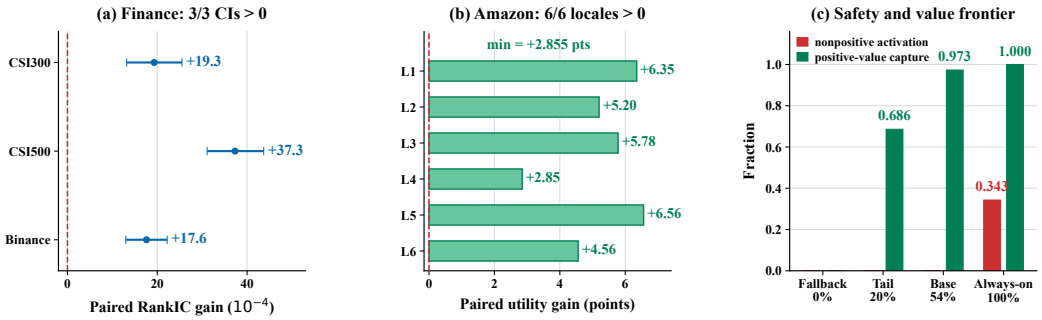

Figure 3: Opportunity and the safety and value frontier. (a,b) Fixed-proposal intervals are positive in all three markets and six Amazon locales. (c) On 35 development units, the nested mean-only and full tail-screened gates authorize 19/35 and 7/35 units and capture 0.973 and 0.686 of positive dual-reference limiting value, respectively. Neither gate authorizes a unit with nonpositive limiting mean. Always-on activates all 35 units, including 12 (34.3%) whose limiting mean is nonpositive.

The domain breakdown shows the role of each check. The state-value gate removes finance and LLM units whose anchor gain also appears under blind allocation. The tail screen then keeps gains that remain nonnegative across the worst blocks. Per-unit results appear in the supplementary material.

E Main Experimental Validation

This section follows the main empirical claims in the order used to check them. It starts with negative open-text controls and the held-out CALIBRI setup. It then covers source-order delayed replay, the full frozen-proposal record, matched baselines, and public multi-model transfer. Transfer-domain protocols are collected in Appendix F.

Section F gives the state, candidate, references, maturity clock, utility, and split for each real-data experiment. All learned statistics use only the chronological training prefix. Stochastic methods use seeds 0–4. The controlled theorem experiment rescales gains to $[-1, 1]$. Real-data studies retain their native RankIC, NDCG, MRR, and quality-after-cost units. Finance reports predictive RankIC separately from long–short returns after transaction costs. Only the CSI500 post-friction analysis is used as evidence of economic value. CSI300 and Binance RankIC remain predictive measures. These real-data intervals are empirical paired bootstrap or normal intervals with zero drift debit. Without a supplied dependence/drift envelope they do not claim the analytic theorem's coverage. For Amazon-M2, we weight Task 1 and Task 2 by their observed rows within each seed. We then

average the five seed-level gains. The reported value therefore differs from the unweighted mean of the six locale summaries.

E.1 Open-text controls and state resolution

Search Arena and WebDev Arena provide full-information delayed feedback. Each proposal outcome becomes available only after its maturity time, including proposals that were not selected. Predictors are fitted on the first 60 percent of the stream and evaluated on the final 20 percent. Across datasets, budgets, and seeds, the dual-reference rule retains the anchor because the lower confidence bounds do not jointly clear zero. Replaying the sequence after changing outcomes that had not yet matured leaves all earlier decisions unchanged, confirming the intended information flow.

Two follow-up analyses separate too little support from a small margin. First, an outcome-free four-cluster state map still splits the useful swaps into groups that are too small for the confidence test. Second, we pool the out-of-fold proposals and reduce four states to two and then one. This gives 260 Search Arena decisions and 825 WebDev Arena decisions enough support to reach the test. However, neither dataset produces two positive lower bounds. Coarser states can reduce the startup cost. They cannot create state value when the dual-reference margin is nonpositive.

The arena-human-preference-140k release shows the same mechanism. The fixed contextual route has mean gains of -0.00229, 0, and -0.00457 at candidate budgets $.1$, $.25$, and $.5$, and every paired interval contains zero. RankShift therefore keeps the anchor in this open-text setting.

E.2 Conditions that produce identifiable LLM routing

The development experiments isolate three prerequisites for useful selective routing. Grouping them by mechanism shows how interface quality, candidate value, and selective support jointly determine whether a proposal becomes actionable.

Requirement	Development observation	Consequence for routing
Interface feasibility	Several long-form tasks fail to produce a bounded, machine-readable answer reliably within the available generation budget.	A candidate cannot be evaluated consistently enough to support authorization.
Candidate value	After output normalization removes formatting failures, candidate regressions outnumber its unique corrections on the tested tasks.	A technically valid candidate still lacks positive value against both references.
Selectivity	Under shorter probes and adaptive stopping, at most one of twenty examples qualifies for a candidate call in the affected tasks.	The matured support remains below the level needed for a positive lower confidence bound.

These observations motivate the CALIBRI evaluation. Executable labels provide a stable utility, and task family supplies a deployment-time state with enough support to measure both margins. The CALIBRI test then evaluates the resulting rule on a separate split.

E.3 CALIBRI setup and held-out evaluation

The model pair and state rule were selected on the CALIBRI training and validation splits before the test outcomes were examined. RankShift-Code is an interpretable heuristic example. Its Python implementation reads the public benchmark and language labels directly. MultiPL-E is a fixed fallback. This places RankShift-Code outside the embedding-only learned-router comparison. The paper's main composition claim freezes those learned proposal scores and changes only their authorization layer. The anchor is Qwen3-Coder-30B-A3B-Instruct and the candidate is GPT-OSS-20B. The router sends all LiveCodeBench origins and the McEval languages $\{$Perl, Python, Rust, Scala$\}$ to the candidate and falls back to the anchor for all other states, including MultiPL-E. Candidate states were required to have positive candidate-minus-anchor development means on both CALIBRI train and validation with train support at least 20 and validation support at least 10.

