PACE: Cost-Efficient LLM Judging with Human-Agreement Control via Pre-Routing

Anonymous authors
Paper under double-blind review

Abstract

LLM-as-a-Judge offers a scalable alternative to human evaluation, but disagreement with human preferences remains a central reliability challenge, and improving agreement through stronger Judges or repeated evaluations can be costly. We introduce PACE (Pre-routing with Agreement-Controlled Evaluation), a selective-evaluation framework that routes each instance to a single Judge tier before evaluation. Using labeled calibration data, PACE estimates each verdict's disagreement risk within its routing group, then pools the verdicts and reports the largest subset with controlled overall disagreement risk. Routing groups may be learned from the unlabeled test batch, introducing dependencies that standard fixed-group analyses do not address. Under leave-one-out routing stability and regularity conditions, we quantify the effect of routing instability on risk estimation and establish asymptotic mFDR control at level α for the reported verdicts. In comparisons with cascades and other baselines, PACE yields competitive agreement–coverage trade-offs with one evaluation record per instance while substantially reducing cost.

1 Introduction

Evaluation is critical infrastructure for developing and deploying large language models. Human assessment remains the reference standard for open-ended outputs, but is costly and difficult to scale consistently (Van Der Lee et al., 2019; Zheng et al., 2023); automatic metrics such as BLEU and ROUGE are inexpensive and repeatable, but lexical overlap can align poorly with human judgments of open-ended generation (Papineni et al., 2002; Lin, 2004; Novikova et al., 2017). To bridge this gap, LLM-as-a-Judge has emerged as an increasingly important and widely used evaluation paradigm. It can score individual responses, compare candidate responses, and produce criterion-specific critiques (Liu et al., 2023; Zheng et al., 2023; Gu et al., 2026).

Despite their broad adoption, LLM Judges do not always agree with human preferences. If left uncontrolled, such disagreement can distort model rankings, misdirect model selection, and propagate erroneous signals into downstream development and deployment decisions (Gu et al., 2026). Disagreement should therefore be treated as a risk that an evaluation system must explicitly control (Badshah et al., 2026; Jung et al., 2025). Let X denote a sample, Y its human preference, and $\widehat{Y}$ the Judge's verdict. Given a user-specified tolerance α, we formalize this objective as controlling the population-level disagreement rate among reported verdicts:

$$\text{Disagreement} := \Pr\left(\widehat{Y} \neq Y \mid X \text{ is reported}\right) \leq \alpha. \tag{1}$$

Relevant approaches fall into two broad strands. The first improves empirical Judge quality through structured prompting (Liu et al., 2023), specialized Judge training (Zhu et al., 2025; Kim et al., 2024), or multi-Judge aggregation (Verga et al., 2024). Although these methods can increase human agreement empirically, they do not guarantee that disagreement among reported verdicts remains below a user-specified tolerance. The second strand differs fundamentally by seeking a statistical guarantee on this disagreement rate. These methods use historical examples paired with human preferences to calibrate a reporting rule that identifies when a Judge's verdict is sufficiently trustworthy. Trust or Escalate (Jung et al., 2025) is a representative example: it queries Judge tiers from weaker to stronger, reports a verdict when the current tier is sufficiently trusted, and otherwise escalates the instance to a stronger Judge. This cascade demonstrates the value of adapting Judge capability and cost to instance

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.

difficulty while targeting a prespecified disagreement level. However, its sequential design presents two limitations for our setting. First, instances reaching a later Judge have already been selected by earlier Judges, so i.i.d.-based guarantees must be applied with great care at later tiers. Second, a difficult instance may require multiple Judge evaluations before it is reported or rejected, increasing evaluation cost.

Together, these limitations motivate a complementary design question:

Can LLM judging control human disagreement with statistical guarantees at lower cost?

In this work, we propose **PACE** (**P**re-routing with **A**greement-**C**ontrolled **E**valuation), a framework that separates Judge assignment from disagreement control. PACE uses a calibration set with human preference labels to support selective reporting on an unlabeled test batch. It proceeds in three steps: (1) A router assigns each instance to exactly one Judge tier before evaluation, without inspecting human labels or Judge outputs. This lets the system match capability and cost to the input: a lower-cost Judge may suffice for an easier instance, while a harder one may warrant a stronger Judge. (2) Human preferences in the calibration set are used to estimate group-conditional disagreement risks for the routed verdicts. (3) PACE pools these risks across tiers and reports the largest subset whose average estimated risk does not exceed the user-specified tolerance α. By invoking only the assigned Judge for each instance, PACE avoids repeated evaluations while controlling disagreement over the pooled reported set.

This construction introduces a theoretical challenge: the routing groups are data dependent, so standard fixed-group analyses do not directly apply. Unlike fixed-group analysis, PACE allows the routing partition to be learned from the unlabeled test batch, while the final reported set is itself selected adaptively from estimated disagreement risks. We develop a stability-based theory for this two-stage adaptivity and establish asymptotic mFDR, equivalently human disagreement, control for the pooled reporting rule. The conditions accommodate a broad class of stable data-driven routers, and we verify them for empirical-quantile routing and regularized learned gates.

Experiments on ChatArena, TL;DR, and MT-Bench show that PACE often meets or closely approaches the target agreement with competitive coverage and lower cost than most baselines. Routing-design and component ablations clarify the roles of Judge assignment, disagreement-risk estimation, and pooled risk budgeting. Additional experiments examine stricter agreement targets, alternative Judge panels, and cross-dataset behavior.

Our contributions are as follows:

1. **PACE.** We introduce a pre-routing framework that combines one-time Judge assignment with pooled risk control to control disagreement at low cost.
2. **Theoretical guarantees.** We establish risk estimator consistency and asymptotic disagreement control under stable data-dependent routing.
3. **Experiments.** Across three datasets, PACE demonstrates competitive agreement–coverage–cost trade-offs relative to external baselines.

2 Related Work

LLMs as Judges. Methods without explicit statistical guarantees improve Judge quality through structured prompting (Liu et al., 2023), specialized training (Zhu et al., 2025; Kim et al., 2024), bias mitigation (Wang et al., 2024), panel aggregation (Verga et al., 2024), or learned correction of human–LLM deviations (Polo et al., 2025). They evaluate improvements empirically rather than certifying a prespecified disagreement level. Methods with statistical guarantees instead use labeled calibration data to determine when a judgment may be reported. Trust or Escalate (Jung et al., 2025) combines fixed-sequence threshold calibration, Simulated Annotators as a confidence estimator, and a weak-to-strong cascade. SCOPE (Badshah et al., 2026) constructs an order-invariant entropy score from two response permutations and calibrates its report threshold under exchangeability.

Statistical risk control. Reliable selective evaluation requires controlling errors among reported outputs. False discovery rate (FDR) and marginal FDR (mFDR) formalize this objective in multiple testing (Benjamini & Hochberg, 1995; Genovese & Wasserman, 2002). Our disagreement criterion

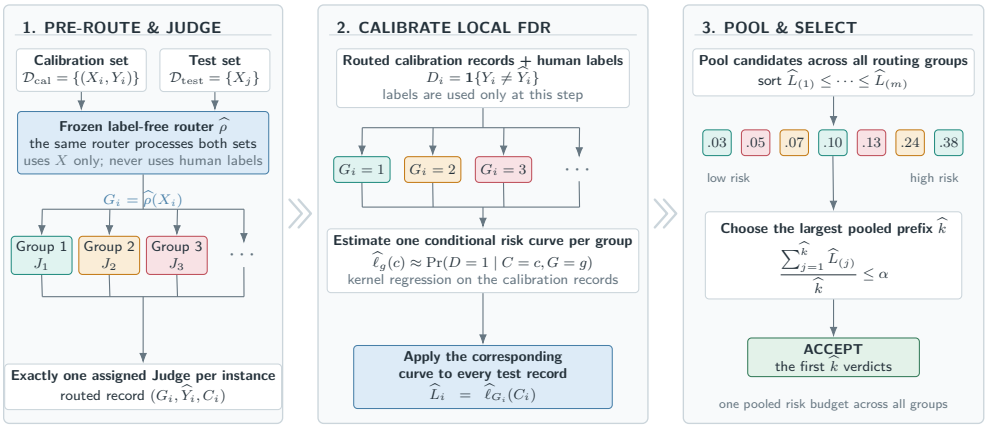

Figure 1: Overview of PACE.

corresponds to mFDR, for which local-FDR methods guide selection using estimated error probabilities under a shared risk budget (Sun & Cai, 2007; Cai & Sun, 2009). For predictive systems, calibration-based methods use labeled data to choose decision rules with guarantees for specified loss functions (Angelopoulos et al., 2025; 2024). Learning routing groups from the same test batch introduces additional dependence, linking our setting to post-selection inference (Bao et al., 2024).

3 Preliminaries

Let $X = (Q, a_1, a_2)$ denote a pairwise evaluation instance, where Q is the user request and a_1, a_2 are the candidate responses. Its human preference label is $Y \in \{1, 2\}$, meaning that a_Y is preferred. We consider a heterogeneous panel $\mathcal{J} = \{J_1, \ldots, J_M\}$ whose Judges differ in capability and cost. Applying the strongest Judge to every instance is unnecessarily expensive when a less costly Judge is already reliable on easier instances. PACE therefore uses a router $\rho : \mathcal{X} \to [M]$ to assign each instance to an appropriate Judge before evaluation. For each instance X_i, let $\widehat{Y}_i \in \{1, 2\}$ and $C_i \in (0, 1)$ denote the preference and confidence produced by its assigned Judge $J_{\rho(X_i)}$.

We observe a labeled calibration set $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$ and a test set $\mathcal{D}_{\text{test}} = \{X_i\}_{i=1}^m$, both drawn from the same distribution. Test labels are unavailable to the procedure. Define the disagreement and report indicators as $D_i = \mathbf{1}\{Y_i \neq \widehat{Y}_i\}$ and $R_i = \mathbf{1}\{i \text{ is reported}\}$. Given the routing assignments, the population-level reporting objective is to maximize expected coverage while controlling the marginal false discovery rate (mFDR) at the tolerance level α:

$$\max_{R_1,\ldots,R_m} \quad \frac{1}{m}\mathbb{E}\left[\sum_{i=1}^m R_i\right] \quad \text{subject to} \quad \text{mFDR} := \frac{\mathbb{E}\left[\sum_{i=1}^m D_i R_i\right]}{\mathbb{E}\left[\sum_{i=1}^m R_i\right]} \leq \alpha.$$

4 Method

PACE operates in the three stages summarized in Figure 1. First, a router assigns each instance to a single Judge tier, so each instance receives exactly one formal evaluation. Second, PACE uses the labeled calibration set to map each group-specific confidence score to an estimated local disagreement risk. Finally, it pools these estimated risks across groups and reports the largest set whose average risk is at most α.

4.1 Pre-Routing to a Heterogeneous Judge Panel

For each instance X_i, PACE sets $G_i = \rho(X_i)$ and invokes only the corresponding Judge J_{G_i}. Writing $\widehat{Y}^{(g)}(x)$ and $C^{(g)}(x)$ for the preference and confidence produced by Judge J_g, the routed evaluation yields

$$\widehat{Y}_i := \widehat{Y}^{(G_i)}(X_i), \qquad C_i := C^{(G_i)}(X_i).$$

Thus, the Judge assignment is fixed before evaluation, and each instance produces exactly one evaluation record. PACE can accommodate any router. Its theoretical guarantees apply when the router satisfies Assumption 1 and the remaining estimation and selection regularity conditions hold. We describe two representative routing designs below:

Score-quantile routing. One simple instantiation is score-quantile routing. Suppose each instance receives a scalar routing score $Z(X)$, with larger values indicating greater evaluation difficulty. Given cumulative group proportions $0 = u_0 < u_1 < \cdots < u_M = 1$, let $\widehat{q}_g$ be the empirical u_g-quantile of $\{Z(X_i)\}_{i=1}^m$, with $\widehat{q}_0 = -\infty$ and $\widehat{q}_M = \infty$. The induced router is

$$\widehat{\rho}(X_i) = g \quad \Longleftrightarrow \quad \widehat{q}_{g-1} < Z(X_i) \le \widehat{q}_g.$$

The empirical cut points are applied to both calibration and test instances. This rule assigns progressively higher-scored instances to stronger Judge tiers while approximately enforcing the desired routing shares $u_g - u_{g-1}$.

Multiclass routing. A lightweight multiclass classifier can directly assign instances to Judge tiers. Let θ_0 denote its parameters, trained on external data independent of the calibration and test samples, and let $s_g(x; \theta_0)$ be its score for assigning instance x to Judge J_g. The routing rule is

$$\rho(x) := \arg\max_{g \in [M]} s_g(x; \theta_0).$$

The classifier is frozen and applied identically to calibration and test instances. Unlike score-quantile routing, this rule assigns Judge tiers directly without imposing prespecified group proportions.

4.2 Group-Conditional Local-FDR Calibration

Raw confidence values are neither necessarily calibrated nor directly comparable across Judge tiers: the same numerical confidence may correspond to different disagreement probabilities. We therefore place all routed verdicts on a common risk scale by defining local FDR (Cai & Sun, 2009)

$$\ell_g(c) = \Pr(D = 1 \mid C = c, G = g) = \mathbb{E}[D \mid C = c, G = g].$$

Intuitively, $\ell_g(c)$ is the posterior probability that an observed verdict disagrees with the human preference after conditioning on its confidence and routed group. This quantity is unavailable for an unlabeled test instance. For each calibration sample, PACE applies the same router and invokes only its assigned Judge, yielding G_i, C_i, and $D_i = \mathbf{1}\{Y_i \neq \widehat{Y}_i\}$. We estimate ℓ_g using Nadaraya–Watson kernel regression with group-specific bandwidth $h_g > 0$ and kernel K (Wand & Jones, 1994):

$$\widehat{\ell}_g(c) = \frac{\sum_{i=1}^n \mathbf{1}\{G_i = g\} K((c - C_i)/h_g) D_i}{\sum_{i=1}^n \mathbf{1}\{G_i = g\} K((c - C_i)/h_g)}. \tag{2}$$

For each test instance, its estimated disagreement risk is $\widehat{L}_i := \widehat{\ell}_{G_i}(C_i)$. Unlike raw Judge confidence, this quantity has the same probabilistic interpretation across groups.

4.3 Pooled Risk-Controlled Selection

Because $\widehat{L}_i$ estimates verdict i's disagreement probability, its average over a reported set estimates that set's disagreement proportion. Among all size-k sets, the k smallest estimated risks minimize this average. Therefore, to maximize estimated coverage under the risk budget, we order $\widehat{L}_{(1)} \le \cdots \le \widehat{L}_{(m)}$ and select the largest feasible cardinality (Cai & Sun, 2009):

$$\widehat{k} = \max\left(\left\{k \in [m] : \frac{1}{k}\sum_{j=1}^{k} \widehat{L}_{(j)} \le \alpha\right\} \cup \{0\}\right), \tag{3}$$

Because the constraint is imposed after pooling all routing groups, low-risk verdicts in one group can offset higher-risk verdicts in another. PACE therefore uses one shared mFDR budget rather than requiring every group to satisfy a separate copy of the target.

Algorithm 1 in Appendix B gives the complete score–quantile implementation analyzed in Section 5 and used in the main experiments.

5 Theoretical Analysis

This section studies a broad instantiation of the general framework in Section 4.1 with data-driven routers. Unlike the prespecified grouping structures considered in Cai & Sun (2009), our routing groups are learned from the unlabeled test covariates. This creates additional dependence between the Judge assignments and the pooled selection, so arguments developed for fixed groups do not apply directly.

Section 5.1 collects the assumptions used to address this dependence. Then we establish consistency of group-conditional local-FDR estimation in Section 5.2, derive asymptotic mFDR control in Section 5.3, and verify the routing-stability assumption for two representative routers in Section 5.4.

5.1 Assumptions and Regularity Conditions

We first state the routing, estimation, and selection conditions used in the analysis. The first condition controls the dependence introduced when a router is learned from the test covariates.

Assumption 1 (Leave-one-out stability of routing)**.** *Let $\widehat{\rho}$ be the router fitted to $\{X_j\}_{j=1}^m$, and let $\widehat{\rho}^{(-i)}$ be its leave-one-out counterpart obtained using the same algorithm. There exist a deterministic population router $\rho^\star : \mathcal{X} \to [M]$, deterministic sequences $\zeta_m, \gamma_m \downarrow 0$ and $C < \infty$ such that,*

$$\frac{1}{m}\sum_{i=1}^{m} \Pr\left\{\widehat{\rho}(X_i) \neq \widehat{\rho}^{(-i)}(X_i)\right\} \leq C\zeta_m,$$

and, for a fresh $X \sim P_X$ independent of the routing data,

$$\frac{1}{m}\sum_{i=1}^{m} \Pr\left\{\widehat{\rho}^{(-i)}(X) \neq \rho^\star(X)\right\} \leq C\gamma_m, \quad \Pr\left\{\widehat{\rho}(X) \neq \rho^\star(X)\right\} \leq C\gamma_m,$$

Here ζ_m controls the self-influence of each test observation. In contrast, γ_m controls how closely the full-sample and leave-one-out routers approach a common population router $\rho^\star$, and therefore measures the population-level instability. Assumption 1 is a routing-specific form of standard algorithmic stability (Bousquet & Elisseeff, 2002; Zrnic & Jordan, 2023) and is satisfied by many practical procedures under mild conditions. In particular, a fixed or independently trained router satisfies the condition trivially.

The second condition imposes kernel-regression regularity only under the deterministic population router $\rho^\star$, rather than uniformly over all possible fitted routers.

Assumption 2 (Kernel regularity)**.** *Let $G^\star := \rho^\star(X)$ and define the population local-FDR function as $\ell_g^\star(c) := \Pr\{D^{(g)} = 1 \mid C^{(g)} = c, G^\star = g\}$. For every $g \in [M]$, the population routing probability is bounded away from zero; the conditional density of $C^{(g)} \mid \rho^\star(X) = g$ is bounded and bounded away from zero on a common compact support and is uniformly Lipschitz; $\ell_g^\star$ has a uniformly bounded Lipschitz derivative; and K is a bounded symmetric probability kernel with finite second moment.*

The final condition ensures that the population reporting cutoff is well behaved and that the reported fraction does not vanish.

Assumption 3 (Regular population reporting cutoff)**.** *Define the population selection criterion $\Psi_\alpha(t) := \mathbb{E}[(L^\star - \alpha)\mathbf{1}\{L^\star \leq t\}]$ for $t \in [0,1]$. The population cutoff*

$$t_\alpha^\star := \sup\{t \in [0,1] : \Psi_\alpha(t) \leq 0\}$$

belongs to $(\alpha, 1)$, is uniquely identified, and satisfies $\Psi_\alpha(t_\alpha^\star) = 0$. Moreover, there exist constants $u_0, c_\Psi, C_L, p_0 > 0$ such that $p_\alpha^\star := \Pr(L^\star \leq t_\alpha^\star) \geq p_0$ and, for every $0 < u \leq u_0$,

$$\Pr\{|L^\star - t_\alpha^\star| \leq u\} \leq C_L u, \qquad \min\left\{-\Psi_\alpha(t_\alpha^\star - u), \Psi_\alpha(t_\alpha^\star + u)\right\} \geq c_\Psi u.$$

Assumption 3 is standard in the multiple-testing literature (Genovese & Wasserman, 2002). It holds in the regular case where $L^\star$ has a density that is continuous in a neighborhood of $t_\alpha^\star > \alpha$.

5.2 Consistency of Local-FDR Estimation

One central question is whether local FDR remains estimable when a data-driven router induces random groups. Standard kernel-regression theory would generally require density and smoothness conditions uniformly over all possible fitted routers (Wand & Jones, 1994). Assumptions 1 and 2 let us avoid this requirement by anchoring the analysis to the deterministic population router $\rho^\star$.

For a fresh observation X, let $\widehat{L}_\rho$ denote the local-FDR value estimated by groupwise Nadaraya–Watson regression, evaluated at its routed Judge confidence under router ρ, and let $L^\star$ denote the corresponding population local-FDR value under $\rho^\star$. The following proposition links the consistency of local-FDR to the routing stability in Assumption 1.

