LEARNING WHEN TO SEEK HUMAN GUIDANCE FOR BEST LLM JUDGE IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM judges make it possible to evaluate model outputs at a scale that would be costly to achieve through human review alone. Before deploying a judge at scale, however, practitioners need to determine whether its judgments agree with human evaluations on the task of interest. Once a reliable judge is identified, it can support large-scale evaluation with substantially less human supervision. The main bottleneck is therefore the human labeling required to distinguish among candidate judges. This problem differs from classical best-arm identification because each human label simultaneously reveals the correctness of all candidate judges on the same item. As a result, the value of labeling an item depends on how much it helps distinguish the judges that are most difficult to separate. Overall disagreement alone does not capture this information, since disagreement among clearly inferior judges has little relevance to identifying the best judge. This shared-information structure motivates Weighted Pairwise Discrepancies (WPD), which assigns higher sampling priority to items that are informative for the most consequential pairwise comparisons. Because such selective labeling changes the distribution of reviewed items, we also develop valid inference for judge accuracies and their differences under the learned allocation. Across simulations and five real-data evaluation panels, WPD improves the probability of correct selection, reduces the number of human labels required to identify the best judge, and maintains valid uncertainty quantification.

1 INTRODUCTION

1.1 MOTIVATION AND CONTRIBUTIONS

LLM judges can evaluate model responses, assess relevance, and detect defects at a scale that would be costly to achieve through human review alone (Zheng et al., 2023; Liu et al., 2023). Before using an LLM judge for large-scale evaluation, however, practitioners need to determine whether its judgments agree with human judgments on the task of interest. This agreement can vary substantially across models, prompts, and evaluation procedures, which often produce different judgments on the same item (Chiang & Lee, 2023). Human labels are therefore needed to compare candidate judges and identify the judge with the highest accuracy. Collecting enough human labels to distinguish the strongest candidates can be expensive, especially when each item requires expert judgment. The central problem is how to identify the best judge with a limited human-labeling budget and quantify the uncertainty in the accuracy differences used to make this choice.

The limited labeling budget makes it important to determine which items should receive human labels. This decision differs from the allocation problem studied in classical best-arm identification, in which each observation reveals the performance of the selected arm (Kaufmann et al., 2016). In our setting, labeling decisions concern items rather than judges. Once an item receives a human label, we learn which judges were correct on that item. An allocation over individual judges therefore cannot capture all the information provided by a single human label. Active model selection methods, on the other hand, do make labeling decisions over items using the judges' predictions (Hara et al., 2024; Okanovic et al., 2025; Kay et al., 2025). However, their criteria can prioritize items that do not help distinguish the strongest judges. For example, large overall disagreement may be driven by weaker judges, so labeling such items may provide little information about the comparison between the strongest judges. Model-based information gain instead depends on assumptions about

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.

the unknown human labels, and errors in these assumptions can misdirect the labeling budget. We therefore need to identify the items that are most informative for the comparisons that determine the best judge, while accounting for the fact that one human label can inform several comparisons at once.

Selective human labeling can make better use of the labeling budget, but it also complicates statistical inference (Hadad et al., 2021; Farquhar et al., 2021). The labeled items may not represent the full item collection because their selection depends on the judges' predictions and, in later stages, on earlier human labels. Directly computing judge accuracies from these selected items may therefore produce invalid estimates of their accuracies on the full item collection. In addition, the same human labels are used to evaluate every candidate judge, so the resulting accuracy estimates and accuracy differences are statistically dependent. These issues matter because identifying the judge with the largest estimated accuracy is not enough to establish that it is meaningfully better than the alternatives. The inference procedure must provide valid confidence intervals for judge accuracies and their differences while accounting for both adaptive item selection and the dependence created by evaluating all judges on the same labeled items.

To address these challenges, we propose an adaptive human labeling framework based on Weighted Pairwise Discrepancies (WPD). Our contributions are as follows.

From a methodological perspective, we derive a labeling rule that combines each item's contributions across judge comparisons, with weights determined by a large-deviation surrogate for correct selection. A uniform pilot estimates the leading judge and its accuracy gaps, which guide the remaining labeling budget. The rule extends to categorical labels and numerical scores through conservative or imputed discrepancies.

From a theoretical perspective, we characterize the oracle allocation and its improvement over uniform labeling under the deviation criterion. We establish convergence of the learned allocation, consistent best-judge identification, joint asymptotic normality of the accuracy estimators, and consistent covariance estimation. These results provide confidence intervals for judge accuracies and their differences using labels from both stages.

From an empirical perspective, we evaluate WPD in simulations with independent and shared judge errors and on five real-data panels covering preference evaluation, relevance assessment, and visual defect detection. WPD improves correct selection and reduces the human labels needed to reach a target PCS. Coverage experiments further support the validity of the proposed inference procedure.