The primary utility is correctness minus 10^{-5} times a deterministic token-equivalent compute proxy. This λ is a cost coefficient and is unrelated to the ridge parameter λ in Theorem 6. Sensitivities at $\lambda \in \{0, 2.5\times10^{-6}, 5\times10^{-6}, 10^{-5}\}$ test the sensitivity to this proxy. The anchor reference always uses Qwen3. The state-independent reference makes the same number of GPT-OSS calls according to a deterministic ordering that does not use state or outcomes. The primary bootstrap resamples 50,000 times over paired $(\text{benchmark}, \text{id})$ blocks, stratified by benchmark, with multilingual rows kept together. The decision rule requires both Bonferroni one-sided 97.5% lower bounds to exceed zero. The evaluation also checks coverage, the number of activated blocks, family-level gains, and cost sensitivity.

The held-out result is a one-shot paired test on an offline full-information trace. It contains 1,618 rows, 732 independent blocks, and 315 activated rows from 315 blocks. Row coverage is 0.19468 and independent block coverage is 0.43033. Mean gains are 0.02184 against the anchor and 0.04889 against the state-independent selector. Corresponding lower bounds are 0.00392 and 0.02966. The selected-model utility charges the token proxy. It does not price acquiring the paired shadow outcome in an online deployment. LiveCodeBench, McEval, and MultiPL-E anchor gains are 0.05076, 0.00605, and exactly zero. Gains remain positive across all ten generation groups and all tested cost sensitivities. Replaying the information schedule confirms that no outcome affects a decision made before that outcome matures.

E.4 Delayed replay and frozen-proposal authorization audit

We additionally test the complete operational sequence on the frozen CALIBRI trace. Within each benchmark, base problems retain their source-file order. The first half forms a calibration stream and the second half is held back as the deployment suffix. During calibration, an outcome-free hash samples half of the base-problem blocks. The anchor serves the request, and a paired candidate outcome is acquired only when the sampled block is selected by either the candidate rule or the call-count-matched blind reference. The authorization look occurs once, after the calibration prefix. Outcomes from the final two blocks of each benchmark remain immature and do not enter the ledger.

The look uses 50,000 paired base-problem bootstrap resamples stratified by benchmark. Each reference receives a one-sided error level of 0.025. Both lower bounds clear zero, so the fixed RankShift proposal is deployed on the source-order suffix. Table 10 reports the complete accounting. The end-to-end value assigns zero incremental output value during startup, subtracts the token proxy for every shadow candidate call, and then adds only the selected-model utility realized on the suffix. This is a replay on real paired outcomes. It does not represent a live-traffic experiment.

Table 10: Delayed partial-feedback replay on CALIBRI. The primary row uses a 50% calibration prefix, 50% shadow sampling, and a two-block feedback delay within each benchmark. Net gain includes startup fallback and every shadow-candidate token cost.

Calib.	Shadow	Delay	LCB_b	LCB_s	Suffix gain	End-to-end gain
50%	50%	2	**+.0050**	**+.0141**	**+.00850**	**+.00335**
50%	25%	2	+.0172	+.0251	+.00850	+.00384
50%	50%	8	+.0047	+.0130	+.00850	+.00341
70%	50%	2	+.0081	+.0271	+.01331	+.00277

The primary replay acquires 1,300 shadow candidate outputs during calibration and makes 1,550 candidate calls on the deployment suffix. Across all 16,180 row-generation events, these correspond to 8.0% evidence-acquisition calls and 9.6% deployed candidate calls. The positive result persists at a longer eight-block delay and under 25% shadow sampling. Under the 100% shadow-sampling sensitivity condition, weaker early-prefix blocks lower the anchor bound, so RankShift retains the deployed anchor. The table is a sensitivity analysis of sampling and early-prefix composition. It makes no causal claim about shadow-call volume and evidence quality. It shows that the deployment decision can change when the observed prefix or feedback delay changes.

Matched router comparison. We evaluate RouteLLM, RouterDC, and Avengers-Pro on the same two-model CALIBRI trace (Ong et al., 2025; Chen et al., 2024b; Zhang et al., 2025). Their adapters retain the routing principle used by each method. They use pairwise preference scoring

for RouteLLM, outcome-weighted contrastive query scoring for RouterDC, and semantic clustering with within-cluster expert ranking for Avengers-Pro. The common interface supplies the frozen BGE-M3 representation, model pool, labels, split, and call budget. The resulting values compare routing rules under one task. They do not compare scores from different original protocols.

All three methods use the same training, validation, and test rows and exactly 315 GPT-OSS calls on test. The primary comparison fixes seed 2027. RouteLLM and Avengers-Pro also have five-seed sensitivity records. RouterDC is primary-seed only. The primary seed receives the same 50,000-resample paired problem-block bootstrap used throughout the unified comparison. The 315-call mask is formed by stable score ordering with row-index tie-breaking after fitting. The test outcomes are not used to choose that mask.