Proposition 1 (Consistency of local-FDR estimation under data-driven router)**.** *Suppose Assumption 2 holds under the population router $\rho^\star$. For any data-dependent routing rule ρ,*

$$\mathbb{E}\left|\widehat{L}_\rho - L^\star\right| \le C\left[\max_{g\in[M]}\left\{h_g^2 + (nh_g)^{-1/2}\right\} + \mathbb{P}\{\rho(X)\neq\rho^\star(X)\}\right].$$

The proof, provided in Appendix C.3, uses an oracle-anchored perturbation argument that separates the stochastic kernel-estimation error from the effect of data-driven routing. This allows the latter to be controlled directly by $\mathbb{P}\{\rho(X)\neq\rho^\star(X)\}$, while imposing kernel regularity only on the population router $\rho^\star$, rather than uniformly over all possible learned routers.

5.3 mFDR Control

We now turn to the mFDR guarantee. Since the final reported set is obtained by thresholding the estimated local FDRs, local-FDR consistency alone does not guarantee stability of the reported set. Assumption 3 supplies the additional regularity needed to transfer local-FDR consistency to the data-dependent pooled cutoff.

Theorem 1 (Asymptotic mFDR control)**.** *Suppose Assumptions 1–3 hold and $h_g \asymp n^{-1/5}$ for all $g \in [M]$. Then, for a constant C,*

$$\mathrm{mFDR} := \frac{\mathbb{E}[V]}{\mathbb{E}[R]} \le \alpha + \frac{C}{p_0}\left(m^{-1/2} + n^{-1/5} + \gamma_m^{1/2} n^{1/10} + \zeta_m^{1/2}\right).$$

Consequently, when $\gamma_m = o(n^{-1/5})$ and $\zeta_m = o(1)$, $\limsup_{n,m\to\infty}\mathrm{mFDR} \le \alpha$.

The proof is provided in Appendix C.4. The bound separates sampling and estimation error from the additional terms due to data-dependent routing. Its proof combines local-FDR estimation with a leave-one-out analysis of the fitted risks and pooled selection.

5.4 Examples of Feasible Routers

The stability requirement in Assumption 1 is satisfied by several practically relevant routing procedures. We give two representative examples below.

Score-quantile routing. Consider the rule in Section 4.1, with a fixed score function $Z(\cdot)$ and population cutoffs $q_g^\star := F_Z^{-1}(u_g)$, $g = 1, \dots, M-1$, where F_Z is the score distribution function. The population router replaces the empirical cutoffs $\widehat{q}_g$ with $q_g^\star$.

Proposition 2 (Stability of score-quantile routing)**.** *Suppose M is fixed and $Z(X)$ has a continuous density bounded above and away from zero near each $q_g^\star$. Then Assumption 1 holds with*

$$\gamma_m = O(m^{-1/2}), \qquad \zeta_m = O(m^{-1}).$$

The proof is provided in Appendix C.5.2. The two rates follow from empirical-quantile estimation and the rank structure of the cutoffs: only observations adjacent to a routing boundary can change their own assignments after deletion.

Multiclass routing. The externally trained multiclass router can be further adapted using only the unlabeled test inputs:

$$\widehat{\theta} := \arg\min_{\theta} \left\{ \frac{1}{m} \sum_{i=1}^{m} \phi(\theta; X_i) + \frac{\lambda}{2} \|\theta - \theta_0\|_2^2 \right\}, \quad \widehat{\rho}(x) := \arg\max_{g \in [M]} s_g(x; \widehat{\theta}).$$

Here ϕ is a label-free adaptation loss. Let $\theta^\star$ minimize the population objective, define $\rho^\star(x) := \arg\max_g s_g(x; \theta^\star)$, and denote the gap between the two largest gate scores by $\Delta_\theta(x)$.

Proposition 3 (Stability of test-adapted multiclass routing)**.** *Suppose the regularized objectives are uniformly strongly convex, $\nabla_\theta \phi(\theta; x)$ is uniformly bounded, and $s_g(x; \theta)$ is uniformly Lipschitz in θ. Assume $\mathbb{E}\|\widehat{\theta} - \theta^\star\| = O(m^{-1/2})$ and $\mathbb{P}\{\Delta_\theta(X) \le u\} \le Cu$ uniformly over the relevant parameter region and all small $u > 0$. Then Assumption 1 holds with $\gamma_m = O(m^{-1/2})$ and $\zeta_m = O(m^{-1})$.*

The proof is provided in Appendix C.5.3. Without test adaptation, $\gamma_m = \zeta_m = 0$ conditional on the external training data. For both adaptive examples, $n \asymp m$ suffices for the routing-rate requirements of Theorem 1.

6 Experiment

6.1 Experimental Setup

We use 5,172 ChatArena samples (Zheng et al., 2023), 5,170 TL;DR samples (Stiennon et al., 2020), and 1,424 MT-Bench samples (Zheng et al., 2023); each contains an instruction, two candidate responses, and a human preference label. Ordered from lower to higher capability and cost, the Judge panel is Mistral-7B-Instruct, Qwen2.5-14B-Instruct, and GPT-5.6-Luna. For the label-free routing experiments, the frozen score model is Qwen2.5-3B-Instruct. Unless otherwise specified, every experiment uses 1,000 independently redrawn random splits, exactly 500 labeled calibration samples per split, and all remaining samples for testing. Calibration labels estimate Judge error risks but are not used to fit the router. Appendix E.1 specifies the experiment settings.

Baselines. **Non-selective methods** are (1) **No selection (Luna)**, which accepts every verdict; and (2) **PoLL**, which invokes all three Judges, averages their class probabilities, and accepts every pooled prediction (Verga et al., 2024). **Selective methods** are (1) **SCOPE-BPE (Model)**, which applies SCOPE's feasible entropy threshold to each named Judge (Badshah et al., 2026); (2) **Three-Judge cascade**, which queries Judges from weakest to strongest and escalates or abstains when the calibrated threshold is not met (Jung et al., 2025); and (3) **Cascaded heuristic selection**, which follows the same path using raw-confidence thresholds (Jung et al., 2025). Appendix E.2 provides more details.

Metrics. We report (1) **Agreement** ($1 - \widehat{\text{mFDR}}$), pooled across splits as the number of correct accepted predictions divided by the total accepted; (2) **Coverage**, the accepted fraction of the test set; (3) **Judge records** per test sample; (4) **Cost** in USD per 1,000 test samples. Lightweight router inference is not counted as a Judge record, but its cost is included in the reported Cost. Appendix E.3 gives the full accounting protocol.

6.2 Main Results

We ask whether PACE improves the agreement–coverage–cost trade-off relative to external baselines on general dialogue evaluation (ChatArena) and summary-preference evaluation (TL;DR), using 85% as the target agreement. We evaluate two label-free quantile routers (Prompt-v1 and Prompt-v2; Appendix E.6) and a pretrained multiclass router (PACE-BGE; Appendix D.2).

Across both datasets, the non-selective baselines and heuristic selection fall short of the agreement target. The fixed-Judge SCOPE variants either miss the target or retain only a small subset, while the cascades require repeated Judge records and substantially higher cost. The label-free PACE variants attain the target with one Judge record per instance. On ChatArena they retain more verdicts than the Three-Judge cascade; on TL;DR they trade lower coverage for lower cost. At the more stringent 90% target on MT-Bench, PACE-BGE attains the target agreement with competitive coverage and lower cost than the Three-Judge cascade (Appendix D.1). Thus PACE supports both label-free and pretrained routing designs, with an operating-point-dependent agreement–coverage–cost trade-off.

Table 1: Main results on ChatArena with target agreement $1 - \alpha = 0.85$.

Method	Agreement	Coverage	Cost	Judge records per test sample			
				Mistral	Qwen	Luna	Total
No selection (Luna)	79.010	100.000	8.884	0.000	0.000	1.000	1.000
PoLL-style probability pooling	79.042	100.000	15.036	1.000	1.000	1.000	3.000
SCOPE-BPE (Mistral)	80.855	11.224	0.384	1.000	0.000	0.000	1.000
SCOPE-BPE (Qwen14B)	84.735	57.440	5.768	0.000	1.000	0.000	1.000
SCOPE-BPE (Luna)	84.678	65.773	8.884	0.000	0.000	1.000	1.000
Three-Judge cascade	86.603	54.993	10.992	1.000	0.876	0.625	2.501
Cascaded heuristic selection	79.030	93.018	4.740	1.000	0.511	0.159	1.669
PACE (ours)							
Prompt-v1 (quantile)	**85.071**	**59.715**	**5.019**	**0.333**	**0.333**	**0.333**	**1.000**
Prompt-v2 (quantile)	**85.233**	**59.764**	**5.020**	**0.333**	**0.333**	**0.333**	**1.000**
PACE-BGE (multiclass)	**84.995**	**59.473**	**5.950**	**0.254**	**0.252**	**0.494**	**1.000**

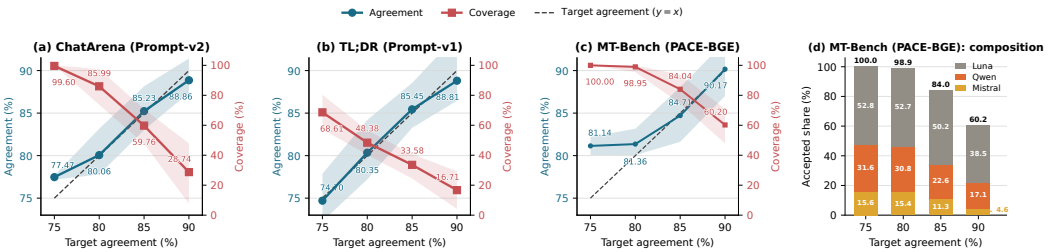

Figure 2: PACE across target agreement levels. Shaded bands contain the central 80% of repeated-split outcomes.

6.3 Target Attainment Across Operating Points

To test whether PACE generalizes across target levels, we sweep target agreement from 75% to 90% for representative routers on three datasets.

Figure 2(a)–(c) shows the intended agreement–coverage response: increasing the target generally raises achieved agreement while reducing the coverage, and achieved agreement broadly follows or exceeds the target line. This expected trade-off confirms that PACE remains effective across target levels rather than only at one chosen operating point. Panel (d) further shows that stricter selection shifts the accepted set toward the strongest Judge Luna, consistent with retaining verdicts with lower estimated disagreement risk.

6.4 Routing-design Analysis

As a common method, quantile grouping has two components: (1) *Routing strategy* that scores inputs; (2) *Group allocation* that determines the proportions assigned to the Judge tiers. We ablate the two components separately under a common ChatArena target agreement 85%.

1. **Routing strategy.** We fix equal-tertile group allocation and compare six grouping strategies: Random, Prompt-v1/v2/v3, Length, and BGE-M3 kNN density/OOD (Chen et al., 2024). Appendix E.1 gives their details.
2. **Group allocation.** We hold the Prompt-v2 grouping strategy fixed and compare three prespecified group allocations: $\mathbf{q_1} = (0.20, 0.80)$, equal tertiles $\mathbf{q_2} = (1/3, 2/3)$, and $\mathbf{q_3} = (0.40, 0.60)$.

Figure 3(a)–(b) shows complementary routing effects. Grouping strategy primarily changes coverage: Prompt-v1 and Prompt-v2 achieve the highest agreement while retaining the most verdicts. Allocation exposes a deployment trade-off: 20/60/20 maximizes coverage, 40/20/40 yields the highest agreement and lowest cost, and equal tertiles provide an intermediate balance. Thus, both design choices provide practical control over PACE's operating point.

Table 2: Main results on TL;DR with target agreement $1 - \alpha = 0.85$.

Method	Agreement	Coverage	Cost	Judge records per test sample			
				Mistral	Qwen	Luna	Total
No selection (Luna)	73.725	100.000	8.884	0.000	0.000	1.000	1.000
PoLL-style probability pooling	72.934	100.000	15.036	1.000	1.000	1.000	3.000
SCOPE-BPE (Mistral)	82.536	10.984	0.384	1.000	0.000	0.000	1.000
SCOPE-BPE (Qwen14B)	84.091	25.718	5.768	0.000	1.000	0.000	1.000
SCOPE-BPE (Luna)	83.003	27.149	8.884	0.000	0.000	1.000	1.000
Three-Judge cascade	86.646	42.029	13.193	1.000	0.937	0.834	2.770
Cascaded heuristic selection	77.983	78.218	9.653	1.000	0.864	0.482	2.346
PACE (ours)							
Prompt-v1 (quantile)	**85.446**	**33.578**	**5.020**	**0.333**	**0.333**	**0.333**	**1.000**
Prompt-v2 (quantile)	**85.210**	**31.330**	**5.021**	**0.333**	**0.333**	**0.333**	**1.000**
PACE-BGE (multiclass)	**84.923**	**43.107**	**7.579**	**0.075**	**0.218**	**0.707**	**1.000**

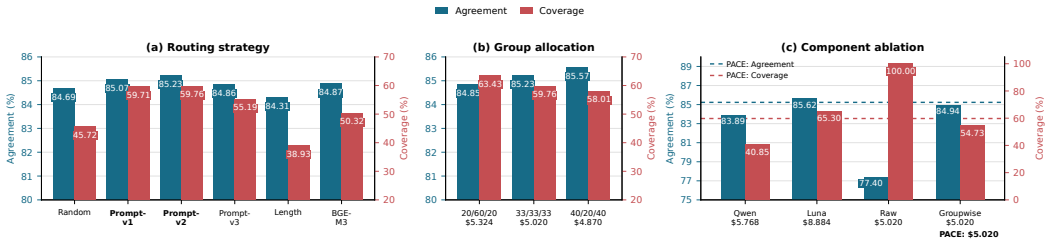

Figure 3: Routing-design (a)–(b) and framework-component (c) ablations on 85% target ChatArena.

6.5 Framework-Component Ablation

PACE has three stages: routing, local-FDR calibration, and pooled risk budgeting. To isolate their roles at the ChatArena 85% target, each ablation removes one stage: fixed Qwen/Luna removes routing, raw-confidence pooling replaces local FDR with $1 - C_i$, and groupwise budgets enforce risk separately by group, preventing unused risk allowance from being shared across groups.

Figure 3(c) shows that fixed Qwen misses the target with lower coverage, while fixed Luna offers modest gains at higher cost. Raw-confidence pooling accepts all verdicts but misses the target, exposing overconfidence and poor cross-Judge comparability. Groupwise budgets reduce coverage by preventing risk reallocation. Each stage contributes to the agreement–coverage–cost trade-off.

Additional experiments. Appendix D reports stricter-target comparisons, learned gates, Judge-panel replacement, calibration-size sensitivity, routing–selection hybrids, cross-dataset evaluation, and allocation sweeps. These studies further support routing portability and clarify the cost–coverage trade-offs, while revealing how calibration size affects finite-sample agreement at stricter targets.

7 Limitations

PACE forms routing quantiles from an offline test batch; its present analysis does not directly cover streaming distribution shifts. Jointly optimizing routing for agreement, coverage, and cost across heterogeneous Judge tiers remains an open direction for future work on data-driven Judge assignment.

8 Conclusion

We introduced PACE, a selective-evaluation framework that combines pre-routing with pooled risk control using one Judge record over heterogeneous panels. We established group-conditional local-FDR consistency under stable data-dependent routing and asymptotic mFDR control for pooled selection. Across three datasets, PACE yields favorable or competitive agreement–coverage operating points at lower cost than most methods, with one Judge evaluation per routed instance.

References

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International conference on learning representations*, volume 2024, pp. 55198–55218, 2024.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

Sher Badshah, Ali Emami, and Hassan Sajjad. Scope: Selective conformal optimized pairwise llm judging. *arXiv preprint arXiv:2602.13110*, 2026.

Yajie Bao, Yuyang Huo, Haojie Ren, and Changliang Zou. Selective conformal inference with false coverage-statement rate control. *Biometrika*, 111(3):727–742, 2024.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

T Tony Cai and Wenguang Sun. Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *Journal of the American Statistical Association*, 104(488):1467–1481, 2009.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the association for computational linguistics: ACL 2024*, pp. 2318–2335, 2024.

Christopher Genovese and Larry Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(3):499–517, 2002.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *The Innovation*, 7(6), 2026.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: Llm judges with provable guarantees for human agreement. In *International Conference on Learning Representations*, volume 2025, pp. 3101–3125, 2025.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353, 2024.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for nlg. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 2241–2252, 2017.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Felipe Maia Polo, Xinhe Wang, Mikhail Yurochkin, Gongjun Xu, Moulinath Banerjee, and Yuekai Sun. Bridging human and llm judgments: Understanding and narrowing the gap. *arXiv preprint arXiv:2508.12792*, 2025.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Wenguang Sun and T Tony Cai. Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912, 2007.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Chris Van Der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th international conference on natural language generation*, pp. 355–368, 2019.

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

Matt P Wand and M Chris Jones. *Kernel smoothing*. CRC press, 1994.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9440–9450, 2024.

Yidong Wang, Yunze Song, Tingyuan Zhu, Xuanwang Zhang, Zhuohao Yu, Hao Chen, Chiyu Song, Qiufeng Wang, Zhen Wu, Xin-Yu Dai, et al. Trustjudge: Inconsistencies of llm-as-a-judge and how to alleviate them. In *International Conference on Learning Representations*, volume 2026, pp. 146399–146421, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. In *International Conference on Learning Representations*, volume 2025, pp. 51257–51296, 2025.

Tijana Zrnic and Michael I Jordan. Post-selection inference via algorithmic stability. *The Annals of Statistics*, 51(4):1666–1691, 2023.

A The usage of AI

In this work, we only use AI for language polishing to improve the clarity and fluency of the articles; it was not used to generate research ideas, problem formulations, framework design, theories, algorithms, data processing, results, or evaluations. All AI-generated content has been carefully reviewed and verified by the authors to ensure its accuracy, rigor, and consistency with the intended meaning.

B PACE Procedure

Algorithm 1 gives the score–quantile instantiation of PACE used in the main experiments. The framework itself permits any frozen label-free router satisfying the conditions stated in the main text.

Algorithm 1 PACE: pre-routing with pooled local-FDR selection

Require: Labeled calibration set $\mathcal{D}_{\text{cal}}$; unlabeled test batch $\mathcal{D}_{\text{test}}$; Judge tiers $\{J_g\}_{g=1}^{M}$; target α
Ensure: Accepted verdict set $\mathcal{R}$

```
Construct routing rule ρ̂ based on D_test
for all i ∈ D_cal ∪ D_test do
    Apply routing rule ρ̂ to obtain G_i
    Invoke only J_{G_i} to obtain (Ŷ_i, C_i)
end for
Compute D_i = 1{Y_i ≠ Ŷ_i} for i ∈ D_cal
Fit ℓ̂_g from the calibration samples in each group using Equation (2)
Compute ℓ̂_{G_i}(C_i) for every i ∈ D_test and sort them increasingly
Choose k̂ by Equation (3), with k̂ = 0 if no positive k is feasible
return the verdicts of the first k̂ test samples in the sorted order as R; return ∅ if k̂ = 0
```

C Detailed Proofs for the Theoretical Analysis

C.1 Notation

Throughout this section, X without a subscript denotes a fresh evaluation observation, independent of both the test sample and the calibration sample. For each Judge $g \in [M]$, let $C^{(g)}$ denote its potential confidence score and define the corresponding disagreement indicator

$$D^{(g)} := \mathbf{1}\left\{Y \neq \widehat{Y}^{(g)}\right\}.$$

Let $\mathcal{F}$ denote all test-sample information and algorithmic randomness used to construct a data-dependent routing rule ρ. And $\mathcal{F}^{-i}$ denotes the same information but excluding X_i. We assume that ρ is $\mathcal{F}$-measurable and that both $\left(X, \{C^{(g)}, D^{(g)}\}_{g=1}^{M}\right)$ and the calibration sample $\left\{\left(X_k^{\text{cal}}, \{C_k^{(g)}, D_k^{(g)}\}_{g=1}^{M}\right)\right\}_{k=1}^{n}$ are independent of $\mathcal{F}$.

Router-specific population quantities. For any $\mathcal{F}$-measurable router ρ, define

$$G_\rho := \rho(X), \qquad C_\rho := C^{(G_\rho)}, \qquad D_\rho := D^{(G_\rho)}.$$

For $g \in [M]$, define the corresponding conditional disagreement risk

$$\ell_{\rho,g}(c) := \mathbb{P}\left(D^{(g)} = 1 \,\middle|\, C^{(g)} = c,\, \rho(X) = g,\, \mathcal{F}\right),$$

and let $L_\rho := \ell_{\rho,G_\rho}(C_\rho)$. Whenever $\rho = \widehat{\rho}^{(-i)}$, the corresponding router-specific quantities are understood conditionally on $\mathcal{F}^{-i}$ rather than the full $\mathcal{F}$.

For the deterministic population router $\rho^\star$, we use the shorthand

$$G^\star := \rho^\star(X), \qquad C^\star := C^{(G^\star)}, \qquad D^\star := D^{(G^\star)},$$

and

$$\ell_g^\star(c) := \mathbb{P}\left(D^{(g)} = 1 \,\middle|\, C^{(g)} = c,\, \rho^\star(X) = g\right), \qquad L^\star := \ell_{G^\star}^\star(C^\star).$$

The conditional population disagreement between a data-dependent router ρ and $\rho^\star$ is

$$d_P(\rho, \rho^\star) := \mathbb{P}\left\{\rho(X) \neq \rho^\star(X) \,\middle|\, \mathcal{F}\right\}. \tag{4}$$

Because X is independent of $\mathcal{F}$, equivalently,

$$d_P(\rho, \rho^\star) = \int \mathbf{1}\left\{\rho(x) \neq \rho^\star(x)\right\} \, \mathrm{d}P_X(x),$$

where the right-hand side is evaluated at the realized router ρ. Thus $d_P(\rho, \rho^\star)$ is $\mathcal{F}$-measurable.

Full-sample, leave-one-out, and population quantities. Let $\widehat{\rho}$ denote the router constructed from the full test sample, and let $\widehat{\rho}^{(-i)}$ denote the corresponding leave-one-out router obtained after removing X_i. For each $i \in [m]$, write

$$\widehat{G}_i := \widehat{\rho}(X_i), \qquad \widetilde{G}_i := \widehat{\rho}^{(-i)}(X_i), \qquad G_i^\star := \rho^\star(X_i),$$

and define the corresponding routed confidence scores

$$\widehat{C}_i := C_i^{(\widehat{G}_i)}, \qquad \widetilde{C}_i := C_i^{(\widetilde{G}_i)}, \qquad C_i^\star := C_i^{(G_i^\star)}.$$

Throughout the proofs, hats, tildes, and stars therefore indicate full-sample, leave-one-out, and population-router quantities, respectively.

Kernel local-FDR estimators. For any $\mathcal{F}$-measurable router ρ, apply ρ to the calibration observations and define

$$\widehat{q}_{\rho,g}(c) := \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}\left\{\rho(X_k^{\mathrm{cal}}) = g\right\} K_{h_g}(c - C_k^{(g)})$$

and

$$\widehat{p}_{\rho,g}(c) := \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}\left\{\rho(X_k^{\mathrm{cal}}) = g\right\} K_{h_g}(c - C_k^{(g)}) D_k^{(g)}.$$

The corresponding Nadaraya–Watson estimator is

$$\widehat{\ell}_{\rho,g}(c) := \begin{cases} \widehat{p}_{\rho,g}(c)/\widehat{q}_{\rho,g}(c), & \widehat{q}_{\rho,g}(c) > 0, \\ 1, & \widehat{q}_{\rho,g}(c) = 0. \end{cases}$$

For a fresh observation, define $\widehat{L}_\rho := \widehat{\ell}_{\rho,G_\rho}(C_\rho)$.

For later use, let

$$q_{\rho,g,h}(c) := \mathbb{E}_{\mathrm{cal}}\left[\widehat{q}_{\rho,g}(c) \,\middle|\, \mathcal{F}\right], \qquad p_{\rho,g,h}(c) := \mathbb{E}_{\mathrm{cal}}\left[\widehat{p}_{\rho,g}(c) \,\middle|\, \mathcal{F}\right].$$

For the test observations, define

$$\widehat{L}_i := \widehat{\ell}_{\widehat{\rho},\widehat{G}_i}(\widehat{C}_i), \qquad \widetilde{L}_i := \widehat{\ell}_{\widehat{\rho}^{(-i)},\widetilde{G}_i}(\widetilde{C}_i), \qquad L_i^\star := \ell_{G_i^\star}^\star(C_i^\star).$$

Thus, $\widehat{L}_i$ and $\widetilde{L}_i$ are the estimated local-FDR values under the full-sample and leave-one-out routers, respectively, whereas $L_i^\star$ is the population local-FDR under $\rho^\star$.

Pooled selection quantities. Define the empirical selection criterion

$$\widehat{\Psi}_{\alpha,m}(t) := \frac{1}{m} \sum_{i=1}^{m} (\widehat{L}_i - \alpha) \mathbf{1}\left\{\widehat{L}_i \leq t\right\}, \qquad t \in [0, 1],$$

and the corresponding empirical cutoff

$$\widehat{t}_\alpha := \sup\left\{t \in [0, 1] : \widehat{\Psi}_{\alpha,m}(t) \leq 0\right\}.$$

Let $\widehat{k}$ be defined by the ordered-prefix rule in (3), with ties broken by a fixed deterministic order. Define the reporting indicators by $R_i := \mathbf{1}\{\mathrm{rank}(\widehat{L}_i) \le \widehat{k}\}$. The quantity $\widehat{t}_\alpha$ is used only as an auxiliary boundary for analyzing this ordered-prefix rule. In particular,

$$\mathbf{1}\{\widehat{L}_i < \widehat{t}_\alpha\} \le R_i \le \mathbf{1}\{\widehat{L}_i \le \widehat{t}_\alpha\}.$$

For notational simplicity, under the canonical bandwidth $h_g \asymp n^{-1/5}$, define $r_{n,m} := n^{-2/5} + \zeta_m + \gamma_m n^{1/5}$.

C.2 Lemmas

The following lemmas establish the main ingredients for the mFDR proof. Lemmas 1–6 control kernel estimation and perturbations induced by changing the router, while Lemmas 7–9 propagate these bounds to the leave-one-out local-FDR values and the final reporting decisions.

Lemma 1 (Ratio perturbation with a possibly small denominator)**.** *Let (A_0, B_0) and (A_1, B_1) be two pairs of nonnegative quantities satisfying*

$$0 \le A_j \le B_j, \qquad j \in \{0, 1\}.$$

Define

$$T_j := \begin{cases} A_j/B_j, & B_j > 0, \\ t_j, & B_j = 0, \end{cases}$$

where t_j can be any value in $[0, 1]$. Then

$$B_0|T_1 - T_0| \le 2|A_1 - A_0| + 2|B_1 - B_0|.$$

Proof. Consider first the good-denominator event $\mathcal{E} := \left\{B_1 \ge \frac{B_0}{2}\right\}$.

On $\mathcal{E}$,

$$|T_1 - T_0| = \left|\frac{A_1}{B_1} - \frac{A_0}{B_0}\right| \le \frac{|A_1 - A_0|}{B_1} + \frac{A_0|B_1 - B_0|}{B_0 B_1} \le \frac{2}{B_0}\left\{|A_1 - A_0| + |B_1 - B_0|\right\},$$

where we used $A_0/B_0 \le 1$.

On the complementary bad-denominator event $\mathcal{E}^c$, we have $|B_1 - B_0| > \frac{B_0}{2}$.

Since $T_0, T_1 \in [0, 1]$,

$$B_0|T_1 - T_0| \le B_0 \le 2|B_1 - B_0|.$$

□

Lemma 2 (Uniform comparability of the oracle smoothed denominator)**.** *Suppose Assumption 2 holds, and let $C = [a, b]$. Define*

$$q^\star_g(c) := \pi^\star_g f^\star_g(c), \qquad q^\star_{g,h}(c) := \int_C K_{h_g}(c - u) q^\star_g(u)\,\mathrm{d}u.$$

Then there exist constants $h_0 > 0$ and $\kappa_0 < \infty$, depending only on K, C, $f_{\min}$, and $f_{\max}$, such that, whenever $\max_{g\in[M]} h_g \le h_0$,

$$q^\star_g(c) \le \kappa_0 q^\star_{g,h}(c)$$

uniformly over $g \in [M]$ and $c \in C$. In particular, $q^\star_{g,h}(c) > 0$ uniformly over $g \in [M]$ and $c \in C$.

Proof. Because K is nonnegative, symmetric, and integrates to one,

$$\int_0^\infty K(v)\,\mathrm{d}v = \int_{-\infty}^0 K(v)\,\mathrm{d}v = \frac{1}{2}.$$

Hence there exists $R > 0$ such that

$$\int_0^R K(v)\,\mathrm{d}v = \int_{-R}^0 K(v)\,\mathrm{d}v \ge \frac{1}{4}.$$

Set $h_0 := \frac{b-a}{2R}$.

For any $c \in [a, b]$, at least one of

$$c - a \geq \frac{b-a}{2} \qquad \text{or} \qquad b - c \geq \frac{b-a}{2}$$

must hold. Moreover,

$$\int_a^b K_h(c-u)\,\mathrm{d}u = \int_{(c-b)/h}^{(c-a)/h} K(v)\,\mathrm{d}v.$$

If $c - a \geq (b-a)/2$ and $h \leq h_0$, then $\frac{c-a}{h} \geq R$, so the integration interval contains $[0, R]$. Therefore,

$$\int_a^b K_h(c-u)\,\mathrm{d}u \geq \int_0^R K(v)\,\mathrm{d}v \geq \frac{1}{4}.$$

Similarly, if $b - c \geq (b-a)/2$, then $\frac{c-b}{h} \leq -R$, and the integration interval contains $[-R, 0]$, yielding the same lower bound. Consequently,

$$\inf_{0<h\leq h_0} \inf_{c\in C} \int_C K_h(c-u)\,\mathrm{d}u \geq \frac{1}{4}.$$

Using the lower density bound in Assumption 2, for every $g \in [M]$, $c \in C$, and $h_g \leq h_0$,

$$\begin{aligned} q^\star_{g,h}(c) &= \pi^\star_g \int_C K_{h_g}(c-u) f^\star_g(u)\,\mathrm{d}u \\ &\geq \pi^\star_g f_{\min} \int_C K_{h_g}(c-u)\,\mathrm{d}u \\ &\geq \frac{1}{4}\pi^\star_g f_{\min}. \end{aligned}$$

On the other hand,

$$q^\star_g(c) = \pi^\star_g f^\star_g(c) \leq \pi^\star_g f_{\max}.$$

It follows that

$$q^\star_g(c) \leq \frac{4 f_{\max}}{f_{\min}} q^\star_{g,h}(c).$$

Thus the stated result holds with $\kappa_0 := \frac{4 f_{\max}}{f_{\min}}$. □

Lemma 3 converts router disagreement into an L_1 perturbation of the routed kernel measures, which is the key bridge between routing stability and local-FDR estimation.

Lemma 3 (Perturbation of routed measures)**.** *For a deterministic router ρ, define the finite measures*

$$Q_{\rho,g}(A) := \mathbb{P}\left(\rho(X) = g,\, C^{(g)} \in A\right)$$

and

$$P_{\rho,g}(A) := \mathbb{E}\left[D^{(g)} \mathbf{1}\left\{\rho(X) = g,\, C^{(g)} \in A\right\}\right],$$

for every Borel set $A \subseteq C$. Thus, $0 \leq P_{\rho,g} \leq Q_{\rho,g}$.

For two routers ρ_1 and ρ_2,

$$\sum_{g=1}^M \left\|Q_{\rho_1,g} - Q_{\rho_2,g}\right\|_{\mathrm{TV}} \leq 2 d_P(\rho_1, \rho_2), \tag{5}$$

$$\sum_{g=1}^M \left\|P_{\rho_1,g} - P_{\rho_2,g}\right\|_{\mathrm{TV}} \leq 2 d_P(\rho_1, \rho_2). \tag{6}$$

Define their kernel-smoothed versions by

$$q_{\rho,g,h}(c) := \int_C K_{h_g}(c-u)\,\mathrm{d}Q_{\rho,g}(u)$$

and

$$p_{\rho,g,h}(c) := \int_C K_{h_g}(c-u)\,\mathrm{d}P_{\rho,g}(u).$$

Then

$$\sum_{g=1}^{M}\int_{\mathbb{R}}\left|q_{\rho_1,g,h}(c)-q_{\rho_2,g,h}(c)\right|\,\mathrm{d}c \le 2d_P(\rho_1,\rho_2), \tag{7}$$

$$\sum_{g=1}^{M}\int_{\mathbb{R}}\left|p_{\rho_1,g,h}(c)-p_{\rho_2,g,h}(c)\right|\,\mathrm{d}c \le 2d_P(\rho_1,\rho_2). \tag{8}$$

Proof. For every measurable function φ satisfying $\|\varphi\|_\infty \le 1$,

$$\begin{aligned}\sum_{g=1}^{M}\left|\int \varphi(c)\,\mathrm{d}\left(Q_{\rho_1,g}-Q_{\rho_2,g}\right)(c)\right| &\le \mathbb{E}\left[\sum_{g=1}^{M}\left|\mathbf{1}\{\rho_1(X)=g\}-\mathbf{1}\{\rho_2(X)=g\}\right|\right]\\ &= 2\mathbb{P}\{\rho_1(X)\neq\rho_2(X)\}.\end{aligned}$$

This proves (5). Since $0 \le D^{(g)} \le 1$, the same argument proves (6).

For any finite signed measure μ, convolution with the nonnegative normalized kernel satisfies

$$\|K_h * \mu\|_{L_1} \le \|\mu\|_{\mathrm{TV}}\int_{\mathbb{R}} K_h(u)\,\mathrm{d}u = \|\mu\|_{\mathrm{TV}}.$$

Applying this contraction property to (5) and (6) proves (7) and (8). □

Lemma 4 (Integrated stochastic fluctuation of routed kernel estimators)**.** *Let ρ be any $\mathcal{F}$-measurable router. Suppose that $C^{(g)} \in C$ almost surely, $0 \le D^{(g)} \le 1$, and K is bounded and integrable. Then, conditional on $\mathcal{F}$,*

$$\sum_{g=1}^{M}\int_C \mathbb{E}_{\mathrm{cal}}\left[\left|\widehat{q}_{\rho,g}(c)-q_{\rho,g,h}(c)\right|+\left|\widehat{p}_{\rho,g}(c)-p_{\rho,g,h}(c)\right|\middle|\mathcal{F}\right]\,\mathrm{d}c \le C\max_{g\in[M]}(nh_g)^{-1/2}.$$

Proof. Conditional on $\mathcal{F}$, the router ρ is fixed and the calibration observations remain independent and identically distributed. For every $g \in [M]$ and $c \in C$,

$$\begin{aligned}&\mathbb{E}_{\mathrm{cal}}\left[\left|\widehat{q}_{\rho,g}(c)-q_{\rho,g,h}(c)\right|\middle|\mathcal{F}\right]\\ &\quad\le\left\{\mathrm{Var}_{\mathrm{cal}}\left(\widehat{q}_{\rho,g}(c)\,\middle|\,\mathcal{F}\right)\right\}^{1/2}\\ &\quad\le\frac{1}{\sqrt{n}}\left\{\mathbb{E}_{\mathrm{cal}}\left[\mathbf{1}\{\rho(X^{\mathrm{cal}})=g\}K_{h_g}^2(c-C^{(g)})\,\middle|\,\mathcal{F}\right]\right\}^{1/2}.\end{aligned}$$

Consequently, Cauchy–Schwarz and Fubini's theorem yield

$$\begin{aligned}&\int_C \mathbb{E}_{\mathrm{cal}}\left[\left|\widehat{q}_{\rho,g}(c)-q_{\rho,g,h}(c)\right|\middle|\mathcal{F}\right]\,\mathrm{d}c\\ &\le\frac{|C|^{1/2}}{\sqrt{n}}\left\{\int_C \mathbb{E}_{\mathrm{cal}}\left[\mathbf{1}\{\rho(X^{\mathrm{cal}})=g\}K_{h_g}^2(c-C^{(g)})\,\middle|\,\mathcal{F}\right]\,\mathrm{d}c\right\}^{1/2}\\ &\le\frac{|C|^{1/2}}{\sqrt{n}}\left\{\int_{\mathbb{R}} K_{h_g}^2(u)\,\mathrm{d}u\right\}^{1/2}\\ &=|C|^{1/2}\|K\|_2(nh_g)^{-1/2}.\end{aligned}$$

Since K is bounded and integrable, $\|K\|_2 < \infty$.

The numerator is handled identically. Since $0 \le D^{(g)} \le 1$,

$$\begin{aligned}&\mathrm{Var}_{\mathrm{cal}}\left(\widehat{p}_{\rho,g}(c)\,\middle|\,\mathcal{F}\right)\\ &\quad\le\frac{1}{n}\mathbb{E}_{\mathrm{cal}}\left[\mathbf{1}\{\rho(X^{\mathrm{cal}})=g\}K_{h_g}^2(c-C^{(g)})(D^{(g)})^2\,\middle|\,\mathcal{F}\right]\\ &\quad\le\frac{1}{n}\mathbb{E}_{\mathrm{cal}}\left[\mathbf{1}\{\rho(X^{\mathrm{cal}})=g\}K_{h_g}^2(c-C^{(g)})\,\middle|\,\mathcal{F}\right].\end{aligned}$$

Thus,

$$\int_C \mathbb{E}_{\text{cal}}\left[\left|\widehat{p}_{\rho,g}(c) - p_{\rho,g,h}(c)\right| \middle| \mathcal{F}\right] \mathrm{d}c \le C(nh_g)^{-1/2}.$$

Summing over g proves the result. □

Lemma 5 shows that replacing one router by a nearby router perturbs the corresponding population disagreement risk only in proportion to their routing disagreement.

Lemma 5 (Integrated stability of routed disagreement risks)**.** *For any two deterministic routers ρ_1, ρ_2,*

$$\mathbb{E}\left|L_{\rho_1} - L_{\rho_2}\right| \le C d_P(\rho_1, \rho_2),$$

where the routed quantities are coupled through the same underlying observation.