Composition with a fixed proposal. We also hold each realized proposal score fixed and apply RankShift as an outer authorization layer. Matured validation outcomes determine which task families may retain the proposal's candidate calls. The test split is used only for the final evaluation, and no future test score or outcome is used by the authorization decision. RankShift authorizes supported subsets of the CSCR, LENS, and StageRoute proposals, retaining 213, 214, and 211 calls. Their resulting gains are $+0.0224$, $+0.0227$, and $+0.0237$, with anchor lower bounds of $+0.0079$, $+0.0081$, and $+0.0095$, respectively. All three remove the proposal's negative worst-family gain. The other nine proposals lack sufficient validation evidence and revert to the anchor. We report these cases as fallbacks, not positive certificates, because both lower bounds equal zero. The supplementary material provides every before-and-after row and verifies that each wrapped test mask is a subset of its original proposal mask. This experiment isolates the value of authorization. RankShift preserves supported proposal decisions without retraining the router or turning unsupported proposals into positive results.

Sequential validation replay. As an execution audit, we replayed the frozen primary proposal actions in increasing request order, with family authorization read solely from the validation ledger and a 315-call upper bound. The replay reads no test utility or test score while making a decision and reproduces 213, 214, and 211 calls for CSCR, LENS, and StageRoute, respectively, and zero calls for the other nine proposals. The bundled privacy-preserving units omit the token-cost proxy, so the receipt reports raw quality gains separately from the ledger's cost-adjusted gains. This sequential execution audit uses the frozen proposal actions and introduces no new router fit.

Table 11: Validation-frozen threshold and quota audit. The target operating ratio is $315/1{,}618$. On the 1,611-row validation split this registers 314 validation calls. On test, the fixed threshold is applied in request order with a hard online cap of 315 calls, so no future test score is ranked or used for tuning. The primary row uses seed 2027. The five-seed columns average the same policy over seeds $2027, 3407, 17, 42, 73$.

	Primary seed 2027			Five-seed mean		
Proposal	Calls	Gain	Wrapped calls	Calls	Gain	Wrapped gain
CSCR	315	+.0177	214	315.0	+.0177	+.0226
LENS	315	+.0195	220	315.0	+.0195	+.0227
StageRoute	314	+.0151	211	314.8	+.0172	+.0242

Table 11 reports three proposals from the five-seed threshold/quota audit. The package contains the complete 12×5 audit using the shared matched implementations, including both proposal and wrapped roles. Some adapters differ from the primary released-code runs. The request-order cap and pass/fail rule are shared. This receipt makes the uniform validation procedure checkable for every method, including rows that remain fallbacks. It does not turn a fallback into a positive certificate.

Table 12: Five-seed validation-only wrapper sensitivity. The same validation pass/fail rule is applied to every learned adapter. "Active" is the number of seeds (out of five) that pass both lower bounds and the worst-family check. Calls and gain are averaged over all five seeds. A zero means validation rejected the proposal and the test policy falls back to the anchor. RouterDC is primary-seed only and is reported in Table 13.

Proposal	Active	Wrapped calls	Wrapped gain	Proposal gain
EmbedLLM-KNN	0/5	0.0	+.0000	+.0142
HybridLLM	2/5	77.0	+.0090	+.0128
Meta-Router	0/5	0.0	+.0000	+.0192
GraphRouter	0/5	0.0	+.0000	+.0166
RouteLLM-MF	0/5	0.0	+.0000	+.0110
Avengers-Pro	0/5	0.0	+.0000	+.0098
CSCR	5/5	213.0	+.0224	+.0165
FrugalGPT	0/5	0.0	+.0000	−.0067
LENS	5/5	214.0	+.0227	+.0167
DiSRouter	0/5	0.0	+.0000	+.0170
StageRoute	5/5	208.6	+.0241	+.0166

Table 13: Fixed proposal routers before and after RankShift authorization on held-out CALIBRI. Each unwrapped proposal row is exactly the same seed-2027 version reported in Table 1. There is no second adapter run or cache. Each pair uses the same frozen proposal scores. RankShift uses matured validation evidence to retain supported candidate calls and uses the test split once for final evaluation. The primary seed is 2027. Qualification requires both one-sided 97.5% lower confidence bounds to be positive and worst-family gain to be nonnegative.

Proposal router	Variant	Calls	Mean	Worst	LCB_b	LCB_s	Decision
CSCR	Proposal	315	+.0165	−.0103	−.0006	+.0249	NQ
	+ RankShift	213	**+.0224**	**.0000**	**+.0079**	**+.0241**	**Qualified**
LENS	Proposal	315	+.0167	−.0103	−.0003	+.0252	NQ
	+ RankShift	214	**+.0227**	**.0000**	**+.0081**	**+.0243**	**Qualified**
StageRoute	Proposal	315	+.0140	−.0166	−.0029	+.0232	NQ
	+ RankShift	211	**+.0237**	**.0000**	**+.0095**	**+.0257**	**Qualified**
EmbedLLM-KNN	Proposal	315	+.0142	−.0080	−.0037	+.0224	NQ
	+ RankShift	0	.0000	.0000	.0000	.0000	Fallback
HybridLLM	Proposal	315	+.0065	−.0221	−.0094	+.0173	NQ
	+ RankShift	0	.0000	.0000	.0000	.0000	Fallback
Meta-Router	Proposal	315	+.0171	−.0032	−.0003	+.0254	NQ
	+ RankShift	0	.0000	.0000	.0000	.0000	Fallback
GraphRouter	Proposal	315	+.0153	−.0195	−.0014	+.0240	NQ
	+ RankShift	0	.0000	.0000	.0000	.0000	Fallback
RouteLLM-MF	Proposal	315	+.0094	−.0152	−.0090	+.0173	NQ
	+ RankShift	0	.0000	.0000	.0000	.0000	Fallback
Avengers-Pro	Proposal	315	+.0115	−.0117	−.0067	+.0189	NQ
	+ RankShift	0	.0000	.0000	.0000	.0000	Fallback
FrugalGPT	Proposal	315	−.0067	−.1111	−.0168	+.0067	NQ
	+ RankShift	0	.0000	.0000	.0000	.0000	Fallback
DiSRouter	Proposal	315	+.0170	−.0091	−.0003	+.0255	NQ
	+ RankShift	0	.0000	.0000	.0000	.0000	Fallback
RouterDC	Proposal	315	+.0161	−.0038	−.0010	+.0246	NQ
	+ RankShift	0	.0000	.0000	.0000	.0000	Fallback

E.5 Baseline coverage and matched evaluation

Table 14 organizes the comparison by scientific role. Direct rows share the fixed Qwen/GPT-OSS pool, deployment-visible inputs, 315-call budget, token proxy, CALIBRI split, and paired block evaluation. Related rows position methods with a different data or training contract.