Proof. For $j = 1, 2$, let $Z_j := (G_{\rho_j}, C_{\rho_j})$, and let P_j denote the law of (Z_j, D_{ρ_j}). Under the natural coupling, the routed tuples coincide whenever $\rho_1(X) = \rho_2(X)$. Hence

$$\|P_1 - P_2\|_{\text{TV}} \le d_P(\rho_1, \rho_2).$$

Let $\eta_j(z) := \mathbb{E}_{P_j}[D \mid Z = z]$. A standard Radon–Nikodym argument gives

$$\int |\eta_1(z) - \eta_2(z)| \, \mathrm{d}(P_1)_Z(z) \le C\|P_1 - P_2\|_{\text{TV}}.$$

Moreover, since $\eta_2 \in [0, 1]$,

$$\mathbb{E}|\eta_2(Z_1) - \eta_2(Z_2)| \le \Pr(Z_1 \ne Z_2) \le d_P(\rho_1, \rho_2).$$

Therefore, by the triangle inequality,

$$\begin{aligned}\mathbb{E}|L_{\rho_1} - L_{\rho_2}| &= \mathbb{E}|\eta_1(Z_1) - \eta_2(Z_2)| \\ &\le \mathbb{E}|\eta_1(Z_1) - \eta_2(Z_1)| + \mathbb{E}|\eta_2(Z_1) - \eta_2(Z_2)| \\ &\le C d_P(\rho_1, \rho_2).\end{aligned}$$

□

Lemma 6 (Integrated smoothing bias of the oracle NW regression)**.** *Suppose Assumption 2 holds. Recall that*

$$q^\star_g(c) := \pi^\star_g f^\star_g(c), \qquad p^\star_g(c) := q^\star_g(c)\ell^\star_g(c),$$

and define

$$q^\star_{g,h} := K_{h_g} * q^\star_g, \qquad p^\star_{g,h} := K_{h_g} * p^\star_g, \qquad \ell^\star_{g,h} := \frac{p^\star_{g,h}}{q^\star_{g,h}}.$$

Then, for all sufficiently small bandwidths,

$$\sum_{g=1}^{M} \int_C q^\star_g(c) \left|\ell^\star_{g,h}(c) - \ell^\star_g(c)\right| \mathrm{d}c \le C \max_{g\in[M]} h_g^2.$$

Equivalently,

$$\mathbb{E}\left|\ell^\star_{G^\star,h}(C^\star) - L^\star\right| \le C \max_{g\in[M]} h_g^2.$$

Proof. Write $C = [a, b]$. Since $p^\star_g = q^\star_g \ell^\star_g$, we have

$$\begin{aligned}p^\star_g(u) - \ell^\star_g(c) q^\star_g(u) &= q^\star_g(u)\{\ell^\star_g(u) - \ell^\star_g(c)\} \\ &= q^\star_g(c)(\ell^\star_g)'(c)(u - c) + O(|u - c|^2),\end{aligned}$$

uniformly over $g \in [M]$ and $u, c \in C$. Indeed, the bounded Lipschitz derivative of $\ell^\star_g$ gives $\ell^\star_g(u) - \ell^\star_g(c) = (\ell^\star_g)'(c)(u-c) + O(|u-c|^2)$. The Lipschitz continuity of $q^\star_g$ then allows $q^\star_g(u)$ in the linear

term to be replaced by $q_g^\star(c)$ at an additional cost of $O(|u-c|^2)$. The remainder is uniform because $q_g^\star$ and $(\ell_g^\star)'$ are uniformly bounded. Consequently,

$$\begin{aligned}&\left|p_{g,h}^\star(c) - \ell_g^\star(c) q_{g,h}^\star(c)\right| \\ &\quad \le C\left|\int_a^b K_{h_g}(c-u)(u-c)\,\mathrm{d}u\right| + C\int_a^b K_{h_g}(c-u)|u-c|^2\,\mathrm{d}u.\end{aligned}$$

The symmetry of K and its finite second moment imply

$$\int_a^b \left|\int_a^b K_{h_g}(c-u)(u-c)\,\mathrm{d}u\right|\mathrm{d}c \le Ch_g^2,$$
$$\int_a^b \int_a^b K_{h_g}(c-u)|u-c|^2\,\mathrm{d}u\,\mathrm{d}c \le Ch_g^2.$$

The first bound follows by extending the inner integral to $\mathbb{R}$, using $\int vK(v)\,\mathrm{d}v = 0$, and bounding the omitted boundary regions by Fubini's theorem. Therefore,

$$\int_C \left|p_{g,h}^\star(c) - \ell_g^\star(c) q_{g,h}^\star(c)\right|\mathrm{d}c \le Ch_g^2.$$

By Lemma 2, $\frac{q_g^\star(c)}{q_{g,h}^\star(c)} \le \kappa_0$, and hence

$$\begin{aligned}&\int_C q_g^\star(c)\left|\ell_{g,h}^\star(c) - \ell_g^\star(c)\right|\mathrm{d}c \\ &\quad = \int_C \frac{q_g^\star(c)}{q_{g,h}^\star(c)}\left|p_{g,h}^\star(c) - \ell_g^\star(c) q_{g,h}^\star(c)\right|\mathrm{d}c \le Ch_g^2.\end{aligned}$$

Summing over the fixed number of groups proves the result. □

Lemma 7 transfers routing stability to the fitted local-FDR values: replacing the full-sample router by its leave-one-out counterpart changes the estimated risk only by the kernel fluctuation and routing-stability terms.

Lemma 7 (Full-versus-LOO local-FDR stability)**.** *Suppose Assumptions 1-2 hold. Then*

$$\frac{1}{m}\sum_{i=1}^m \mathbb{E}\left|\widehat{L}_i - \widetilde{L}_i\right| \le C\left[\max_{g\in[M]}(nh_g)^{-1/2} + \zeta_m + \frac{\gamma_m}{\min_{g\in[M]} h_g}\right]$$

and

$$\frac{1}{m}\sum_{i=1}^m \mathbb{E}|\widehat{L}_i - L_i^\star| \le C\left[\max_{g\in[M]}(h_g^2 + nh_g)^{-1/2} + \zeta_m + \frac{\gamma_m}{\min_{g\in[M]} h_g}\right].$$

Proof. Let $E_i := \left\{\widehat{\rho}(X_i) = \widehat{\rho}^{(-i)}(X_i)\right\}$.

On E_i, the two procedures evaluate their kernel estimators at the same group and confidence value.

Conditional on the test sample, for every g and c,

$$\begin{aligned}&\mathbb{E}_{\mathrm{cal}}\left|\widehat{q}_{\widehat{\rho},g}(c) - \widehat{q}_{\widehat{\rho}^{(-i)},g}(c)\right| \\ &\quad \le \frac{\|K\|_\infty}{h_g} d_P\left(\widehat{\rho}, \widehat{\rho}^{(-i)}\right),\end{aligned} \tag{9}$$

and, because $D \in [0,1]$,

$$\begin{aligned}&\mathbb{E}_{\mathrm{cal}}\left|\widehat{p}_{\widehat{\rho},g}(c) - \widehat{p}_{\widehat{\rho}^{(-i)},g}(c)\right| \\ &\quad \le \frac{\|K\|_\infty}{h_g} d_P\left(\widehat{\rho}, \widehat{\rho}^{(-i)}\right).\end{aligned} \tag{10}$$

We next make explicit the denominator argument. Define

$$A_i := \left\{ \widehat{q}_{\widehat{\rho}^{(-i)},\widetilde{G}_i}(\widetilde{C}_i) \geq \frac{1}{2} q^\star_{\widetilde{G}_i,h}(\widetilde{C}_i) \right\}.$$

By Assumption 2 and Lemma 2, there exists a constant $\underline{q} > 0$ such that, for all sufficiently small bandwidths, $q^\star_{g,h}(c) \geq \underline{q}$ uniformly over $g \in [M]$ and $c \in C$.

On $E_i \cap A_i$, Lemma 1 gives

$$\begin{aligned} &|\widehat{L}_i - \widetilde{L}_i| \mathbf{1}(E_i \cap A_i) \\ &\leq C \Bigg[\left| \widehat{p}_{\widehat{\rho},\widetilde{G}_i}(\widetilde{C}_i) - \widehat{p}_{\widehat{\rho}^{(-i)},\widetilde{G}_i}(\widetilde{C}_i) \right| \\ &\qquad + \left| \widehat{q}_{\widehat{\rho},\widetilde{G}_i}(\widetilde{C}_i) - \widehat{q}_{\widehat{\rho}^{(-i)},\widetilde{G}_i}(\widetilde{C}_i) \right| \Bigg]. \end{aligned} \tag{11}$$

Indeed, on A_i, $\widehat{q}_{\widehat{\rho}^{(-i)},\widetilde{G}_i}(\widetilde{C}_i) \geq \frac{q}{2}$, so the denominator appearing in Lemma 1 is uniformly bounded away from zero.

Combining (9), (10), and (11), we obtain

$$\begin{aligned} &\mathbb{E}\left[|\widehat{L}_i - \widetilde{L}_i| \mathbf{1}(E_i \cap A_i) \right] \\ &\leq \frac{C}{\min_{g \in [M]} h_g} \mathbb{E} d_P\left(\widehat{\rho}, \widehat{\rho}^{(-i)} \right). \end{aligned} \tag{12}$$

It remains to control the bad-denominator event A_i^c. The key point is that $\widehat{\rho}^{(-i)}$ is constructed without using X_i. Thus, conditional on the leave-one-out test sample, X_i is an independent evaluation observation.

Fix the leave-one-out router and the calibration sample, and define

$$B_i(g,c) := \mathbf{1}\left\{ \widehat{q}_{\widehat{\rho}^{(-i)},g}(c) < \frac{1}{2} q^\star_{g,h}(c) \right\}.$$

Under the natural coupling, whenever $\widehat{\rho}^{(-i)}(X_i) = \rho^\star(X_i)$, the routed group and confidence value under the two routers coincide. Therefore,

$$\mathbb{E}_{X_i} B_i\left(\widehat{\rho}^{(-i)}(X_i), C_i^{(\widehat{\rho}^{(-i)}(X_i))} \right) \leq \mathbb{E}_{X_i} B_i\left(\rho^\star(X_i), C_i^{(\rho^\star(X_i))} \right) + d_P\left(\widehat{\rho}^{(-i)}, \rho^\star \right).$$

Taking expectation over the calibration sample,

$$\begin{aligned} \mathbb{P}(A_i^c) \leq &\mathbb{E}\left[\sum_{g=1}^M \int_C q^\star_g(c) \mathbf{1}\left\{ \widehat{q}_{\widehat{\rho}^{(-i)},g}(c) < \frac{1}{2} q^\star_{g,h}(c) \right\} dc \right] \\ &+ \mathbb{E} d_P\left(\widehat{\rho}^{(-i)}, \rho^\star \right). \end{aligned} \tag{13}$$

Whenever $\widehat{q}_{\widehat{\rho}^{(-i)},g}(c) < \frac{1}{2} q^\star_{g,h}(c)$, we necessarily have

$$\left| \widehat{q}_{\widehat{\rho}^{(-i)},g}(c) - q^\star_{g,h}(c) \right| \geq \frac{1}{2} q^\star_{g,h}(c).$$

Hence

$$\mathbf{1}\left\{ \widehat{q}_{\widehat{\rho}^{(-i)},g}(c) < \frac{1}{2} q^\star_{g,h}(c) \right\} \leq \frac{2 \left| \widehat{q}_{\widehat{\rho}^{(-i)},g}(c) - q^\star_{g,h}(c) \right|}{q^\star_{g,h}(c)}.$$

Using Lemma 2, $\frac{q^\star_g(c)}{q^\star_{g,h}(c)} \leq C$, and therefore

$$\mathbb{E}\left[\sum_{g=1}^M \int_C q^\star_g(c) \mathbf{1}\left\{ \widehat{q}_{\widehat{\rho}^{(-i)},g}(c) < \frac{1}{2} q^\star_{g,h}(c) \right\} dc \right] \leq C \sum_{g=1}^M \int_C \mathbb{E}\left| \widehat{q}_{\widehat{\rho}^{(-i)},g}(c) - q^\star_{g,h}(c) \right| dc.$$

We decompose

$$\widehat{q}_{\widehat{\rho}^{(-i)},g} - q^{\star}_{g,h} = \left(\widehat{q}_{\widehat{\rho}^{(-i)},g} - q_{\widehat{\rho}^{(-i)},g,h}\right) + \left(q_{\widehat{\rho}^{(-i)},g,h} - q^{\star}_{g,h}\right).$$

By Lemma 4,

$$\sum_{g=1}^{M} \int_C \mathbb{E}\left|\widehat{q}_{\widehat{\rho}^{(-i)},g}(c) - q_{\widehat{\rho}^{(-i)},g,h}(c)\right| dc \le C \max_{g\in[M]} (nh_g)^{-1/2},$$

whereas Lemma 3 gives

$$\begin{aligned}&\sum_{g=1}^{M} \int_C \left|q_{\widehat{\rho}^{(-i)},g,h}(c) - q^{\star}_{g,h}(c)\right| dc \\ &\quad \le C d_P\left(\widehat{\rho}^{(-i)}, \rho^{\star}\right). \end{aligned} \tag{14}$$

Combining (13)–(14), we obtain

$$\mathbb{P}(A_i^c) \le C\left[\max_{g\in[M]} (nh_g)^{-1/2} + \mathbb{E} d_P\left(\widehat{\rho}^{(-i)}, \rho^{\star}\right)\right]. \tag{15}$$

Since $\widehat{L}_i, \widetilde{L}_i \in [0,1]$,

$$\mathbb{E}\left[|\widehat{L}_i - \widetilde{L}_i| \mathbf{1}(E_i \cap A_i^c)\right] \le \mathbb{P}(A_i^c).$$

Together with (12) and (15), this gives

$$\mathbb{E}\left[|\widehat{L}_i - \widetilde{L}_i| \mathbf{1}(E_i)\right] \le C\left[\max_{g\in[M]} (nh_g)^{-1/2} + \mathbb{E} d_P\left(\widehat{\rho}^{(-i)}, \rho^{\star}\right) + \frac{\mathbb{E} d_P\left(\widehat{\rho}, \widehat{\rho}^{(-i)}\right)}{\min_{g\in[M]} h_g}\right].$$

On the complementary event, since $\widehat{L}_i, \widetilde{L}_i \in [0,1]$,

$$\mathbb{E}\left[|\widehat{L}_i - \widetilde{L}_i| \mathbf{1}(E_i^c)\right] \le \mathbb{P}(E_i^c).$$

Therefore,

$$\begin{aligned}&\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left|\widehat{L}_i - \widetilde{L}_i\right| \\ &\quad \le C\Bigg[\max_{g\in[M]} (nh_g)^{-1/2} + \frac{1}{m}\sum_{i=1}^{m} \mathbb{E} d_P\left(\widehat{\rho}^{(-i)}, \rho^{\star}\right) \\ &\qquad\qquad + \frac{1}{\min_{g\in[M]} h_g} \frac{1}{m}\sum_{i=1}^{m} \mathbb{E} d_P\left(\widehat{\rho}, \widehat{\rho}^{(-i)}\right) + \frac{1}{m}\sum_{i=1}^{m} \mathbb{P}(E_i^c)\Bigg].\end{aligned}$$

By Assumption 1, $\frac{1}{m}\sum_{i=1}^{m} \mathbb{P}(E_i^c) \le C\zeta_m$.

Moreover, the triangle inequality for d_P gives

$$d_P\left(\widehat{\rho}, \widehat{\rho}^{(-i)}\right) \le d_P\left(\widehat{\rho}, \rho^{\star}\right) + d_P\left(\widehat{\rho}^{(-i)}, \rho^{\star}\right).$$

Hence Assumption 1 further implies

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{E} d_P\left(\widehat{\rho}, \widehat{\rho}^{(-i)}\right) \le C\gamma_m,$$

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{E} d_P\left(\widehat{\rho}^{(-i)}, \rho^{\star}\right) \le C\gamma_m.$$

It follows that

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left|\widehat{L}_i-\widetilde{L}_i\right| \le C\left[\max_{g\in[M]}(nh_g)^{-1/2}+\gamma_m+\zeta_m+\frac{\gamma_m}{\min_{g\in[M]}h_g}\right]$$
$$\le C\left[\max_{g\in[M]}(nh_g)^{-1/2}+\zeta_m+\frac{\gamma_m}{\min_{g\in[M]}h_g}\right],$$

where the last inequality uses $\min_{g\in[M]} h_g \le 1$ for all sufficiently large n. This proves the first part of the lemma.

Moreover, as

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}|\widehat{L}_i-L_i^\star| \le \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}|\widehat{L}_i-\widetilde{L}_i| + \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}|\widetilde{L}_i-L_i^\star|.$$

As X_i is independent of $\hat{\rho}^{-i}$ and $\rho^\star$, invoking Proposition 1 with $\hat{\rho}^{-i}$ and $\rho^\star$ gives the bound of the second term, which proves the result. □

Lemma 8 propagates local-FDR estimation error through the pooled thresholding step, showing that both the empirical cutoff and the reported set approach their population counterparts.

Lemma 8 (Consistency of the pooled selection rule)**.** *Suppose Assumptions 1–3 hold. Then*

$$\mathbb{E}\left|\widehat{t}_\alpha - t_\alpha^\star\right| \le C(m^{-1/2}+r_{n,m}^{1/2}).$$

Moreover, letting $R_i^\star := \mathbf{1}\left\{L_i^\star \le t_\alpha^\star\right\}$*, with the fixed label-free tie-breaking convention described above, we have*

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left|R_i-R_i^\star\right| \le C(m^{-1/2}+r_{n,m}^{1/2}).$$

Proof. For $t\in[0,1]$, define $\phi_t(\ell) := (\ell-\alpha)\mathbf{1}\{\ell\le t\}$.

Then $\widehat{\Psi}_{\alpha,m}(t) = \frac{1}{m}\sum_{i=1}^{m}\phi_t(\widehat{L}_i)$, whereas the oracle empirical criterion is $\Psi^\star_{\alpha,m}(t) := \frac{1}{m}\sum_{i=1}^{m}\phi_t(L_i^\star)$.

Let

$$F_m^\star(t) := \frac{1}{m}\sum_{i=1}^{m}\mathbf{1}\{L_i^\star\le t\}, \qquad F^\star(t) := \mathbb{P}(L^\star\le t),$$

and define $\Delta_m := \sup_{t\in[0,1]}\left|F_m^\star(t)-F^\star(t)\right|$.

Because $L_1^\star,\dots,L_m^\star$ are i.i.d., the Dvoretzky–Kiefer–Wolfowitz inequality gives $\mathbb{E}[\Delta_m]\le Cm^{-1/2}$.

An integration-by-parts argument gives

$$\Psi^\star_{\alpha,m}(t)-\Psi_\alpha(t) = \int_{[0,t]}(\ell-\alpha)\,d\{F_m^\star-F^\star\}(\ell)$$
$$= (t-\alpha)\{F_m^\star(t)-F^\star(t)\} - \int_0^t\{F_m^\star(u)-F^\star(u)\}\,du.$$

Consequently,

$$\sup_{t\in[0,1]}\left|\Psi^\star_{\alpha,m}(t)-\Psi_\alpha(t)\right| \le 2\Delta_m. \tag{16}$$

We next compare the full-sample criterion with its oracle empirical counterpart. Write

$$e_i := \left|\widehat{L}_i-L_i^\star\right|, \qquad \overline{e}_m := \frac{1}{m}\sum_{i=1}^{m}e_i.$$

Then by Lemma 7

$$\mathbb{E}[\overline{e}_m] \le Cr_{n,m}. \tag{17}$$

Fix $v > 0$ and t satisfying $t - 2v > \alpha$. If $e_i \le v$, a direct case analysis gives

$$\phi_{t-v}(L_i^\star) - e_i \le \phi_t(\widehat{L}_i) \le \phi_{t+v}(L_i^\star) + e_i.$$

Indeed, when $\widehat{L}_i \le t$, one has $L_i^\star \le t + v$, whereas when $L_i^\star \le t - v$, one has $\widehat{L}_i \le t$. The condition $t - 2v > \alpha$ ensures that observations whose inclusion status differs across the two thresholds have nonnegative contributions.

Since $|\phi_t| \le 1$, the preceding inequality can be extended to all i as

$$\phi_{t-v}(L_i^\star) - e_i - 2\mathbf{1}\{e_i > v\} \le \phi_t(\widehat{L}_i) \le \phi_{t+v}(L_i^\star) + e_i + 2\mathbf{1}\{e_i > v\}.$$

Averaging over i and using $\frac{1}{m}\sum_{i=1}^m \mathbf{1}\{e_i > v\} \le \frac{\overline{e}_m}{v}$, we obtain

$$\Psi^\star_{\alpha,m}(t - v) - \overline{e}_m - \frac{2\overline{e}_m}{v} \le \widehat{\Psi}_{\alpha,m}(t) \le \Psi^\star_{\alpha,m}(t + v) + \overline{e}_m + \frac{2\overline{e}_m}{v}. \tag{18}$$

We now use (18) to localize the empirical cutoff. Define

$$d_m := \Delta_m + \overline{e}_m^{1/2}, \qquad v_m := \overline{e}_m^{1/2},$$

with the convention that $\overline{e}_m/v_m = 0$ when $\overline{e}_m = 0$. Choose a sufficiently large constant $A > 1$ such that $c_\Psi(A - 1) > 3$, and put $u_m := Ad_m$.

Consider first the event on which

$$u_m \le u_0, \qquad (A + 2)d_m < t_\alpha^\star - \alpha. \tag{19}$$

On this event,

$$t_\alpha^\star - u_m - 2v_m > \alpha.$$

Applying the upper bound in (18) with $t = t_\alpha^\star - u_m$ gives

$$\begin{aligned}\widehat{\Psi}_{\alpha,m}(t_\alpha^\star - u_m) &\le \Psi^\star_{\alpha,m}(t_\alpha^\star - u_m + v_m) + \overline{e}_m + 2\overline{e}_m^{1/2} \\ &\le \Psi_\alpha(t_\alpha^\star - (u_m - v_m)) + 2\Delta_m + \overline{e}_m + 2\overline{e}_m^{1/2}.\end{aligned}$$

Since $u_m - v_m \ge (A - 1)d_m$, the population crossing condition in Assumption 3 yields

$$\Psi_\alpha(t_\alpha^\star - (u_m - v_m)) \le -c_\Psi(u_m - v_m) \le -c_\Psi(A - 1)d_m.$$

Moreover, because $\overline{e}_m \le \overline{e}_m^{1/2}$,

$$2\Delta_m + \overline{e}_m + 2\overline{e}_m^{1/2} \le 3d_m.$$

Therefore,

$$\widehat{\Psi}_{\alpha,m}(t_\alpha^\star - u_m) \le -\{c_\Psi(A - 1) - 3\}d_m < 0. \tag{20}$$

Similarly, applying the lower bound in (18) with $t = t_\alpha^\star + u_m$ gives

$$\begin{aligned}\widehat{\Psi}_{\alpha,m}(t_\alpha^\star + u_m) &\ge \Psi^\star_{\alpha,m}(t_\alpha^\star + u_m - v_m) - \overline{e}_m - 2\overline{e}_m^{1/2} \\ &\ge \Psi_\alpha(t_\alpha^\star + (u_m - v_m)) - 2\Delta_m - \overline{e}_m - 2\overline{e}_m^{1/2} \\ &\ge \{c_\Psi(A - 1) - 3\}d_m > 0.\end{aligned} \tag{21}$$

Because $\widehat{\Psi}_{\alpha,m}(t)$ is nondecreasing for $t > \alpha$, (20)–(21) imply

$$t_\alpha^\star - u_m \le \widehat{t}_\alpha \le t_\alpha^\star + u_m.$$

Thus, on the event (19), $\left|\widehat{t}_\alpha - t_\alpha^\star\right| \le Ad_m$.

On the complement of (19), d_m is bounded below by a fixed positive constant depending only on u_0, A, and $t_\alpha^\star - \alpha$. Since both cutoffs belong to $[0, 1]$, after enlarging the constant A above, we obtain

$$\left|\widehat{t}_\alpha - t_\alpha^\star\right| \le C\left(\Delta_m + \overline{e}_m^{1/2}\right).$$

Taking expectations and applying Jensen's inequality, (16), and (17), we conclude that

$$\begin{aligned}\mathbb{E}\left|\widehat{t}_\alpha - t^\star_\alpha\right| &\le C\left\{\mathbb{E}[\Delta_m] + \mathbb{E}[\overline{e}_m^{1/2}]\right\}\\ &\le C\left\{m^{-1/2} + \left(\mathbb{E}\overline{e}_m\right)^{1/2}\right\}\\ &\le C\left(m^{-1/2} + r_{n,m}^{1/2}\right).\end{aligned} \tag{22}$$

Markov's inequality further gives $\left|\widehat{t}_\alpha - t^\star_\alpha\right| = O_{\mathbb{P}}(m^{-1/2} + r_{n,m}^{1/2})$.

It remains to compare the corresponding reporting indicators. Let $T_m := \left|\widehat{t}_\alpha - t^\star_\alpha\right|$.

For any deterministic $v > 0$, if

$$e_i \le v \qquad \text{and} \qquad |L_i^\star - t^\star_\alpha| > T_m + v,$$

then $\widehat{L}_i$ and $L_i^\star$ lie on the same side of their respective cutoffs. Hence

$$\left|R_i - R_i^\star\right| \le \mathbf{1}\{e_i > v\} + \mathbf{1}\left\{|L_i^\star - t^\star_\alpha| \le T_m + v\right\}.$$

Averaging gives

$$\frac{1}{m}\sum_{i=1}^m \left|R_i - R_i^\star\right| \le \frac{\overline{e}_m}{v} + \frac{1}{m}\sum_{i=1}^m \mathbf{1}\left\{|L_i^\star - t^\star_\alpha| \le T_m + v\right\}. \tag{23}$$

The anti-concentration condition in Assumption 3, together with the trivial bound by one outside the neighborhood of $t^\star_\alpha$, implies that there exists $C_0 < \infty$ such that

$$\mathbb{P}\left\{|L^\star - t^\star_\alpha| \le w\right\} \le C_0 w \qquad \text{for every } w > 0.$$

Consequently, by the definition of Δ_m,

$$\frac{1}{m}\sum_{i=1}^m \mathbf{1}\left\{|L_i^\star - t^\star_\alpha| \le w\right\} \le C_0 w + 2\Delta_m \qquad \text{for every } w > 0.$$

Applying it to the random radius $w = T_m + v$ in (23) yields

$$\frac{1}{m}\sum_{i=1}^m \left|R_i - R_i^\star\right| \le \frac{\overline{e}_m}{v} + C_0(T_m + v) + 2\Delta_m.$$

Take $v = r_{n,m}^{1/2}$, with the result for $r_{n,m} = 0$ obtained by taking a limit. Taking expectations in the above inequality and using (17), (22), and $\mathbb{E}\Delta_m \le Cm^{-1/2}$ gives

$$\frac{1}{m}\sum_{i=1}^m \mathbb{E}\left|R_i - R_i^\star\right| \le C\left\{m^{-1/2} + r_{n,m}^{1/2}\right\}.$$

This proves the result. □

Finally, we introduce Lemma 9. This final stability step replaces the actual reporting indicator by a leave-one-out counterpart that is decoupled from the i-th test observation, enabling the conditional-expectation argument in the proof of Theorem 1.

Lemma 9 (Leave-one-out stability of the reporting indicators)**.** *For each $i \in [m]$, define the true local disagreement risk under the leave-one-out router by $L_i^{(-i)} := \ell_{\widehat{\rho}^{(-i)},\widetilde{G}_i}(\widetilde{C}_i)$, and define the auxiliary leave-one-out reporting indicator $R_i^{(-i)} := \mathbf{1}\left\{L_i^{(-i)} \le t^\star_\alpha\right\}$.*

Let π be the permutation that orders the estimated risks $\widehat{L}_i$ increasingly, with ties broken by a fixed, label-free deterministic order. Let $\widehat{k}$ be defined by the ordered-prefix rule in Equation (3), and retain the reporting indicators

$$R_i := \mathbf{1}\{\pi^{-1}(i) \le \widehat{k}\}, \qquad R_i^\star := \mathbf{1}\{L_i^\star \le t^\star_\alpha\}.$$

Here $\widehat{t}_\alpha$ is only an auxiliary boundary:

$$\mathbf{1}\{\widehat{L}_i < \widehat{t}_\alpha\} \le R_i \le \mathbf{1}\{\widehat{L}_i \le \widehat{t}_\alpha\}.$$

We do not identify R_i with the inclusive-threshold indicator at $\widehat{t}_\alpha$.

Then

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left|R_i^{(-i)} - R_i^\star\right| \le C\gamma_m^{1/2},$$

and

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left|R_i - R_i^{(-i)}\right| \le C\left(m^{-1/2} + r_{n,m}^{1/2}\right).$$

In particular, since $r_{n,m} \ge \gamma_m$ for all sufficiently large n, m,

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left|R_i - R_i^{(-i)}\right| \le C(m^{-1/2} + r_{n,m}^{1/2}) = o(1).$$

Proof. Conditional on $\mathcal{F}^{-i}$, the router $\widehat{\rho}^{(-i)}$ is fixed and X_i is an independent evaluation observation. Applying the integrated stability of conditional disagreement risks to $\widehat{\rho}^{(-i)}$ and $\rho^\star$ gives

$$\mathbb{E}\left[\left|L_i^{(-i)} - L_i^\star\right| \middle| \mathcal{F}^{-i}\right] \le C d_P\left(\widehat{\rho}^{(-i)}, \rho^\star\right).$$

Therefore,

$$\delta_m := \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left|L_i^{(-i)} - L_i^\star\right| \le C\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}d_P\left(\widehat{\rho}^{(-i)}, \rho^\star\right) \le C\gamma_m.$$

Fix $0 < u \le u_0$. If $R_i^{(-i)} \ne R_i^\star$,

then either $L_i^\star$ lies within distance u of the population cutoff or the leave-one-out risk differs from its oracle counterpart by more than u. More precisely,

$$\left\{R_i^{(-i)} \ne R_i^\star\right\} \subseteq \left\{|L_i^\star - t_\alpha^\star| \le u\right\} \cup \left\{|L_i^{(-i)} - L_i^\star| > u\right\}.$$

Consequently,

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left|R_i^{(-i)} - R_i^\star\right| &\le \frac{1}{m}\sum_{i=1}^{m}\mathbb{P}\left\{|L_i^\star - t_\alpha^\star| \le u\right\} \\
&\quad + \frac{1}{m}\sum_{i=1}^{m}\mathbb{P}\left\{|L_i^{(-i)} - L_i^\star| > u\right\} \\
&\le C_L u + \frac{\delta_m}{u} \\
&\le C\left(u + \frac{\gamma_m}{u}\right),
\end{aligned}$$

where the second inequality follows from the anti-concentration condition and Markov's inequality. Taking $u = \gamma_m^{1/2}$ gives

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left|R_i^{(-i)} - R_i^\star\right| \le C\gamma_m^{1/2}. \tag{24}$$

By the triangle inequality,

$$\left|R_i - R_i^{(-i)}\right| \le \left|R_i - R_i^\star\right| + \left|R_i^{(-i)} - R_i^\star\right|.$$

Averaging over i and applying Lemma 8 and (24), we obtain

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left|R_i - R_i^{(-i)}\right| \le C\left(m^{-1/2} + r_{n,m}^{1/2}\right).$$

□

C.3 Proof of Proposition 1

Proof. Write $\Delta_\rho := d_P(\rho, \rho^\star)$.

All the arguments below are conditional on $\mathcal{F}$, so that ρ and Δ_ρ are fixed.

We work on the kernel estimator. The key point is that we do not apply a population smoothing-bias argument to the potentially irregular distribution induced by ρ. Instead, we apply only the distribution-free stochastic part of the kernel calculation to ρ, while using the regular oracle-smoothed regression as an anchor for the ratio comparison.

Recall that

$$q_g^\star(c) := \pi_g^\star f_g^\star(c), \qquad p_g^\star(c) := q_g^\star(c)\ell_g^\star(c),$$

and define

$$q_{g,h}^\star := \int_C K_{h_g}(c-u)q_g^\star(u)\,\mathrm{d}u, \qquad p_{g,h}^\star := \int_C K_{h_g}(c-u)p_g^\star(u)\,\mathrm{d}u, \qquad \ell_{g,h}^\star(c) := \frac{p_{g,h}^\star(c)}{q_{g,h}^\star(c)}.$$

By Lemma 2, $q_g^\star(c) \le \kappa_0 q_{g,h}^\star(c)$ uniformly over $g \in [M]$ and $c \in C$.

Apply Lemma 1 with $(A_0, B_0) = (p_{g,h}^\star, q_{g,h}^\star)$ and $(A_1, B_1) = (\widehat{p}_{\rho,g}, \widehat{q}_{\rho,g})$.

Since all the corresponding ratios take values in $[0,1]$, we obtain

$$\begin{aligned}
&\sum_{g=1}^M \int_C q_g^\star(c)\mathbb{E}_{\mathrm{cal}}\left[\left|\widehat{\ell}_{\rho,g}(c) - \ell_{g,h}^\star(c)\right| \middle| \mathcal{F}\right] \mathrm{d}c \\
&\quad \le C\sum_{g=1}^M \int_C \mathbb{E}_{\mathrm{cal}}\left[\left|\widehat{p}_{\rho,g}(c) - p_{g,h}^\star(c)\right| + \left|\widehat{q}_{\rho,g}(c) - q_{g,h}^\star(c)\right| \middle| \mathcal{F}\right] \mathrm{d}c. \qquad (25)
\end{aligned}$$

For the denominator, write

$$\widehat{q}_{\rho,g} - q_{g,h}^\star = \left(\widehat{q}_{\rho,g} - q_{\rho,g,h}\right) + \left(q_{\rho,g,h} - q_{g,h}^\star\right),$$

and similarly,

$$\widehat{p}_{\rho,g} - p_{g,h}^\star = \left(\widehat{p}_{\rho,g} - p_{\rho,g,h}\right) + \left(p_{\rho,g,h} - p_{g,h}^\star\right).$$

Lemma 4 gives

$$\begin{aligned}
&\sum_{g=1}^M \int_C \mathbb{E}_{\mathrm{cal}}\left[\left|\widehat{q}_{\rho,g}(c) - q_{\rho,g,h}(c)\right| + \left|\widehat{p}_{\rho,g}(c) - p_{\rho,g,h}(c)\right| \middle| \mathcal{F}\right] \mathrm{d}c \\
&\quad \le C \max_{g\in[M]} (nh_g)^{-1/2}. \qquad (26)
\end{aligned}$$

On the other hand, Lemma 3 and the L_1-contraction property of convolution imply

$$\sum_{g=1}^M \int_C \left\{|q_{\rho,g,h}(c) - q_{g,h}^\star(c)| + |p_{\rho,g,h}(c) - p_{g,h}^\star(c)|\right\} \mathrm{d}c \le C\Delta_\rho. \qquad (27)$$

Combining (25), (26), and (27), we obtain

$$\begin{aligned}
&\sum_{g=1}^M \int_C q_g^\star(c)\mathbb{E}_{\mathrm{cal}}\left[\left|\widehat{\ell}_{\rho,g}(c) - \ell_{g,h}^\star(c)\right| \middle| \mathcal{F}\right] \mathrm{d}c \\
&\quad \le C\left\{\max_{g\in[M]} (nh_g)^{-1/2} + \Delta_\rho\right\}. \qquad (28)
\end{aligned}$$

We next transfer (28) from the fixed subgroup g to the data-driven subgroup indicator G_ρ induced by ρ. Fix $\mathcal{F}$ and a realization of the calibration sample, and define

$$H(g,c) := \left|\widehat{\ell}_{\rho,g}(c) - \ell_{g,h}^\star(c)\right|.$$

For the fixed calibration sample, H is a deterministic function taking values in $[0, 1]$.

Under the natural coupling, whenever $\rho(X) = \rho^\star(X)$, we have $(G_\rho, C_\rho) = (G^\star, C^\star)$.

It follows pointwise that

$$H(G_\rho, C_\rho) \le H(G^\star, C^\star) + \mathbf{1}\left\{\rho(X) \ne \rho^\star(X)\right\}.$$

Taking expectation over the independent evaluation observation X, conditional on $\mathcal{F}$ and with the calibration sample fixed, gives

$$\begin{aligned}\mathbb{E}\left[H(G_\rho, C_\rho)\,\middle|\,\mathcal{F}\right] &\le \mathbb{E}\left[H(G^\star, C^\star)\,\middle|\,\mathcal{F}\right] + \Delta_\rho \\ &= \sum_{g=1}^{M}\int_C q_g^\star(c) H(g, c)\,\mathrm{d}c + \Delta_\rho,\end{aligned}$$

where the last equality follows from $q_g^\star(c)\,\mathrm{d}c = \mathbb{P}\left\{G^\star = g,\ C^\star \in \mathrm{d}c\right\}$.

Averaging the preceding inequality over the calibration sample yields

$$\begin{aligned}&\mathbb{E}\left[\left|\widehat{\ell}_{\rho,G_\rho}(C_\rho) - \ell^\star_{G_\rho,h}(C_\rho)\right|\,\middle|\,\mathcal{F}\right] \\ &\quad\le \sum_{g=1}^{M}\int_C q_g^\star(c)\,\mathbb{E}_{\mathrm{cal}}\left[\left|\widehat{\ell}_{\rho,g}(c) - \ell^\star_{g,h}(c)\right|\,\middle|\,\mathcal{F}\right]\mathrm{d}c + \Delta_\rho.\end{aligned}$$

Therefore, (28) implies

$$\mathbb{E}\left[\left|\widehat{L}_\rho - \ell^\star_{G_\rho,h}(C_\rho)\right|\,\middle|\,\mathcal{F}\right] \le C\left\{\max_{g\in[M]}(nh_g)^{-1/2} + \Delta_\rho\right\}. \tag{29}$$

By Lemma 6,

$$\mathbb{E}\left|\ell^\star_{G^\star,h}(C^\star) - L^\star\right| \le C\max_{g\in[M]} h_g^2. \tag{30}$$

Under the natural coupling, whenever $\rho(X) = \rho^\star(X)$, we have $(G_\rho, C_\rho) = (G^\star, C^\star)$.

Since $\ell^\star_{g,h}(c) \in [0, 1]$, it follows that

$$\begin{aligned}&\mathbb{E}\left[\left|\ell^\star_{G_\rho,h}(C_\rho) - \ell^\star_{G^\star,h}(C^\star)\right|\,\middle|\,\mathcal{F}\right] \\ &\qquad\le \mathbb{P}\left\{\rho(X) \ne \rho^\star(X)\,\middle|\,\mathcal{F}\right\} = \Delta_\rho.\end{aligned} \tag{31}$$

Combining (29), (31), and (30), we obtain

$$\begin{aligned}&\mathbb{E}\left[\left|\widehat{L}_\rho - L^\star\right|\,\middle|\,\mathcal{F}\right] \\ &\quad\le \mathbb{E}\left[\left|\widehat{L}_\rho - \ell^\star_{G_\rho,h}(C_\rho)\right|\,\middle|\,\mathcal{F}\right] \\ &\qquad+ \mathbb{E}\left[\left|\ell^\star_{G_\rho,h}(C_\rho) - \ell^\star_{G^\star,h}(C^\star)\right|\,\middle|\,\mathcal{F}\right] \\ &\qquad+ \mathbb{E}\left|\ell^\star_{G^\star,h}(C^\star) - L^\star\right| \\ &\quad\le C\left[\max_{g\in[M]}\left\{h_g^2 + (nh_g)^{-1/2}\right\} + \Delta_\rho\right].\end{aligned}$$

Taking expectation with respect to $\mathcal{F}$ and using $\mathbb{E}\Delta_\rho = \mathbb{P}\left\{\rho(X) \ne \rho^\star(X)\right\}$ proves the theorem.

□

C.4 Proof of Theorem 1

Proof. Define $V = \sum_{j=1}^m D_j R_j$ and $R = \sum_{j=1}^m R_j$, thereby mFDR $= \mathbb{E}[V]/\mathbb{E}[R]$. By construction of the pooled reporting rule, $\sum_{i=1}^m (\widehat{L}_i - \alpha) R_i \le 0$ for every realization of the calibration and test

samples. Therefore,

$$\begin{aligned}\mathbb{E}[V] - \alpha\mathbb{E}[R] &= \sum_{i=1}^{m} \mathbb{E}\left[(D_i - \alpha)R_i\right] \\ &= \sum_{i=1}^{m} \mathbb{E}\left[(D_i - \widehat{L}_i)R_i\right] + \mathbb{E}\left[\sum_{i=1}^{m}(\widehat{L}_i - \alpha)R_i\right] \\ &\leq \sum_{i=1}^{m} \mathbb{E}\left[(D_i - \widehat{L}_i)R_i\right]. \end{aligned} \tag{32}$$

Let $\mathcal{H}_i^{(-i)} := \sigma\left(\mathcal{F}_{-i}, \widetilde{G}_i, \widetilde{C}_i\right)$.

If an auxiliary label-free tie-breaking variable is used, it is also included in $\mathcal{H}_i^{(-i)}$. By construction, $R_i^{(-i)}$ is $\mathcal{H}_i^{(-i)}$-measurable. Moreover, $L_i^{0,(-i)} = \mathbb{E}[D_i \mid \mathcal{H}_i^{(-i)}]$.

We next control the right-hand side of (32) using leave-one-out quantities. For every i,

$$\begin{aligned}(D_i - \widehat{L}_i)R_i = (D_i - L_i^{0,(-i)})R_i^{(-i)} &\\ + (L_i^{0,(-i)} - \widetilde{L}_i)R_i^{(-i)} &\\ + (D_i - \widetilde{L}_i)(R_i - R_i^{(-i)}) &\\ + (\widetilde{L}_i - \widehat{L}_i)R_i. &\end{aligned} \tag{33}$$

It follows that

$$\mathbb{E}\left[(D_i - L_i^{0,(-i)})R_i^{(-i)}\right] = \mathbb{E}\left[R_i^{(-i)}\mathbb{E}\left[D_i - L_i^{0,(-i)} \,\middle|\, \mathcal{H}_i^{(-i)}\right]\right] = 0.$$

Since $0 \leq D_i, \widehat{L}_i, \widetilde{L}_i, R_i, R_i^{(-i)} \leq 1$, the remaining three terms in (33) satisfy

$$\left|(L_i^{0,(-i)} - \widetilde{L}_i)R_i^{(-i)}\right| \leq \left|L_i^{0,(-i)} - \widetilde{L}_i\right|,$$

$$\left|(D_i - \widetilde{L}_i)(R_i - R_i^{(-i)})\right| \leq \left|R_i - R_i^{(-i)}\right|, \tag{34}$$

$$\left|(\widetilde{L}_i - \widehat{L}_i)R_i\right| \leq \left|\widetilde{L}_i - \widehat{L}_i\right|. \tag{35}$$

As for (33), by the triangle inequality,

$$\left|L_i^{0,(-i)} - \widetilde{L}_i\right| \leq \left|L_i^{0,(-i)} - L_i^{(-i)}\right| + \left|L_i^{(-i)} - \widetilde{L}_i\right|.$$

To bound the first term, let $E_i := \left\{\widehat{\rho}(X_i) = \widehat{\rho}^{(-i)}(X_i)\right\}$.