Table 14: Baseline coverage is organized by scientific role and protocol match. "Direct" denotes a numerical comparison under the common CALIBRI contract.

Role	Methods	Evidence	Reason for inclusion or separation
Simple controls	Anchor, always candidate, always fallback, uniform random, registered state-independent hash	Direct	Establishes the deployed floor, full-budget ceiling, system-abstention floor, and blind allocation frontier.
Mechanism-aligned	RouterDC, RouteLLM, Avengers-Pro	Direct, labeled	Central published mechanisms trained and evaluated under the common trace.
Released methods	EmbedLLM, GraphRouter, Meta-Router	Direct adaptation	Public algorithm, architecture, or scoring components connected to the common two-model interface.
Public code	HybridLLM, CSCR	Direct adaptation	Public framework or training components evaluated with the same split and budget.
Paper rebuilds	FrugalGPT, LENS, DiSRouter, StageRoute	Direct, labeled	Published routing principles evaluated through the common interface.
Adjacent routers	Confidence Tokens, unified routing and cascading (Chuang et al., 2025; Dekoninck et al., 2025)	Related	Require learned confidence tokens or a distinct joint routing-cascade training contract.
Dynamic pools	UniRoute (Jitkrittum et al., 2026)	Related	Targets unseen-model transfer, which the fixed execution-labeled pair does not cover.
Risk controls	LEC, TRACER (Wang et al., 2026; Rida, 2026)	Related	Control selection-conditioned error or teacher parity. They do not target delayed dual-reference utility.
Delayed agents	TRACE-Router (Raj et al., 2026)	Related	Assigns an entire agentic task using terminal reward. Its decision unit and latency objective differ from request-level code routing.

The expanded comparison uses the strongest available implementation source for each matched adapter. EmbedLLM uses its released KNN algorithm, GraphRouter its released graph architecture, and Meta-Router, HybridLLM, and CSCR their public scoring, framework, or training components. FrugalGPT, LENS, DiSRouter, and StageRoute use implementations of their published routing principles. Table 14 records the source category for each row.

UniRoute addresses a dynamic pool that can contain models unseen during training (Jitkrittum et al., 2026), a condition absent from the fixed Qwen/GPT-OSS trace. LEC and TRACER are adjacent authorization-related controls. They target selection-conditioned risk or a one-reference parity criterion (Wang et al., 2026; Rida, 2026). The present decision targets delayed dual-reference utility. These methods define adjacent risk-control directions. The numerical table focuses on routers that admit the same information, model pool, call budget, and evaluation contract. Theorem 1 explains the structural difference. Under the same-budget policy class, a single baseline-adjusted utility test

cannot in general certify both production improvement and comparator-relative state attribution. The joint rule remains distinct from either one-reference family.

AutoMix and LLM-Blender are important adjacent mixture and ensemble systems (Aggarwal et al., 2024; Jiang et al., 2023), but their verifier/ensemble outputs, multi-model pools, and training contracts do not fit the frozen two-model, paired-outcome CALIBRI interface. We therefore record them as an unmatched comparison limitation and omit a non-comparable number. For the same reason, the table omits matched numbers for delayed/censored-feedback and conservative-policy baselines. Their action units, feedback contracts, or safety targets differ from this record. The omission records a scope limitation and supplies no dominance claim over those methods.

The transfer domains use the same structure. Table 15 marks methods evaluated on the same trace as "Direct." It marks frozen experts as "Pool" and methods with a different data or training contract as "Related." Every numerical comparison uses the domain's own frozen trace and metric.

Table 15: Cross-domain baseline ladder.

Domain	Tier	Methods	Evidence
Finance	Simple	rolling window, exponential decay, elastic model, LightGBM	Direct/pool
Finance	Canonical	ALSTM, Transformer	Direct
News	Simple	best fixed, fixed share, delayed Hedge, state-free GBDT, CTR/freshness	Direct/pool
News	Canonical	Chronological NRMS (EMNLP 2019), NewsRecLib-MINER (ACL Findings 2022)	Direct
Commerce	Simple	global/session popularity, transition or Markov, recency, metadata rules	Direct/pool
Commerce	Canonical	GRU4Rec (Hidasi et al., 2016), NARM (Li et al., 2017), SR-GNN (Wu et al., 2019b)	Pool

Tables 16–18 report the protocol-matched transfer comparison. All values use the common chronological trace and native metric for each dataset. Finance uses mean RankIC and an adapted MASTER architecture on our Qlib split. MIND uses the same impressions and NDCG@10 for every method, including our chronological NRMS implementation and the pinned NewsRecLib-MINER implementation. Coveo freezes SR-GNN as the anchor and exposes NARM and GRU4Rec to the router. Amazon-M2 uses its strongest fixed expert as both registered references. Hence $s_t \equiv b_t$ there and Gate 2 is mathematically identical to Gate 1. This separate experiment should be read as anchor-risk control, not as an additional state-attribution contrast.