On E_i, the full-sample and leave-one-out routers invoke the same Judge, and hence the corresponding disagreement variables coincide. Therefore,

$$\begin{aligned}\left|L_i^{0,(-i)} - L_i^{(-i)}\right| &\leq \mathbb{E}\left[\left|D_i - D_i^{(\widetilde{G}_i)}\right| \mid \mathcal{H}_i^{(-i)}\right] \\ &\leq \mathbb{P}\left(E_i^c \mid \mathcal{H}_i^{(-i)}\right).\end{aligned}$$

Averaging over i gives

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left|L_i^{0,(-i)} - L_i^{(-i)}\right| \leq \frac{1}{m}\sum_{i=1}^{m}\mathbb{P}(E_i^c) \leq C\zeta_m.$$

For the second term, invoking Lemma 5 and Proposition 1 with $\hat{\rho}^{-i}$ and $\rho^\star$ gives

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left|L_i^{(-i)} - \widetilde{L}_i\right| &\leq \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left|L_i^{(-i)} - L_i^\star\right| + \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left|\widetilde{L}_i - L_i^\star\right| \\ &\leq C\left[\max_{g\in[M]}\left\{h_g^2 + (nh_g)^{-1/2}\right\} + \gamma_m\right].\end{aligned}$$

Since X_i is independent of the data used to construct $\widehat{\rho}^{(-i)}$, the results apply conditionally on the leave-one-out test sample

Consequently,

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left|L_i^{0,(-i)}-\widetilde{L}_i\right| \le C\left[\max_{g\in[M]}\left\{h_g^2+(nh_g)^{-1/2}\right\}+\gamma_m+\zeta_m\right].$$

As for (34) and (35), they are bounded directly by Lemma 9 and Lemma 7, respectively.

Combining (33)–(35) and averaging over i, we obtain

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[(D_i-\widehat{L}_i)R_i\right] \\
&\quad\le \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left|L_i^{0,(-i)}-\widetilde{L}_i\right|+\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left|R_i-R_i^{(-i)}\right| \\
&\qquad+\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left|\widetilde{L}_i-\widehat{L}_i\right| \\
&\quad\le C\left(m^{-1/2}+r_{n,m}^{1/2}\right). \qquad (36)
\end{aligned}$$

Substituting (36) into (32) gives

$$\mathbb{E}[V]-\alpha\mathbb{E}[R]\le Cm(m^{-1/2}+r_{n,m}^{1/2}). \qquad (37)$$

It remains to lower-bound the expected number of reported verdicts. Let $R_i^\star := \mathbf{1}\left\{L_i^\star \le t_\alpha^\star\right\}$.

By Assumption 3, $\mathbb{E}[R_i^\star]=\mathbb{P}\left(L^\star\le t_\alpha^\star\right)=p_\alpha^\star\ge p_0$.

As it is proved by Lemma 8, the consistency of the pooled selection rule gives $\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left|R_i-R_i^\star\right| = o(1)$.

Consequently,

$$\begin{aligned}
\frac{\mathbb{E}[R]}{m} &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[R_i] \\
&\ge \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[R_i^\star]-\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left|R_i-R_i^\star\right| \\
&\ge p_0-o(1)\ge\frac{p_0}{2} \qquad (38)
\end{aligned}$$

for all sufficiently large n, m.

Finally, combining (37) and (38),

$$\mathrm{mFDR}-\alpha=\frac{\mathbb{E}[V]-\alpha\mathbb{E}[R]}{\mathbb{E}[R]}\le\frac{Cm(m^{-1/2}+r_{n,m}^{1/2})}{mp_0/2}\le\frac{C}{p_0}\left(m^{-1/2}+r_{n,m}^{1/2}\right),$$

after absorbing numerical constants into C. Since $m^{-1/2}+r_{n,m}^{1/2}\to 0$, the desired asymptotic mFDR control follows. □

C.5 Verification of Routing Stability

We first give a general sufficient condition for score-based routing, and then verify Propositions 2 and 3.

C.5.1 A general stability result for score-threshold routing

Consider a data-dependent score $\widehat{S} : \mathcal{X} \to \mathbb{R}$ and thresholds

$$-\infty = \widehat{t}_0 < \widehat{t}_1 < \cdots < \widehat{t}_{M-1} < \widehat{t}_M = \infty,$$

which induce the router

$$\widehat{\rho}(x) = g \quad \Longleftrightarrow \quad \widehat{t}_{g-1} < \widehat{S}(x) \le \widehat{t}_g.$$

Let $S^\star$ and

$$-\infty = t_0^\star < t_1^\star < \cdots < t_{M-1}^\star < t_M^\star = \infty$$

define the corresponding population router

$$\rho^\star(x) = g \quad \Longleftrightarrow \quad t_{g-1}^\star < S^\star(x) \le t_g^\star.$$

Let $\widehat{S}^{(-i)}$ and $\widehat{t}_g^{(-i)}$ denote their leave-one-out counterparts.

Proposition 4 (A general sufficient condition for stable score routing)**.** *Suppose M is fixed and the population routing boundaries satisfy the margin condition*

$$\mathbb{P}\left\{ \min_{g\in[M-1]} |S^\star(X) - t_g^\star| \le u \right\} \le Cu$$

for all sufficiently small $u > 0$.

Suppose further that, for some deterministic sequence $a_m \to 0$,

$$\mathbb{E}\left[\|\widehat{S} - S^\star\|_\infty + \max_{g\in[M-1]} |\widehat{t}_g - t_g^\star| \right] \le Ca_m,$$

and

$$\frac{1}{m}\sum_{i=1}^m \mathbb{E}\left[\|\widehat{S}^{(-i)} - S^\star\|_\infty + \max_{g\in[M-1]} |\widehat{t}_g^{(-i)} - t_g^\star| \right] \le Ca_m.$$

Then

$$\mathbb{P}\left\{ \widehat{\rho}(X) \ne \rho^\star(X) \right\} \le Ca_m$$

and

$$\frac{1}{m}\sum_{i=1}^m \mathbb{P}\left\{ \widehat{\rho}^{(-i)}(X) \ne \rho^\star(X) \right\} \le Ca_m.$$

Moreover, suppose that for some deterministic $b_m \to 0$,

$$\|\widehat{S} - \widehat{S}^{(-i)}\|_\infty + \max_{g\in[M-1]} |\widehat{t}_g - \widehat{t}_g^{(-i)}| \le Cb_m$$

almost surely for every i, and that, conditional on the leave-one-out test sample,

$$\mathbb{P}\left\{ \min_{g\in[M-1]} \left| \widehat{S}^{(-i)}(X) - \widehat{t}_g^{(-i)} \right| \le u \,\middle|\, \mathcal{F}_{-i} \right\} \le Cu$$

almost surely for all sufficiently small $u > 0$. Then

$$\frac{1}{m}\sum_{i=1}^m \mathbb{P}\left\{ \widehat{\rho}(X_i) \ne \widehat{\rho}^{(-i)}(X_i) \right\} \le Cb_m.$$

Consequently, Assumption 1 holds with

$$\gamma_m = O(a_m), \qquad \zeta_m = O(b_m).$$

Proof. Define

$$\delta := \|\widehat{S} - S^\star\|_\infty + \max_{g\in[M-1]} |\widehat{t}_g - t_g^\star|.$$

If $\widehat{\rho}(X) \ne \rho^\star(X)$, then the two routing rules place X on different sides of at least one routing boundary. Therefore, $\min_{g\in[M-1]} |S^\star(X) - t_g^\star| \le \delta$.

Conditional on the test sample, δ is fixed and X is independent of the learned router. Hence the margin condition gives $d_P(\widehat{\rho}, \rho^\star) \le C\delta$ for sufficiently small δ. Enlarging C covers the remaining values of δ, since the left-hand side is bounded by one. Taking expectations yields $\mathbb{P}\{\widehat{\rho}(X) \neq \rho^\star(X)\} \le Ca_m$.

The same argument applied to each $\widehat{\rho}^{(-i)}$, followed by averaging over i, gives

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{P}\left\{\widehat{\rho}^{(-i)}(X) \neq \rho^\star(X)\right\} \le Ca_m.$$

We next consider self-influence. If $\widehat{\rho}(X_i) \neq \widehat{\rho}^{(-i)}(X_i)$, then, for at least one $g \in [M-1]$,

$$\left|\widehat{S}^{(-i)}(X_i) - \widehat{t}_g^{(-i)}\right| \le \|\widehat{S} - \widehat{S}^{(-i)}\|_\infty + \max_{g\in[M-1]} |\widehat{t}_g - \widehat{t}_g^{(-i)}| \le Cb_m.$$

Since X_i is independent of $\mathcal{F}_{-i}$, the conditional margin condition therefore gives

$$\mathbb{P}\left\{\widehat{\rho}(X_i) \neq \widehat{\rho}^{(-i)}(X_i) \,\middle|\, \mathcal{F}_{-i}\right\} \le Cb_m.$$

Taking expectations and averaging over i proves

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{P}\left\{\widehat{\rho}(X_i) \neq \widehat{\rho}^{(-i)}(X_i)\right\} \le Cb_m.$$

□

C.5.2 Proof of Proposition 2

Proof. Write

$$Z_i := S(X_i), \qquad i = 1, \dots, m,$$

and let $Z_{(1)} < \cdots < Z_{(m)}$ denote their order statistics. The strict inequalities hold almost surely because $S(X)$ has a continuous density.

For definiteness, take the empirical u_g-quantile to be

$$\widehat{q}_g := Z_{(k_g)}, \qquad k_g := \lceil mu_g \rceil.$$

Other standard rank-based definitions lead to the same rates.

Let $q_g^\star := F_S^{-1}(u_g)$. Since the density is bounded away from zero and bounded above in a neighborhood of $q_g^\star$, the standard empirical-quantile bound gives $\mathbb{E}|\widehat{q}_g - q_g^\star| = O(m^{-1/2})$ for every $g \in [M-1]$. Since M is fixed,

$$\mathbb{E}\max_{g\in[M-1]} |\widehat{q}_g - q_g^\star| = O(m^{-1/2}).$$

Moreover, the bounded density implies the boundary margin condition

$$\mathbb{P}\left\{\min_{g\in[M-1]} |S(X) - q_g^\star| \le u\right\} \le Cu.$$

Applying Proposition 4 with the fixed score $\widehat{S} = S^\star = S$, therefore gives $\mathbb{P}\{\widehat{\rho}(X) \neq \rho^\star(X)\} = O(m^{-1/2})$.

The same argument for the leave-one-out empirical quantiles, which are computed from $m-1$ i.i.d. observations, gives

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{P}\left\{\widehat{\rho}^{(-i)}(X) \neq \rho^\star(X)\right\} = O(m^{-1/2}).$$

Hence the first part of Assumption 1 holds with $\gamma_m = O(m^{-1/2})$.

We next establish the sharper self-influence rate. For each i, let $R_i \in [m]$ denote the rank of Z_i in the full sample. For the g-th routing boundary, define $k_g^{(-)} := \lceil (m-1)u_g \rceil$.

Since $k_g - k_g^{(-)} \in \{0, 1\}$, removing one observation can move the rank defining the empirical quantile by at most one position.

More precisely,

$$Z_i \le \widehat{q}_g \quad \Longleftrightarrow \quad R_i \le k_g.$$

If the i-th observation is removed and the resulting empirical quantile is evaluated against the removed value Z_i, then

$$Z_i \le \widehat{q}_g^{(-i)} \quad \Longleftrightarrow \quad R_i \le k_g^{(-)}.$$

Consequently,

$$\mathbf{1}\left\{\mathbf{1}\{Z_i \le \widehat{q}_g\} \ne \mathbf{1}\{Z_i \le \widehat{q}_g^{(-i)}\}\right\} \le \mathbf{1}\left\{k_g^{(-)} < R_i \le k_g\right\}.$$

Because $k_g - k_g^{(-)} \le 1$,

$$\sum_{i=1}^{m} \mathbf{1}\left\{\mathbf{1}\{Z_i \le \widehat{q}_g\} \ne \mathbf{1}\{Z_i \le \widehat{q}_g^{(-i)}\}\right\} \le 1$$

almost surely.

A routing label can change only if the observation changes sides of at least one of the $M-1$ routing boundaries. Thus,

$$\sum_{i=1}^{m} \mathbf{1}\left\{\widehat{\rho}(X_i) \ne \widehat{\rho}^{(-i)}(X_i)\right\} \le M-1$$

almost surely. Taking expectations and dividing by m gives

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{P}\left\{\widehat{\rho}(X_i) \ne \widehat{\rho}^{(-i)}(X_i)\right\} \le \frac{M-1}{m}.$$

Since M is fixed, $\zeta_m = O(m^{-1})$.

This proves the result. □

C.5.3 Proof of Proposition 3

For clarity, write $\rho_\theta(x) := \arg\max_{g\in[M]} s_g(x;\theta)$.

Let

$$\widehat{Q}_m(\theta) := \frac{1}{m}\sum_{j=1}^{m} \phi(\theta; X_j) + \frac{\lambda}{2}\|\theta - \theta_0\|_2^2,$$

and let $\widehat{\theta}$ minimize $\widehat{Q}_m$. Define the leave-one-out objective by

$$\widehat{Q}_m^{(-i)}(\theta) := \frac{1}{m-1}\sum_{j\ne i} \phi(\theta; X_j) + \frac{\lambda}{2}\|\theta - \theta_0\|_2^2,$$

with minimizer $\widehat{\theta}^{(-i)}$.

We use the following slightly more explicit version of the conditions in Proposition 3. Suppose that $\widehat{Q}_m$ and $\widehat{Q}_m^{(-i)}$ are uniformly μ-strongly convex, that $\|\nabla_\theta \phi(\theta; x)\| \le B$, and that $|s_g(x;\theta) - s_g(x;\theta')| \le L\|\theta - \theta'\|$ uniformly over g, x, θ, θ'. We further assume that $\mathbb{P}\{\Delta_\theta(X) \le u\} \le C_\Delta u$ uniformly over the parameter region attained by $\widehat{\theta}$, $\widehat{\theta}^{(-i)}$, and $\theta^\star$.

Proof. We first control the population routing discrepancy. Suppose $\rho_{\widehat{\theta}}(X) \ne \rho_{\theta^\star}(X)$.

Let $g^\star = \rho_{\theta^\star}(X)$. Since each routing score is L-Lipschitz in θ, every score changes by at most $L\|\widehat{\theta} - \theta^\star\|$.

Hence the identity of the largest score can change only if the population routing margin satisfies

$$\Delta_{\theta^\star}(X) \le 2L\|\widehat{\theta} - \theta^\star\|.$$

Conditional on the test sample, $\widehat{\theta}$ is fixed and X is a fresh observation. The margin condition therefore implies

$$d_P\left(\rho_{\widehat{\theta}}, \rho_{\theta^\star}\right) \le C\|\widehat{\theta} - \theta^\star\|.$$

Taking expectations and using

$$\mathbb{E}\|\widehat{\theta} - \theta^\star\| = O(m^{-1/2})$$

gives

$$\mathbb{P}\left\{\widehat{\rho}(X) \neq \rho^\star(X)\right\} = O(m^{-1/2}).$$

We next establish leave-one-out stability of the parameter estimate. By strong convexity,

$$\mu \left\|\widehat{\theta} - \widehat{\theta}^{(-i)}\right\| \leq \left\|\nabla \widehat{Q}_m\left(\widehat{\theta}^{(-i)}\right) - \nabla \widehat{Q}_m(\widehat{\theta})\right\|.$$

Since $\nabla \widehat{Q}_m(\widehat{\theta}) = 0$ and $\nabla \widehat{Q}_m^{(-i)}(\widehat{\theta}^{(-i)}) = 0$, the right-hand side is bounded by

$$\begin{aligned}
&\left\|\nabla \widehat{Q}_m(\widehat{\theta}^{(-i)}) - \nabla \widehat{Q}_m^{(-i)}(\widehat{\theta}^{(-i)})\right\| \\
&= \left\|\frac{1}{m}\nabla\phi(\widehat{\theta}^{(-i)}; X_i) - \frac{1}{m(m-1)}\sum_{j\neq i}\nabla\phi(\widehat{\theta}^{(-i)}; X_j)\right\| \\
&\leq \frac{2B}{m}.
\end{aligned}$$

Therefore, $\left\|\widehat{\theta} - \widehat{\theta}^{(-i)}\right\| \leq \frac{2B}{\mu m}$.

This also implies that the leave-one-out estimator has the same population rate:

$$\begin{aligned}
\mathbb{E}\left\|\widehat{\theta}^{(-i)} - \theta^\star\right\| &\leq \mathbb{E}\left\|\widehat{\theta} - \theta^\star\right\| + \mathbb{E}\left\|\widehat{\theta} - \widehat{\theta}^{(-i)}\right\| \\
&= O(m^{-1/2}).
\end{aligned}$$

Repeating the preceding population-disagreement argument with $\widehat{\theta}^{(-i)}$ therefore gives

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{P}\left\{\widehat{\rho}^{(-i)}(X) \neq \rho^\star(X)\right\} = O(m^{-1/2}).$$

Thus the first part of Assumption 1 holds with $\gamma_m = O(m^{-1/2})$.

Finally, consider the routing assignment of X_i itself. If $\rho_{\widehat{\theta}}(X_i) \neq \rho_{\widehat{\theta}^{(-i)}}(X_i)$, then the largest routing score under $\widehat{\theta}^{(-i)}$ must be within $2L\left\|\widehat{\theta} - \widehat{\theta}^{(-i)}\right\|$ of the second-largest score. Hence $\Delta_{\widehat{\theta}^{(-i)}}(X_i) \leq \frac{4LB}{\mu m}$, where we used (C.5.3).

Conditional on the leave-one-out test sample, $\widehat{\theta}^{(-i)}$ is fixed and X_i is independent of it. Therefore, the uniform margin condition gives

$$\begin{aligned}
&\mathbb{P}\left\{\widehat{\rho}(X_i) \neq \widehat{\rho}^{(-i)}(X_i)\,\middle|\,\mathcal{F}_{-i}\right\} \\
&\qquad \leq \mathbb{P}\left\{\Delta_{\widehat{\theta}^{(-i)}}(X_i) \leq \frac{4LB}{\mu m}\,\middle|\,\mathcal{F}_{-i}\right\} \leq \frac{C}{m}.
\end{aligned}$$

Taking expectations and averaging over i yields

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{P}\left\{\widehat{\rho}(X_i) \neq \widehat{\rho}^{(-i)}(X_i)\right\} = O(m^{-1}).$$

Therefore, $\zeta_m = O(m^{-1})$, which completes the proof. □

D Additional Experimental Results

This appendix complements the main experiments with seven studies that examine PACE beyond the primary 85% comparisons. First, we compare PACE with the full set of baselines on MT-Bench at the stricter 90% target, testing the framework at a more demanding operating point (Appendix D.1).

Second, we detail the learned PACE-BGE gate and evaluate whether PACE extends beyond scalar-quantile routing to learned multiclass group assignments (Appendix D.2). Third, we replace the first two Judges to test whether frozen routing signals remain useful under a different Judge panel (Appendix D.3). Fourth, we vary the number of labeled calibration samples to examine finite-sample stability (Appendix D.4). Fifth, we combine pre-routing and cascading with alternative reliability backends to separate their contributions to cost and selective utility (Appendix D.5). Sixth, we apply the same Prompt-v1 router unchanged across all three datasets to examine cross-dataset behavior without dataset-specific router selection (Appendix D.6). Finally, we sweep the target under multiple group allocations to determine whether the routing-allocation trade-off observed in the main text persists across operating points (Appendix D.7). Together, these experiments evaluate the robustness of PACE to the target level, router family, Judge-panel composition, calibration-set size, dataset, and group allocation, while isolating the roles of one-shot routing and pooled risk selection.

D.1 MT-Bench Baseline Comparison

As a supplementary stress test beyond the primary 85% operating point, we compare PACE and other baselines on MT-Bench (Zheng et al., 2023) at stricter target $1 - \alpha = 0.90$.

Table 3: Baseline comparison on MT-Bench with target agreement $1 - \alpha = 0.90$.

Method	**Agreement**	**Coverage**	**Cost**	**Judge records per test sample**			
				Mistral	Qwen	Luna	Total
No selection (Luna)	82.618	100.000	8.884	0.000	0.000	1.000	1.000
PoLL-style probability pooling	83.390	100.000	15.036	1.000	1.000	1.000	3.000
SCOPE-BPE (Mistral)	88.763	23.322	0.384	1.000	0.000	0.000	1.000
SCOPE-BPE (Qwen14B)	89.915	61.299	5.768	0.000	1.000	0.000	1.000
SCOPE-BPE (Luna)	89.015	49.856	8.884	0.000	0.000	1.000	1.000
Three-Judge cascade	90.918	61.800	11.546	1.000	0.936	0.649	2.585
Cascaded heuristic selection	87.211	85.232	7.911	1.000	0.818	0.316	2.134
PACE (ours)							
Prompt-v2 (quantile)	**89.450**	**55.961**	**5.021**	**0.333**	**0.333**	**0.333**	**1.000**
Prompt-v3 (quantile)	**89.844**	**53.547**	**5.022**	**0.333**	**0.333**	**0.333**	**1.000**
PACE-BGE (multiclass)	**90.166**	**60.197**	**6.584**	**0.156**	**0.316**	**0.528**	**1.000**

At this stricter operating point, non-selective evaluation and heuristic selection fall well short of the 90% agreement target. All fixed-Judge SCOPE variants also miss it, including SCOPE with the strongest Judge (Luna). Under their respective evaluation protocols, only the Three-Judge cascade and PACE-BGE meet the target, while the two lower-cost quantile routers fall slightly below it. Because PACE-BGE reserves 30% of each dataset for router development, Appendix D.2 provides a direct comparison in which PACE-BGE and representative baselines use exactly the same held-out evaluation pool and repeated splits.

D.2 Learned BGE Multiclass Gate

Beyond quantile routing, we ask whether PACE also applies to learned multiclass groups (Section 5.4). We treat PACE-BGE as a pretrained router: its gate is fitted on a disjoint labeled development split and frozen before PACE calibration and testing. No calibration or test labels are used to fit, select, or adapt the gate.

The additional development samples may be reserved from the same underlying dataset; they need not come from another data source. What matters for this protocol is separation from calibration and testing, with all gate fitting and selection completed before evaluation. PACE-BGE consumes human preferences and the three Judges' correctness outcomes on the development split, unlike the label-free prompt routers.

BGE-M3 encodes the request and the two candidate answers separately. We form an answer-order-invariant feature map from the request embedding, the mean and absolute difference of the answer embeddings, their elementwise product, symmetric request–answer differences, cosine similarities,

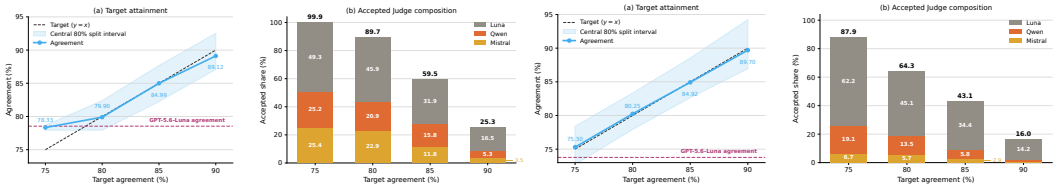

Figure 4: PACE-BGE learned multiclass gate on ChatArena (left two panels) and TL;DR (right two panels). Shaded bands show the central 80% of repeated-split agreement outcomes.

and text lengths. We retain the three binary correctness outcomes $Y_g \in \{0, 1\}$ and fit three logistic heads,

$$\widehat{p}_g(x) = \sigma\left(w_g^\top \Phi(x) + b_g\right), \qquad s_g(x) = \operatorname{logit} \widehat{p}_g(x) - \eta \widetilde{C}_g, \qquad \rho(x) = \arg\max_{g \in \{1,2,3\}} s_g(x),$$

where $\widetilde{C}_g$ is normalized Judge cost. PCA dimension, logistic regularization, and η are selected on router validation by routed accuracy minus 0.05 times normalized cost; the transform and heads are then refitted on router train plus validation and frozen. For each dataset, a seed-42 split reserves 20% for router training, 10% for router validation, and 70% as the PACE evaluation pool. Repeated PACE evaluation uses 500 calibration samples on ChatArena and TL;DR and 300 on MT-Bench, with the remainder serving as test samples in each of 1,000 splits.

Figure 4 tests whether learned multiclass routing preserves the target-responsive behavior observed with quantile grouping. On both datasets, higher targets raise empirical agreement while reducing coverage, and the accepted set shifts toward Luna. The agreement curves closely track the target line. Together with the MT-Bench result, this supports the effectiveness of PACE's risk-estimation and pooled-selection stages beyond scalar-quantile routing.

Matched-pool baseline comparison. PACE-BGE is evaluated only on the held-out 70% pool left after router development, whereas the main baseline tables use the full datasets. To separate method differences from evaluation-pool composition, we rerun two representative baselines—SCOPE-BPE with the strongest single Judge (Luna) and the Three-Judge cascade—on exactly the same evaluation IDs. Within each of 1,000 runs, all methods share the same calibration and test indices: 500 calibration samples on ChatArena and TL;DR, and 300 on the smaller MT-Bench pool.

Table 4: Matched-pool comparison on ChatArena at target agreement $1 - \alpha = 0.85$.

Method	**Agreement**	**Coverage**	**Cost**	**Judge records per test sample**			
				Mistral	Qwen	Luna	Total
SCOPE-BPE (Luna)	84.287	57.078	8.884	0.000	0.000	1.000	1.000
Three-Judge cascade	86.220	55.112	10.962	1.000	0.870	0.626	2.496
PACE-BGE (multiclass)	**84.995**	**59.473**	**5.950**	**0.254**	**0.252**	**0.494**	**1.000**

Table 5: Matched-pool comparison on TL;DR at target agreement $1 - \alpha = 0.85$.

Method	**Agreement**	**Coverage**	**Cost**	**Judge records per test sample**			
				Mistral	Qwen	Luna	Total
SCOPE-BPE (Luna)	83.495	34.318	8.884	0.000	0.000	1.000	1.000
Three-Judge cascade	86.776	42.160	13.186	1.000	0.936	0.833	2.769
PACE-BGE (multiclass)	**84.923**	**43.107**	**7.579**	**0.075**	**0.218**	**0.707**	**1.000**

Table 6: Matched-pool comparison on MT-Bench at target agreement $1 - \alpha = 0.90$.

Method	Agreement	Coverage	Cost	Judge records per test sample			
				Mistral	Qwen	Luna	Total
SCOPE-BPE (Luna)	88.607	41.782	8.884	0.000	0.000	1.000	1.000
Three-Judge cascade	90.995	60.636	11.337	1.000	0.886	0.658	2.544
PACE-BGE (multiclass)	**90.166**	**60.197**	**6.584**	**0.156**	**0.316**	**0.528**	**1.000**

On ChatArena and TL;DR, PACE-BGE comes within 0.08 percentage points of the 85% target, retains the highest coverage among these methods, and costs less than both SCOPE-Luna and the cascade. On MT-Bench, it attains the stricter 90% target with coverage nearly identical to the cascade, while using one Judge record instead of 2.544 and reducing online cost by 42%. These matched-pool results show that PACE-BGE's favorable agreement–coverage–cost trade-off is not an artifact of evaluating it on a different sample pool.

D.3 Judge-Panel Replacement Robustness

To test whether PACE depends on the original Mistral–Qwen14B–Luna panel, we replace its first two tiers with Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Gemma-2-9B-IT (Team et al., 2024), retaining GPT-5.6-Luna as the strongest tier. We reuse the frozen Prompt-v1/v2/v3, Length, TF–IDF, and BGE-M3 scores without adapting the routers to labels or outputs from the replacement panel.

Table 7: Grouping-strategy robustness on ChatArena after replacing the Judge panel with Llama-3.1-8B, Gemma-2-9B, and GPT-5.6-Luna.

Method	Target 75%		Target 80%		Target 85%		Target 90%	
	Agreement	Coverage	Agreement	Coverage	Agreement	Coverage	Agreement	Coverage
Random tertiles	76.674	98.672	80.408	77.290	85.786	48.747	89.400	23.063
Prompt-v1	77.872	99.726	80.324	85.969	85.791	58.749	90.283	34.064
Prompt-v2	77.516	99.559	80.374	83.557	85.991	56.210	90.404	31.100
Prompt-v3	77.555	99.591	80.390	84.025	85.765	56.144	90.462	31.605
Length	75.922	95.965	80.152	68.734	85.070	40.963	88.429	17.223
TF–IDF answer similarity	75.918	96.754	80.671	71.090	85.589	41.677	87.953	15.544
BGE-M3 kNN density/OOD	76.877	98.906	80.282	78.733	85.704	50.136	89.518	25.174

After panel replacement, the prompt routers continue to meet every target and retain the strongest coverage at most targets, whereas some alternative routers miss the 90% target. This supports panel portability: PACE remains effective after replacing two Judges without retraining the frozen routing scores on Judge outcomes or human labels.

D.4 Calibration-Set Size Sensitivity

The main experiments use 500 labeled calibration samples, but the amount of labeled data available in practice may vary. We examine whether increasing the calibration-set size stabilizes PACE's finite-sample behavior and brings its agreement closer to the target without collapsing coverage. We evaluate Prompt-v2 with equal-tertile routing on ChatArena at both the 85% target and the stricter 90% target. The stricter target leaves a smaller error budget and motivates examining a larger calibration budget for more precise local-FDR estimation.

Both experiments use 1,000 repeated random runs on the 5,172 ChatArena samples. At the 85% target, each run reserves 1,000 samples as a calibration reservoir and uses the remaining 4,172 as a fixed test set; calibration sizes are $n \in \{400, 500, 600, 700, 800, 900, 1000\}$. At the 90% target, each run instead reserves 2,000 calibration candidates and uses the remaining 3,172 as a fixed test set; calibration sizes are $n \in \{500, 750, 1000, 1250, 1500, 1750, 2000\}$. Within either experiment, the calibration sets are nested prefixes of the same reservoir and share the test-derived routing thresholds and test set. The two targets therefore use different held-out test sets and are not paired comparisons across targets.

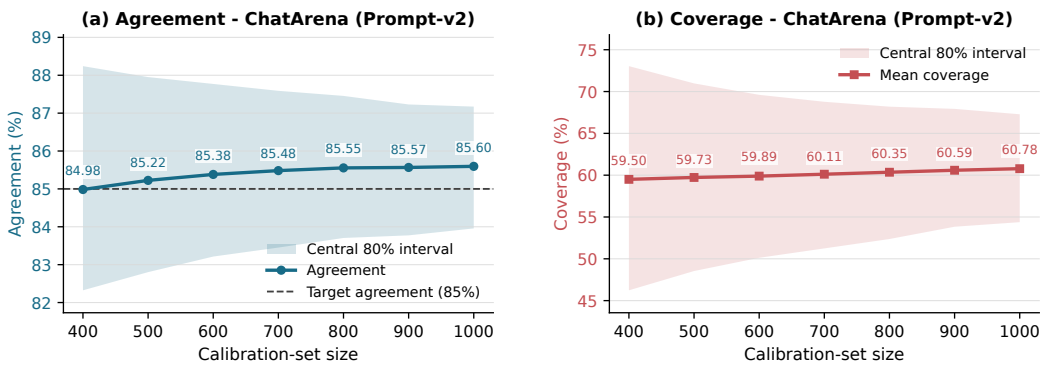

Figure 5: Calibration-set size sensitivity for Prompt-v2 on ChatArena at the 85% target. Shaded bands show the central 80% of the split outcomes.

At the 85% target, agreement increases from 84.98% at $n = 400$ to 85.60% at $n = 1000$, exceeding the target from $n = 500$ onward (Figure 5). The splitwise interval narrows as calibration size increases, while mean coverage remains near 60%.

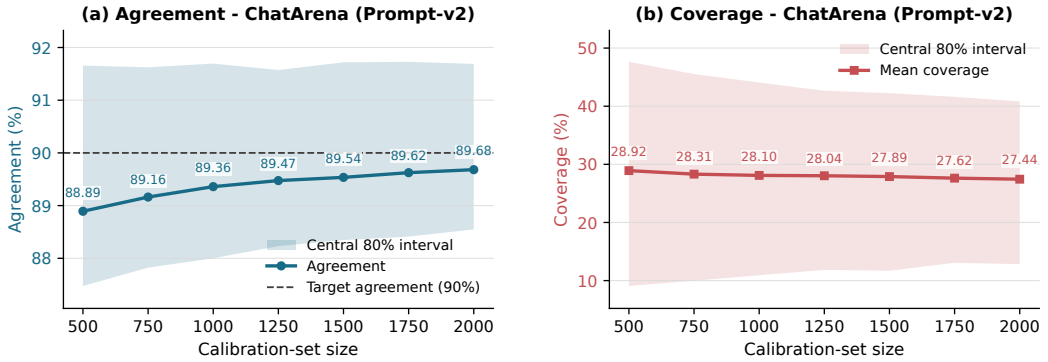

Figure 6: Calibration-set size sensitivity for Prompt-v2 on ChatArena at the stricter 90% target. Shaded bands show the central 80% of the split outcomes.

At the stricter 90% target, agreement gradually approaches the target as calibration size grows, increasing from 88.89% at $n = 500$ to 89.68% at $n = 2000$ (Figure 6). Mean coverage changes from 28.92% to 27.44%. The target is not attained at the tested sizes; the observed finite-sample trend does not by itself establish asymptotic convergence or require every split to satisfy the target.

D.5 Pre-Routing and Cascade Hybrids

The main comparison changes both the Judge-assignment mechanism and the reliability backend: PACE pre-routes each sample and applies pooled local-FDR selection, whereas the native cascade invokes Judges sequentially and uses stagewise binomial upper-bound thresholds. To examine the contribution of each design choice, we construct two targeted hybrids on ChatArena at 85% target as the main experiment.

Pre-routed cascade retains Prompt-v2 equal-tertile routing but uses it only to choose the starting Judge. Samples routed to Mistral may escalate to Qwen and Luna, samples routed to Qwen may escalate to Luna, and Luna-routed samples start at Luna. At each tier, the confidence threshold is calibrated with the native cascade's binomial upper-bound rule using calibration samples that start at or reach that tier; a test sample accepts the first reached Judge that passes its threshold and otherwise abstains after Luna. This hybrid compares stagewise cascade certification with pooled local-FDR selection under the same initial pre-routing.

Cascade-PACE instead retains PACE's local-FDR estimator and pooled selection but replaces Prompt-v2 routing with a sequential confidence gate. Every sample first calls Mistral; the highest-confidence test tertile stops there. The remainder calls Qwen, whose highest-confidence half stops at Qwen, while the final third reaches Luna. The same test-derived thresholds assign calibration samples to their final tiers, after which local FDR is estimated within each induced group and pooled exactly as in PACE. This hybrid compares pre-routing with confidence-gated cascading under the same risk-estimation and selection backend.

Table 8: Routing–selection hybrids on ChatArena at 85% target.

Method	Initial assignment	Reliability backend	Agreement	Coverage	Cost	Calls
PACE (Prompt-v2)	**Prompt-v2 pre-routing**	**Local FDR + pooled selection**	**85.233**	**59.764**	**5.020**	**1.000**
Pre-routed cascade	Prompt-v2 pre-routing	Cascade upper bounds	88.327	51.991	10.183	1.676
Cascade-PACE	Confidence-gated cascade	Local FDR + pooled selection	82.753	54.801	7.189	2.000
Three-Judge cascade	Confidence-gated cascade	Cascade upper bounds	86.603	54.993	10.992	2.501

Table 8 separates two effects. With Prompt-v2 as the initial routing signal, the pre-routed cascade exceeds the target but retains fewer samples and costs roughly twice as much as PACE. Its stagewise upper-bound thresholds are more conservative at this nominal target than pooled risk budgeting, which can reallocate risk allowance across routed groups. With the local-FDR backend fixed, Cascade-PACE uses twice as many Judge records and costs 43% more than PACE, yet misses the target and retains fewer samples. In this setting, sequential raw-confidence gating produces groups with a less effective finite-sample risk ranking than the frozen text-based pre-router.

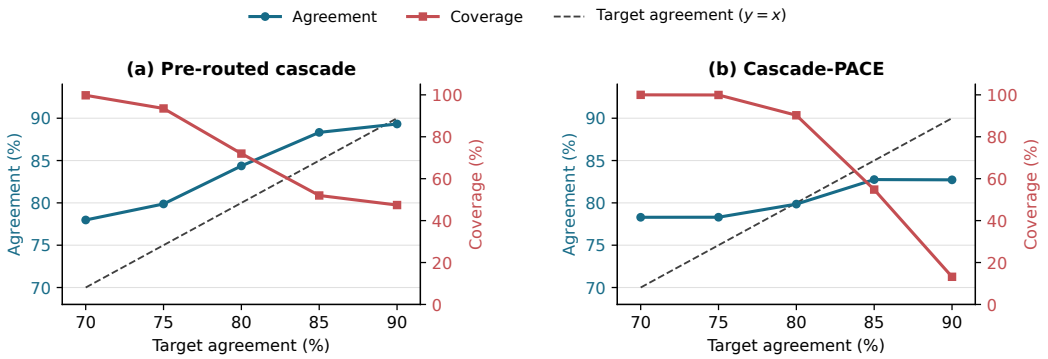

Figure 7: Target sweep for the two routing–selection hybrids on ChatArena. Dashed lines mark target agreement ($y = x$).

Figure 7 shows that the pre-routed cascade initially raises agreement as the target tightens, but approaches a plateau near the highest target while coverage continues to fall. Cascade-PACE's agreement levels off earlier, and its coverage drops sharply at the most demanding target. Neither reaches 90% agreement at the 90% target. Together with Table 8, these trends support complementary roles for PACE's design choices: pre-routing avoids repeated Judge calls and gives the local-FDR estimator a more useful risk ranking than sequential raw-confidence gating in this setting, while pooled selection makes less conservative use of the risk budget than stagewise thresholds at the 85% target.

D.6 Cross-Dataset Evaluation

The main experiments use dataset-specific operating points. To test whether the framework's behavior depends on selecting a new router for each dataset, we fix Prompt-v1 and evaluate it unchanged on ChatArena, TL;DR, and MT-Bench over target agreements from 75% to 90%.

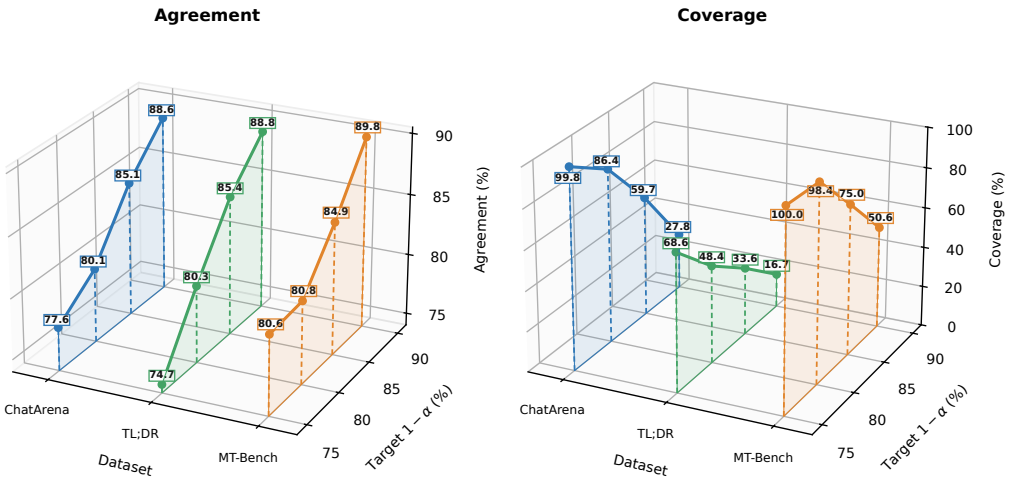

Figure 8: Cross-dataset PACE results with the same Prompt-v1 router.

Figure 8 shows the expected agreement–coverage trade-off on all three datasets: tightening the target generally increases agreement while reducing the accepted set. Thus, the same frozen router and selection pipeline provide usable operating points without dataset-specific adaptation, while the differing coverage levels make clear that the attainable trade-off remains dataset dependent.