Table 16: RankShift leads all protocol-matched finance models. Mean RankIC on the same Qlib splits and five seeds. Adapted MASTER uses the released architecture with the common causal feature panel.

Dataset	Anchor	ALSTM	Transformer	Adapted MASTER	RankShift
CSI300	.07405	.07288	.07073	.04324	**.07598**
CSI500	.09552	.08797	.09242	.06287	**.09925**
Binance	.13511	.13577	.13367	.10915	**.13687**

Table 17: RankShift leads the protocol-matched MIND baselines. All values are mean NDCG@10 on the identical chronological impression trace. NewsRecLib-MINER and Chronological NRMS use the same split and inputs.

Method	NDCG@10	Method	NDCG@10
Best fixed / state-free GBDT	.35022	Delayed Hedge	.34631
Fixed Share	.30992	Chronological NRMS	.29606
NewsRecLib-MINER	.29681	**RankShift**	**.36303**

Table 18: Commerce comparisons preserve the deployed pool. Models in the pool are not reported as separate routing policies. RankShift selects among them and is evaluated against the frozen anchor on each local trace. These gains are not directly comparable with scores reported on other commerce datasets.

Dataset	Anchor/control	Available expert pool	RankShift gain
Coveo	SR-GNN	SR-GNN, NARM, GRU4Rec	$+.\mathbf{02395}$
Amazon-M2	best fixed	fixed, Markov, metadata experts	$+.\mathbf{05635}$

Table 1 reports the full seed-2027 comparison. Its twelve learned rows are exactly the unwrapped proposal rows in Table 13. No second adapter version or cache is used. None combines two positive lower bounds with nonnegative gain in every family. The separately labeled RankShift-Code heuristic does. Five-seed sensitivity is stored for eleven learned adapters. RouterDC has only the frozen primary-seed version, so no five-seed RouterDC claim is made.

At the fixed 315-call budget, a state-independent policy cannot rank requests from prompt, benchmark, language, or outcome information. The registered hash is one deterministic blind allocation, and the main-table registered uniform draw is one fixed random allocation. The 10,000 post-analysis uniform subsets measure the corresponding allocation distribution. The registered draw remains the primary comparator. Learning a request score would reintroduce state and would therefore cease to be a state-independent comparator. This distinction is why the blind reference and learned proposal routers occupy separate baseline tiers.

Table 1 separates aggregate proposal quality from the three quantities required by the joint criterion. The strongest proposal mean loses value on one benchmark family. RankShift-Code clears both reference bounds and stays nonnegative in every family.

Joint-criterion gain. For policy π, let $\Delta_b(\pi)$ be its mean gain over the anchor, $L_b(\pi)$ and $L_s(\pi)$ its lower bounds against the anchor and blind reference, and $\Delta_f(\pi)$ its mean gain in benchmark family f. We define

$$\mathrm{JG}(\pi) = \begin{cases} \Delta_b(\pi), & L_b(\pi) > 0,\ L_s(\pi) > 0,\ \min_f \Delta_f(\pi) \geq 0, \\ -, & \text{otherwise.} \end{cases}$$

This rule does not average the three requirements into a new score. The two lower bounds implement the dual-reference mean test, while the family condition is a separate empirical stability check outside Theorem 3. A dash means that at least one requirement fails. RankShift is the only policy with joint-criterion gain in this comparison, at $+0.02184$.

Call-count-matched random allocation. After the primary analysis, we drew 10,000 uniform subsets of 315 rows. Their mean gain was -0.01833 with central 95% interval $[-0.02882, -0.00784]$. No draw reached the RankShift gain $+0.02184$. Sampling one row per block gave mean -0.02855 with central interval $[-0.04172, -0.01545]$, again with zero draws reaching RankShift.

Authorization-layer latency. The microbenchmark updates both gain summaries, performs the state lookup and two LCB calculations, and returns route or fallback. On an Intel Core i5-14600KF CPU under Python 3.12.13, five runs give median batch-amortized latency 0.239–0.241 μs and batch 95th percentiles 0.250–0.277 μs. The median is 0.240 μs, or about 4.2 million in-memory decisions per second. This hot-path measurement excludes feature extraction, persistence, model inference, and shadow-outcome acquisition.

E.6 Public LLMRouterBench transfer

We test whether the same dual-reference identification principle holds on the public LLMRouterBench performance and cost release (Li et al., 2026). We generate no new model responses. Seven tasks contain complete aligned outcomes for all 13 flagship models, yielding 10,917 aligned prompts after excluding one prompt mismatch. A stable prompt-hash split with seed 3407 assigns 60% of prompts to fitting, 20% to authorization, and 20% to the final test. The fit split chooses the best

Table 19: Complete public LLMRouterBench model and policy comparison. All rows use the same 2,238 test prompts and registered cost-adjusted utility. Cost is mean dollars per request multiplied by 10^3. Δ_b and LCB_b compare with the fit-optimal fixed anchor. "Worst" is the minimum task-level gain. RankShift is the only state-conditioned policy whose switches are fixed by positive dual-reference validation bounds.

Policy	Role	Cost	Δ_b	LCB_b	Worst
Claude Sonnet 4	fixed model	6.609	−.1822	−.2019	−.4580
DeepSeek V3 0324	fixed model	.547	−.1142	−.1341	−.2975
DeepSeek V3.1 Terminus	fixed model	.779	−.1003	−.1202	−.3774
DeepSeek R1 0528	fixed model	12.351	−.1039	−.1238	−.2535
Gemini 2.5 Flash	fixed model	9.129	−.1237	−.1439	−.2781
Gemini 2.5 Pro	fixed model	52.824	−.1164	−.1384	−.4240
GPT-5 Chat	fixed model	3.869	−.0688	−.0879	−.1532
GPT-5	fixed model	27.044	−.0287	−.0500	−.0677
Qwen3-235B	fixed anchor	.733	.0000	.0000	.0000
Qwen3-235B Thinking	fixed model	6.537	−.0327	−.0506	−.0953
GLM-4.6	fixed model	15.263	−.1558	−.1771	−.2978
Kimi K2 0905	fixed model	1.079	−.1027	−.1229	−.2325
Intern-S1	fixed model	2.552	−.1718	−.1924	−.3922
Fit-optimized state-free mixture	comparator	.733	.0000	.0000	.0000
Task-aware proposal	not authorized	1.826	+.0198	+.0085	.0000
RankShift	**authorized**	1.799	**+.0105**	**+.0045**	**.0000**
Per-prompt oracle	hindsight only	4.246	+.2730	+.2550	+.0619

model within each task for the state-conditioned proposal and the best fixed model for the anchor. It also solves a linear program over all 13 models to obtain the state-independent mixture with the highest fit-split mean utility subject to its fit-split mean dollar cost not exceeding that of the proposal. The linear program ranges over all 13 models. The selected mixture is frozen before the authorization data are read and is a fitted comparator rather than a known population optimum.

For the registered normalized cost preference 0.1, Table 20 shows the two validation bounds. Only GPQA and LiveCodeBench clear both. The policy therefore executes the proposal on those task cells and uses the exact anchor elsewhere. On the untouched test, this switches 10.28% of requests. The fit-optimized mixture places all mass on Qwen3-235B, so the registered state-free policy coincides with the fit-optimal fixed anchor. Thus $s_t \equiv b_t$ and Gate 2 coincides with Gate 1 in this public 13-model transfer. RankShift acts here as an anchor-risk controller with registered bounds. RouterBench GSM8K shows the separate attribution role of Gate 2 in Table 22. The financial units show the same role in Table 8, where $s_t \neq b_t$. The deployed gain is $+0.0105$, and both lower bounds equal $+0.0045$. Before gating, the proposal gains $+0.0198$ against the anchor. The difference comes from task cells whose two margins remain unresolved. Across five auxiliary prompt-hash split seeds, all five deployed policies retain strictly positive lower bounds. Their mean gain is $+0.0235$, and the smallest lower bound is $+0.00046$.

Table 19 places every model and policy on the same test rows. The fixed-model rows compare the registered comparator with every model on the test split. The ungated proposal has more test-set gain, but five of its seven task cells do not clear the validation rule. RankShift converts the supported part of that opportunity into positive deployed value and preserves the anchor on the remaining task cells.

Table 20: Dual-reference authorization on the public 13-model LLMRouterBench trace. One-sided 97.5% paired normal bounds use only the 20% authorization split. The final test is untouched until the task decisions in the last column are fixed.

Task cell	Validation n	LCB_b	LCB_s	Authorize
AIME	12	−.0861	−.0861	no
GPQA	38	**+.1186**	**+.1186**	**yes**
HLE	434	−.0567	−.0567	no

continued on next page

Table 20 (continued)

Task cell	Validation n	LCB_b	LCB_s	Authorize
LiveCodeBench	198	$+.\mathbf{0568}$	$+.\mathbf{0568}$	**yes**
LiveMathBench	23	0.0000	0.0000	no
MMLU-Pro	605	$-.0374$	$-.0374$	no
SimpleQA	834	0.0000	0.0000	no

The task-cell pattern reveals two distinct mechanisms. AIME and LiveMathBench contain only 12 and 23 authorization prompts, so limited support is a plausible bottleneck. HLE and MMLU-Pro contain 434 and 605 prompts but retain negative lower bounds, so limited support is not the sole explanation. SimpleQA has 834 prompts and a zero margin, indicating no detectable state-conditioned advantage. GPQA and LiveCodeBench are the two cells that combine adequate support with positive margins against both references. The gate therefore distinguishes small-support cells from large-support tasks whose current proposal has a weak or zero margin.

E.7 Public MMLU and GSM8K extension

We next test whether the same dual-reference identification principle holds on two public tasks that are widely used in model-routing studies. We use the zero-shot RouterBench trace (Hu et al., 2024), which provides aligned correctness and cost records for MMLU (Hendrycks et al., 2021) and GSM8K (Cobbe et al., 2021). The experiment compares Mixtral-8x7B-Chat with GPT-4-1106-preview. All twelve proposals share the base prompt features. FrugalGPT also uses the observed anchor response through its cascade view. The data split, utility, and 25% candidate-call budget are shared. We vary only the training seed. RankShift uses a separate authorization split and one fixed, call-matched state-independent allocation shared by every proposal. These rows test the same mechanism under one protocol. They do not reproduce the original numbers reported by the baseline papers.

Table 21 shows a broad positive result on MMLU. Every proposal clears both references for all five training seeds. RankShift therefore retains the proposal decisions and obtains a held-out gain between $+.0653$ and $+.0861$ over the Mixtral anchor. The result is stable across different routing mechanisms. This suggests that the available prompt signal is strong enough to support deployment under this model pair and budget.

GSM8K shows a different pattern. All twelve proposals improve the anchor in mean. Eleven do not clear the state-independent lower bound. Their gains can therefore be explained by spending the same candidate budget without a reliably better prompt-dependent allocation. FrugalGPT is the only proposal that clears both references for all five seeds, and RankShift deploys it with a held-out gain of $+.0627$. This contrast shows the role of the second reference. It separates access to the stronger model from evidence that the routing rule uses the request state well.