D.7 Group-Allocation Target Sweep

The main text compares Prompt-v2 allocations at the common ChatArena 85% operating point. To test whether that comparison is specific to a single target, we evaluate the same three prespecified allocations over target agreements $1 - \alpha \in \{0.75, 0.80, 0.85, 0.90\}$.

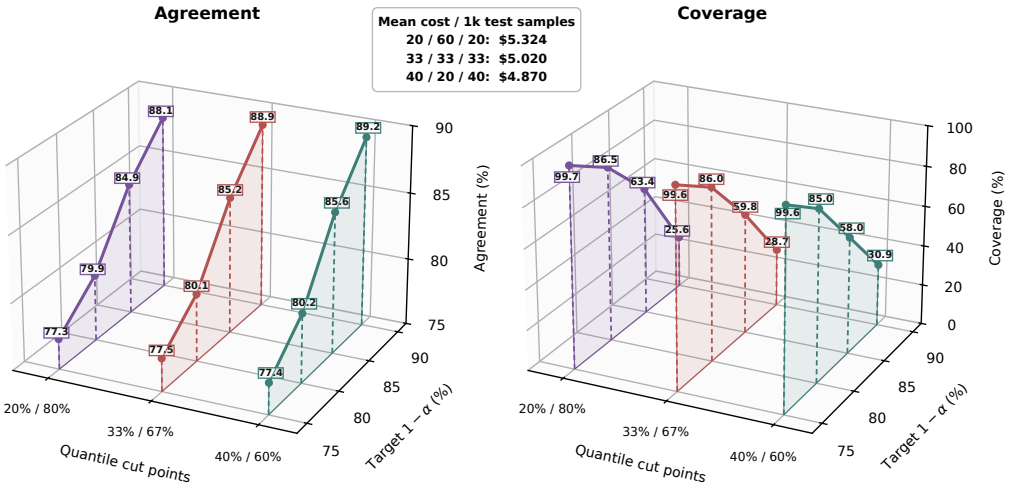

Figure 9: Quantile-boundary sensitivity for Prompt-v2 on ChatArena across target agreement levels.

Figure 9 shows the expected monotone target trade-off under all three allocations: tightening the target raises agreement and reduces coverage. Their relative differences persist over the sweep, indicating that the main-text allocation comparison reflects a stable deployment trade-off rather than an artifact of the 85% target.

E Experimental Implementation Details

E.1 PACE Implementation Details

Within routing group g, we use the normalized Gaussian kernel

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \tag{39}$$

and bandwidth

$$h_g = \max\{0.1 n_g^{-1/5}, h_{g,\text{Silverman}}\}, \qquad h_{g,\text{Silverman}} = 1.8 \min\left(\sigma_g, \frac{\text{IQR}_g}{1.349}\right) n_g^{-1/5}, \tag{40}$$

where n_g is the calibration-group size and σ_g and IQR_g are the dispersion estimates used by the kernel estimator in group g. The factor 1.8 is an empirical coefficient in this Silverman-type bandwidth rule.

For numerical degeneracy, the scale is the smaller strictly positive value among σ_g and $\text{IQR}_g/1.349$ (values below 10^{-12} are ignored); if neither is positive, the scale is set to one. The lower safeguard decreases as $n_g^{-1/5}$, so it does not obstruct the bandwidth rate assumed in the theory when group sample sizes are proportional to n and the population scale is nondegenerate. The theorem's distributional and stability assumptions remain necessary.

Details of Grouping Strategies. Prompt-v1, Prompt-v2, and Prompt-v3 use a frozen Qwen2.5-3B router to score overall comparison difficulty, Judge failure risk, and the minimum required Judge tier, respectively. Each produces a five-level distribution and the continuous score $Z = \sum_{k=1}^{5} k p_k$. Length uses $\log(1+x)$, where x is the total number of Unicode characters in the instruction and both answers. BGE-M3 kNN density/OOD uses the mean cosine distance to a sample's 20 nearest neighbors in the BGE-M3 embedding space (Chen et al., 2024).

E.2 Baseline Implementation Details

Shared Judge-inference protocol. One Judge record is constructed by multiple internal inference steps. All methods reuse the same cache generated with $N = 3$ simulated annotators, each conditioned on $K = 3$ labeled demonstrations. Each simulated annotator receives both candidate orders, (A, B) and (B, A). The second output is first mapped back to the original order, after which the class probabilities are averaged across directions. This procedure corresponds to the balanced-position calibration of FairEval (Wang et al., 2024) and, for the present binary Judge outputs without ties, is also consistent with the likelihood-aware bidirectional aggregation of TrustJudge (Wang et al., 2026). The internal simulated-annotator and bidirectional inference steps are included in the cost assigned to that record. Except for the multi-Judge aggregation and cascade methods explicitly listed in the tables, each sample receives exactly one Judge evaluation record. Lightweight router inference is accounted for separately in the reported cost and does not produce a formal verdict. Appendix E.4 gives the exact prompt, probability aggregation, and cached-record format.

Non-selective Baselines. No selection (Luna) accepts all Luna outputs. PoLL-style pooling takes the arithmetic mean of the full class probabilities from the three Judges and accepts every pooled prediction (Verga et al., 2024). Both methods therefore have 100% coverage independent of the target agreement.

Selective Baselines. For each named Judge, SCOPE-BPE computes binary preference entropy from its bidirectionally averaged preference probabilities and uses SCOPE's finite-sample linear constraint to choose the largest feasible entropy threshold (Badshah et al., 2026). The Three-Judge cascade follows the publicly released Trust or Escalate implementation (Jung et al., 2025) in the order Mistral, Qwen2.5-14B, GPT-5.6-Luna, and abstain; cascaded heuristic selection follows the same path using raw-confidence thresholds. We set the cascade's $\delta = 0.1$, as in the released code's usage example. The implementation passes this full value to the binomial upper-bound calculation at each tier, rather than allocating $\delta/3$ across tiers. Each tier calibrates its confidence threshold on calibration samples not accepted by earlier tiers, searching 5,000 candidates in $[0, 1]$ and retaining only candidates that accept at least 30 calibration samples. If even the highest retained candidate fails the

upper-bound check at the requested α, the released implementation returns that candidate instead of forcing the tier to abstain. We retain this threshold-selection behavior; repeated-split evaluation on cached Judge records uses a vectorized reproduction of the released rule. The same δ and rule are used for the pre-routed cascade in Appendix D.5, with calibration restricted to samples that start at or reach each tier.

E.3 Cost Accounting

All costs are reported as test-time online inference cost in monetary-equivalent USD per 1,000 test samples. The metric includes the formal Judge records and any router inference performed on the test workload, but excludes the calibration set, which is treated as an offline resource prepared before deployment. For local model j, let T_j^{GPU} be the GPU-seconds aggregated over all participating devices, N_j the number of samples actually attempted in the timing run, and r_j the on-demand hourly price of a cloud GPU of the same type. Then

$$c_{j,1000}^{\mathrm{local}} = 1000 \times \frac{T_j^{\mathrm{GPU}}}{3600 N_j} r_j. \tag{41}$$

We use GPU-seconds rather than wall-clock time alone: if an inference occupies two GPUs simultaneously, T_j^{GPU} is twice the wall-clock duration in seconds. For a synchronous API model with a bill of B_j for the full workload,

$$c_{j,1000}^{\mathrm{API}} = 1000 \times \frac{B_j}{N_j}. \tag{42}$$

Mistral-7B processes 5,178 actual attempts on one V100 in 43 minutes 17 seconds, corresponding to 0.721 GPU-hours. Qwen2.5-14B takes 5 hours 25 minutes 22 seconds on two V100s, corresponding to 10.846 aggregate GPU-hours. The main estimate uses the Linux on-demand price of 22.032 per hour for Azure `Standard_ND40rs_v2`; the instance contains eight 32GB V100s, yielding 2.754 per GPU-hour. Excluding model loading and CUDA graph capture but retaining the actual inference pipeline gives costs of 0.384 and 5.768 per 1,000 samples for Mistral and Qwen14B, respectively. GPT-5.6-Luna is billed at 46 for the same complete workload under synchronous API deployment, or 8.884 per 1,000 samples.

Router costs are accounted for separately using the same formula. Qwen2.5-3B and the single-embedding BGE-M3 router process all 5,172 samples on an RTX 4090. Timing excludes model loading but includes tokenization, host-to-device transfer, forward computation, and score postprocessing. Prompt-v1/v2/v3 cost 0.00827, 0.00946, and 0.01025 per 1,000 samples, respectively; single-embedding BGE-M3 encoding and score postprocessing cost 0.000996 per 1,000 samples. PACE-BGE instead encodes the request and two answers separately before applying its frozen PCA and linear heads. The complete path processes the 5,172 ChatArena samples in 84.621 seconds on one V100; using the same V100 replacement price of 2.754 per GPU-hour gives 0.012516 per 1,000 samples. We use this ChatArena-measured routing charge for PACE-BGE on all datasets. Each tabulated method cost is the sum of its test-time formal Judge cost and online routing cost; offline router fitting and calibration construction are excluded from the online-cost column.

E.4 Pairwise Judge Prompt and Cached-Record Construction

Prompt template. The main experimental caches use the $N = 3, K = 3$ simulated-annotator protocol. Figure 10 reproduces the complete prompt structure used for each simulated annotator: the system message and task description are followed by three labeled demonstrations and an unlabeled target query whose preferred-response field is left blank.

Figure 11 shows one complete labeled demonstration from the fixed $N = 3, K = 3$ few-shot set used to construct the Judge prompts. The released code repository includes all nine demonstrations, grouped by simulated annotator.

The target sample's human preference is never included in its query. The Judge is evaluated once in the original answer order and once after swapping the answers. Only outputs that can be parsed as `[[A]]` or `[[B]]` and for which both directions are available are retained for that simulated annotator.

Prompt for Pairwise Judge Evaluation

[System]
You are a helpful assistant.

[Task]
Given an instruction and two assistant's responses to the instruction, an annotator chose which assistant's answer is more preferred. Given examples of the annotator's decision, predict the annotators' verdict on the given example. If Assistant A's response is more preferred than Assistant B's, the annotator chose "[[A]]". If Assistant B's response is more preferred than Assistant A's, the annotator chose "[[B]]".

[Labeled demonstration; repeated K = 3 times]

[Instruction]	`<instruction_example_k>`
[Assistant A's response]	`<answer_a_example_k>`
[Assistant B's response]	`<answer_b_example_k>`
[Preferred response]	`<[[A]] or [[B]]>`
	... remaining labeled demonstrations ...

[Unlabeled target query]

[Instruction]	`<instruction>`
[Assistant A's response]	`<answer_a>`
[Assistant B's response]	`<answer_b>`
[Preferred response]	`<blank; Judge returns [[A]] or [[B]]>`

Figure 10: Pairwise Judge prompt template. Each of the $N = 3$ simulated annotators receives $K = 3$ labeled demonstrations before evaluating the unlabeled target query.

An Actual Labeled Demonstration

Annotator 3, demonstration 2 of 3

[Instruction]	classify the named entities in this text: George Washington and his troops crossed the Delaware River on December 25, 1776 during the American Revolutionary War.
[Assistant A's response]	Person: George Washington Organization: troops Date: December 25, 1776 Action: crossed the Delaware River
[Assistant B's response]	The named entities in this text are George Washington, his troops, the Delaware River, December 25, 1776, and the American Revolutionary War.
[Preferred response]	[[A]]

Figure 11: An actual labeled ChatArena demonstration supplied to the third simulated annotator (its second example). The preferred response is Assistant A.

Probability aggregation and record format. Let $p^{\rightarrow}_{i,n}(A)$ and $p^{\rightarrow}_{i,n}(B)$ be the token probabilities for simulated annotator n in the original order, and let $p^{\leftarrow}_{i,n}(A)$ and $p^{\leftarrow}_{i,n}(B)$ denote the probabilities from the swapped query after mapping the labels back to the original order. For the valid paired annotators $\mathcal{V}_i$, the stored probabilities are

$$\tilde{p}_i(y) = \frac{1}{2|\mathcal{V}_i|} \sum_{n \in \mathcal{V}_i} \left[p^{\rightarrow}_{i,n}(y) + p^{\leftarrow}_{i,n}(y) \right], \qquad y \in \{A, B\}. \tag{43}$$

The predicted verdict is $\arg\max_y \tilde{p}_i(y)$, and the raw confidence used by the main experiments is $C_i^{\text{raw}} = \max_y \tilde{p}_i(y)$. Because top-token probabilities need not exhaust unit mass, the cache also stores the A/B-normalized probabilities $\tilde{p}_i(y)/(\tilde{p}_i(A) + \tilde{p}_i(B))$ and the corresponding normalized confidence. Each released text-free record contains the sample identifier, human label, ordered Judge names, per-Judge probabilities/prediction/confidence/error, and frozen routing scores. Source instructions and candidate responses are omitted from the public table-reproduction cache and can be reconstructed from the cited source datasets.

E.5 Qualitative Examples

The following examples illustrate three possible outcomes of the pipeline. They come from the first ChatArena split generated with seed 42, Prompt-v2, calibration size 500,and target agreement 85%. Text is shortened only for presentation; routing and selection use the complete records. These examples illustrate behavior and are not used as performance evidence.

Low-tier, accepted (`chatarena_000056`). The request asks for the square root of 1,000,000. Answer A gives 1, whereas Answer B gives 1,000. Prompt-v2 assigns score 3.043, placing the sample in the low group and routing it to Mistral. Mistral predicts B, agreeing with the human label, with raw confidence 0.9986. The estimated local FDR is 0.0660, so the verdict is accepted.

High-tier, accepted (`chatarena_001136`). The request asks for a command-line expression that searches shell history for "his." Answer A proposes `history | grep his`, whereas Answer B proposes a nonstandard command. Prompt-v2 assigns score 4.969, placing the sample in the high group and routing it to Luna. Luna predicts A, agreeing with the human label, with raw confidence 1.0000. The estimated local FDR is 0.1048, so the verdict remains within the pooled risk budget and is accepted.

High-tier, abstained (`chatarena_001625`). The request poses the three-ducks riddle. Answer A gives three ducks, whereas Answer B gives four. Prompt-v2 assigns score 4.891 and routes the sample to Luna. Luna predicts A, which agrees with the human label, but its raw confidence is only 0.7187 and the estimated local FDR is 0.3037. The procedure therefore abstains. This example also shows that selective control can reject an individually correct verdict when its estimated risk is too high.

E.6 Zero-Shot Routing Prompts

All three versions use a frozen Qwen2.5-3B-Instruct model and access neither human labels, Judge outputs, nor Judge correctness. The model computes probabilities only over the next-token set $\{1, 2, 3, 4, 5\}$ and uses $Z = \sum_{k=1}^{5} k\, p_k$ as the continuous routing score. The complete system prompts used in the experimental code follow.

Prompt-v1: Comparison Difficulty

[System]

You estimate the difficulty of a pairwise LLM-as-a-Judge task.
Your task is NOT to decide whether Answer A or Answer B is better. Estimate how difficult it would be for another language-model Judge to reliably agree with a careful human annotator about which answer better satisfies the user request.

Use exactly one of these ordered levels:
1 = Very easy: one answer is clearly superior or clearly broken; little factual checking, expertise, or multi-step reasoning is needed.
2 = Easy: the quality gap is fairly clear, with only light checking or reasoning.
3 = Moderate: the answers have competing strengths, or some factual/domain/reasoning checks are needed, but a careful comparison is still manageable.
4 = Hard: the answers are close, the prompt is ambiguous, or reliable comparison needs substantial expertise, factual verification, code/math analysis, or multi-step reasoning.
5 = Very hard: the difference is extremely subtle or uncertain and dependable human agreement itself may require deep expertise or external verification.

Consider answer similarity and trade-offs, factual-verification burden, specialized knowledge, reasoning depth, and ambiguity. Ignore superficial length or style differences unless they materially affect answer quality.

Return only the single digit 1, 2, 3, 4, or 5.

Figure 12: Complete system prompts for Prompt-v1 (comparison difficulty)

Prompt-v2: Judge Failure-Risk Rubric

[System]

You are a label-free failure-risk estimator for a pairwise LLM-as-a-Judge task.
Do NOT decide whether Answer A or Answer B is better. Estimate how likely a low- or medium-capability Judge is to disagree with a careful human when comparing the two answers.

Use the following fixed rubric. Consider all five factors jointly:
1. Decisiveness: is one answer plainly better, or are their strengths close?
2. Verification burden: are specialized facts, external knowledge, citations, calculations, or code behavior difficult to verify from the text alone?
3. Reasoning burden: is multi-step logical, mathematical, causal, or code reasoning required to compare the answers?
4. Instruction ambiguity: is the user request underspecified, subjective, or composed of several constraints whose importance is unclear?
5. Bias susceptibility: could verbosity, style, answer position, confident wording, or superficial fluency mislead a Judge?

Assign exactly one ordered risk level:
1 = Minimal risk: the comparison is decisive and none of the five factors is materially challenging.
2 = Low risk: the quality gap is fairly clear; only light checking is needed.
3 = Moderate risk: at least one factor is meaningful, but a medium Judge should usually compare the answers reliably.
4 = High risk: several factors are substantial, or the answers are close enough that a medium Judge may disagree with a careful human.
5 = Extreme risk: the comparison is exceptionally close, ambiguous, technical, verification-heavy, or strongly exposed to Judge bias.

Judge the difficulty of COMPARING the supplied answers, not the difficulty of writing a new answer. Do not treat length alone as evidence of quality or risk.
Return only the single digit 1, 2, 3, 4, or 5.

Figure 13: Complete system prompts for Prompt-v2 (Judge failure-risk rubric).

Prompt-v3: Required Judge Tier

[System]

You are a label-free routing estimator for a pairwise LLM-as-a-Judge task. Your task is NOT to decide whether Answer A or Answer B is better. Estimate the weakest Judge capability tier that would reliably agree with a careful human annotator about which answer better satisfies the user request.

The available Judge tiers are:
- Tier 1, Mistral-7B: a small, low-cost Judge. It is suitable when one answer is obviously better, broken, irrelevant, or clearly less useful. It is less reliable for close comparisons, specialized facts, subtle instruction following, long multi-step reasoning, mathematics, and code analysis.
- Tier 2, Qwen2.5-14B: a medium Judge. It is more reliable than Tier 1 for multilingual text, long answers, mathematics, code, multi-step reasoning, and nuanced comparisons, but can still be unreliable on very close, ambiguous, highly specialized, or verification-heavy cases.
- Tier 3, GPT-5.6-Luna: the strongest and most expensive Judge. It should be reserved for cases where Tier 1 or Tier 2 may fail because the answers are extremely close, technically demanding, ambiguous, or require deep factual and logical verification.

Use exactly one of these ordered routing levels:
1 = Tier 1 is clearly sufficient; the decisive quality difference is obvious.
2 = Tier 1 is probably sufficient, although Tier 2 would be safer.
3 = Tier 1 may be unreliable; Tier 2 is probably sufficient.
4 = Tier 2 may also be unreliable; Tier 3 would be substantially safer.
5 = Tier 3 is clearly needed, and even Tier 3 may find the comparison demanding.

Judge the capability required for COMPARING the answers, not how difficult it would be to write an answer from scratch. Consider the closeness of the two answers, factual-verification burden, specialized knowledge, mathematical or code reasoning, instruction-following subtleties, ambiguity, and susceptibility to superficial style or verbosity bias. Do not assume a longer answer is better.

Return only the single digit 1, 2, 3, 4, or 5.

Figure 14: Complete system prompt for Prompt-v3 (required Judge tier).

All three versions share the same user-message structure. Their version-specific opening sentences and answer prefixes are:

- Prompt-v1: `Estimate only the comparison difficulty for the following sample.` and `Difficulty level:`
- Prompt-v2: `Apply the fixed five-factor rubric and estimate the risk that a low- or medium-capability Judge will disagree with a careful human.` and `Judge failure-risk level:`
- Prompt-v3: `Estimate the weakest Judge capability tier required to compare the following sample reliably.` and `Required-tier routing level:`

The router reuses only the instruction-and-two-answers payload illustrated in Figure 10, but labels the fields as `[User request]`, `[Answer A]`, and `[Answer B]`. Unlike the formal Judge prompt, the routing message contains neither labeled demonstrations nor a `[Preferred response]` field; it begins and ends with the version-specific task sentence and answer prefix listed above, and returns a single routing level in $\{1, 2, 3, 4, 5\}$.