This extension uses paired empirical normal bounds on a frozen public trace. It tests cross-task behavior under a controlled offline protocol. Live API traffic and the theorem's dependence-and-drift assumptions lie outside its scope.

Table 21: MMLU extension on the public RouterBench trace. Raw and deployed gains are held-out utility differences from the fixed anchor. LCBs are five-seed means computed only on the authorization split. "Pass" is the number of training seeds that clear both one-sided 97.5% paired bounds.

Proposal	Raw Δ_b	LCB_b	LCB_s	Pass	Deployed Δ_b
HybridLLM	$+.0813$	$+.0527$	$+.0091$	$5/5$	$+.0813$
RouterDC	$+.0824$	$+.0518$	$+.0085$	$5/5$	$+.0824$
RouteLLM-MF	$+.0834$	$+.0575$	$+.0141$	$5/5$	$+.0834$
GraphRouter	$+.0857$	$+.0562$	$+.0125$	$5/5$	$+.0857$
FrugalGPT	$+.0821$	$+.0537$	$+.0100$	$5/5$	$+.0821$

continued on next page

Table 21 (continued)

Proposal	Raw Δ_b	LCB_b	LCB_s	Pass	Deployed Δ_b
CSCR	$+.0861$	$+.0590$	$+.0155$	$5/5$	$+.0861$
EmbedLLM-KNN	$+.0813$	$+.0562$	$+.0126$	$5/5$	$+.0813$
Avengers-Pro	$+.0830$	$+.0616$	$+.0180$	$5/5$	$+.0830$
Meta-Router	$+.0853$	$+.0584$	$+.0150$	$5/5$	$+.0853$
LENS	$+.0861$	$+.0590$	$+.0155$	$5/5$	$+.0861$
DiSRouter	$+.0653$	$+.0505$	$+.0065$	$5/5$	$+.0653$
StageRoute	$+.0856$	$+.0599$	$+.0164$	$5/5$	$+.0856$

Table 22: GSM8K extension on the public RouterBench trace. The protocol and columns match Table 21. A zero deployed gain means that RankShift retains the registered anchor.

Proposal	Raw Δ_b	LCB_b	LCB_s	Pass	Deployed Δ_b
HybridLLM	$+.0426$	$+.0339$	$-.0053$	$0/5$	.0000
RouterDC	$+.0411$	$+.0321$	$-.0067$	$0/5$	.0000
RouteLLM-MF	$+.0415$	$+.0340$	$-.0049$	$0/5$	.0000
GraphRouter	$+.0446$	$+.0319$	$-.0074$	$0/5$	.0000
FrugalGPT	$+.0627$	$+.0529$	$+.0142$	$5/5$	$+.0627$
CSCR	$+.0433$	$+.0343$	$-.0048$	$0/5$	.0000
EmbedLLM-KNN	$+.0425$	$+.0298$	$-.0096$	$0/5$	.0000
Avengers-Pro	$+.0392$	$+.0297$	$-.0095$	$0/5$	.0000
Meta-Router	$+.0431$	$+.0343$	$-.0048$	$0/5$	.0000
LENS	$+.0423$	$+.0320$	$-.0071$	$0/5$	.0000
DiSRouter	$+.0431$	$+.0289$	$-.0104$	$0/5$	.0000
StageRoute	$+.0430$	$+.0333$	$-.0058$	$0/5$	.0000

E.8 Finance cost sensitivity

Table 23 evaluates every registered transaction-cost level under both raw and winsorized returns. This separates gains that persist under cost from conclusions that depend on a particular treatment of return tails.

Table 23: Financial returns under four transaction-cost assumptions.

data	bps	return type	gain	LCB	UCB	p
Binance	0	raw	3.41e-06	-0.0001	0.0001	0.9666
Binance	0	winsorized	4.62e-05	-6.20e-05	0.0001	0.4246
Binance	5	raw	-1.86e-06	-0.0001	0.0001	0.9044
Binance	5	winsorized	4.10e-05	-6.64e-05	0.0001	0.4894
Binance	10	raw	-7.12e-06	-0.0001	0.0001	0.8524
Binance	10	winsorized	3.57e-05	-7.29e-05	0.0001	0.5646
Binance	20	raw	-1.77e-05	-0.0002	0.0001	0.7448
Binance	20	winsorized	2.52e-05	-7.90e-05	0.0001	0.6678
CSI300	0	raw	-3.62e-06	-0.0001	8.90e-05	0.8426
CSI300	0	winsorized	1.38e-06	-0.0001	8.98e-05	0.9244
CSI300	5	raw	1.78e-06	-0.0001	9.66e-05	0.8988
CSI300	5	winsorized	6.78e-06	-0.0001	9.40e-05	0.9840
CSI300	10	raw	7.18e-06	-0.0001	0.0001	0.9990
CSI300	10	winsorized	1.22e-05	-9.54e-05	0.0001	0.9134
CSI300	20	raw	1.80e-05	-9.38e-05	0.0001	0.8486
CSI300	20	winsorized	2.30e-05	-8.47e-05	0.0001	0.7342
CSI500	0	raw	0.0002	9.48e-05	0.0003	0.0000
CSI500	0	winsorized	0.0002	8.95e-05	0.0003	0.0000
CSI500	5	raw	0.0002	0.0001	0.0003	0.0000

continued on next page

Table 23 (continued)

data	bps	return type	gain	LCB	UCB	p
CSI500	5	winsorized	0.0002	9.64e-05	0.0003	0.0000
CSI500	10	raw	0.0002	0.0001	0.0003	0.0000
CSI500	10	winsorized	0.0002	0.0001	0.0003	0.0000
CSI500	20	raw	0.0002	0.0001	0.0003	0.0000
CSI500	20	winsorized	0.0002	0.0001	0.0003	0.0000

F TRANSFER EXPERIMENTS AND PROTOCOLS

These experiments test whether the same dual-reference identification principle holds when utilities, expert pools, state maps, and feedback clocks change. Their role is to test transfer of one decision principle.

Throughout this section, c, s, and b denote the state-conditioned proposal, state-independent reference, and fallback anchor. Development and held-out results use separate chronological splits.

Finance. The candidate is a calibrated-anchor delayed ridge with at most 0.25 adaptive mass. Its state summarizes the current expert-prediction slate through cross-sectional size, dispersion, distance from the anchor, and correlation with the anchor. The state-independent reference uses the same rule with the state vector set to zero. The anchor is the expert with the largest mean RankIC on the preceding calibration block. Feedback arrives one trading day later, and daily utilities form 36 monthly blocks per development unit. CSI300 and CSI500 use development years 2014–2016 and held-out years 2017–2020. Binance uses 2019–2021 and 2022–2025. The primary metric is daily cross-sectional RankIC. A separate economic analysis reports raw and winsorized long–short returns after 0, 5, 10, and 20 bps transaction costs.

MIND. Per-expert HistGradientBoosting models predict advantage from a 45-dimensional, label-free state containing time, slate, history, novelty, item-age, and cross-expert disagreement features. The state-independent models use only summaries from the previous mature block, and the anchor is the best fixed expert selected on the core blocks only, as specified by the frozen preregistration. Feedback comes from the preceding one-hour block. Each development unit contains 24 blocks and 31,494 impressions. The held-out analysis uses an 80% chronological prefix and 20% suffix. Utility is impression NDCG@10, with an eight-hour moving-block statistic and no model-inference cost term.

EB-NeRD. The candidate uses per-expert HistGradientBoosting models with an 18-dimensional state built from user, device, time, slate, novelty, exposure, CTR, article-age, and expert-disagreement features. The state-independent reference uses the same architecture on lagged utilities, and the anchor is the best fixed expert. Feedback comes from the preceding 30-minute impression block. Each development unit contains 67 blocks and 131,825 impressions. The chronological split is 60% fit, 20% development, and 20% held out. Utility is impression NDCG@10 with an eight-block moving statistic.

Coveo. A multi-output ridge model selects the expert with the largest predicted advantage above 0.0096724 and otherwise uses SR-GNN. Its 34-dimensional state combines session-origin features with agreement among SR-GNN, NARM, and GRU4Rec. The state-independent reference chooses by fit-period mean advantage under the same expert pool and threshold. Utility matures at the target event and is aggregated by target-time hour. Each development unit contains 433 hour blocks and 224,922 aligned sessions. Utility is session MRR@20. The held-out suffix does not update the routing rule.

Amazon-M2. The candidate uses a state-aware HistGradientBoosting rule when its predicted advantage over the state-independent selector clears the development threshold. The state contains locale, session length, candidate popularity, metadata, price, brand, expert-margin, and disagreement features. Feedback is delayed by eight slates and summarized over a 64-slate rolling window. Statistics use 20 contiguous blocks per locale. The evaluation reports sampled-candidate MRR@100

for five seeds across six Task 1 and Task 2 locales. This experiment is separate from the 35-unit development analysis.

Expert pools. Finance compares rolling-window, exponential-decay, LightGBM, elastic, ALSTM, and Transformer experts, giving 15 eligible experts on CSI300/CSI500 and 13 on Binance. MIND combines CTR, history-affinity, and freshness experts. Its anchor is the best fixed expert selected on the core blocks only, as specified by the frozen preregistration. It is not a global test-period CTR oracle. EB-NeRD uses causal all-history and rolling CTR. Coveo combines popularity, transition, recency, SR-GNN, NARM, and GRU4Rec. Amazon-M2 uses popularity, Markov, brand-price, and title-based experts.

G SUPPORT AND COMPLEXITY ANALYSIS

The controlled study tests the mechanism predicted by the activation-boundary results in Section 4. RankShift authorizes a route only after each reference has accumulated enough mature evidence to resolve its margin. Increasing state complexity divides that evidence across more cells or more feature directions, while delay and noise reduce the effective information in each observation. The result is a clear change in activation. Simple state maps become identifiable early. Richer representations require stronger support before the same decision can be authorized.

Table 24 summarizes the case shown in Figure 2. The finite-state controller reaches the target with 128 mature blocks when the state is pooled into one cell. The continuous controller scales gradually from 128 blocks in one dimension to 512 blocks in five dimensions. More complex settings do not reach the target within the tested range of 512 blocks. These outcomes follow the theoretical prediction that the evidence requirement grows with state complexity and shrinks with the identifiable dual-reference margin.

Table 24: State complexity determines the mature support needed for authorization. Entries report the minimum tested support needed to reach 80% activation power in the case shown in Figure 2. > 512 means that the target was not reached within the tested range.

Controller	State complexity	Minimum mature support
Finite-state	$K = 1$	128
Finite-state	$K \in \{20, 80, 6480\}$	> 512
Continuous-feature	$d = 1$	128
Continuous-feature	$d = 3$	256
Continuous-feature	$d = 5$	512
Continuous-feature	$d = 10$	> 512

This transition matters in practice. RankShift shows the cost of a richer state representation before deployment. Users can choose fewer states when support is scarce and add detail as mature evidence grows. The method ties state design to the available support. The supplementary results store the full sweep over delay, noise, effect size, and complexity in machine-readable form. This record reproduces the summarized boundary without turning the appendix into an experiment log.