1.2 Related literature

Best-arm identification and large-deviation methods. Best-arm identification studies how to allocate observations to select the strongest candidate under a fixed budget or reach a prescribed confidence level (Audibert et al., 2010; Garivier & Kaufmann, 2016). Large-deviation methods compare allocations through the decay of incorrect-selection probabilities and emphasize the hardest comparisons (Glynn & Juneja, 2004; Wang et al., 2023). We use this perspective to construct an item-level labeling objective. Here, one human label informs several judges simultaneously, so comparison variances depend on a common allocation over items. The quadratic criterion uses the moderate-deviation behavior of these comparisons (Merlevède & Peligrad, 2009).

Active model selection and shared feedback. Active model selection uses candidate predictions to choose informative examples for labeling. These methods share our observation structure because one reference label evaluates every candidate on the queried item. Hara et al. (2024) minimize the aggregate variance of pairwise performance differences, leading to disagreement-based acquisition. Model Selector (Okanovic et al., 2025) selects labels through information gain about the best model, while CODA (Kay et al., 2025) uses a probabilistic model of consensus and disagreement to guide acquisition. Our allocation links each item's contribution to the accuracy gaps and the comparisons that limit a large-deviation criterion for correct selection, and combines this allocation with joint inference for judge performance.

Inference under adaptive data collection. Adaptive experiments use accumulated observations to guide later decisions, which changes the behavior of conventional estimators and confidence intervals. Zhang et al. (2020) develop inference for batched bandit data, and Hadad et al. (2021) use adaptive weighting to obtain confidence intervals for policy evaluation. Martingale limit theory pro-

vides a foundation for these analyses (Hall & Heyde, 1980). We study a fixed item collection whose human labels are selectively revealed. The learned labeling probabilities and shared feedback enter the joint covariance of the accuracy estimators, allowing the same labels to support best-judge identification and inference for performance differences.

2 Problem Setup

In this section, we formulate the problem of best LLM judge identification with selective human labeling. We consider N items, such as pairs of text and AI-generated images, evaluated by M LLM judges. For each item i, judge j provides a predicted label g_{ij}, and Y_i denotes the human label, which serves as the ground truth. All judge labels are available before human labeling, whereas Y_i is observed only if item i is selected for human labeling. We begin with binary labels $g_{ij}, Y_i \in \{0, 1\}$ and extend the framework to non-binary labels and evaluation scores in Section 5.

We aim to identify the judge whose labels most closely agree with human decisions. We measure judge j's accuracy across the N items by $\mu_j = N^{-1} \sum_{i=1}^{N} \mathbf{1}\{g_{ij} = Y_i\}$ and define the best judge as $b = \arg\max_{1 \le j \le M} \mu_j$. Once identified, this judge can be deployed for large-scale evaluations at lower cost than repeated human labeling. We also extend the target to top-K judge identification in Appendix B.

Because collecting human labels for all items is expensive, we use the observed judge labels to decide which items to label. For each item i, let $R_i \in \{0, 1\}$ indicate whether its human label is collected, with $R_i = 1$ if it is collected and $R_i = 0$ otherwise. Let $\pi_i := \Pr(R_i = 1)$ denote its labeling probability. Once Y_i is collected, the correctness indicators $\mathbf{1}\{g_{ij} = Y_i\}$ are observed for every judge j. Thus, one human label provides information about all judges on the selected item.

We use inverse-probability weighting (IPW) (Horvitz & Thompson, 1952) to estimate judge accuracy from these selectively collected human labels. For a fixed design $\pi := (\pi_1, \ldots, \pi_N)$, we estimate each μ_j by $\widehat{\mu}_j = N^{-1} \sum_{i=1}^{N} (R_i/\pi_i) \mathbf{1}\{g_{ij} = Y_i\}$, for $j = 1, \ldots, M$. We then select the judge with the highest estimated accuracy, $\widehat{b} = \arg\max_{1 \le j \le M} \widehat{\mu}_j$. Our goal is to choose labeling probabilities that maximize the probability of correct selection (PCS), $\mathrm{PCS}(\pi) := \Pr_\pi(\widehat{b} = b)$, under a labeling budget:

$$\max_{\pi \in (0,1]^N} \mathrm{PCS}(\pi) \qquad \text{subject to} \qquad \sum_{i=1}^{N} \pi_i \le B. \tag{1}$$

We assign unit cost to each human label, so B is the expected labeling budget.

3 Labeling Objective and Oracle Probabilities

In this section, we develop a tractable surrogate for PCS maximization in Section 3.1. We then obtain the corresponding oracle labeling probabilities in Section 3.2, assuming that the best judge and its accuracy gaps are known.

3.1 A large-deviation perspective on the design objective

Directly optimizing PCS is difficult because it depends on the joint distribution of all estimated judge accuracies. Its large-sample limit also provides little guidance for allocation. Many designs achieve PCS approaching one as more labels are collected, even though they use labels with different efficiency. Following the large-deviation approach to optimal selection (Glynn & Juneja, 2004), we therefore compare how quickly their probabilities of incorrect selection decrease. This gives us a criterion for choosing among designs that eventually identify the same best judge.

We build this criterion from comparisons between the best judge and each competing judge. For each $c \neq b$, let $\Delta_c := \mu_b - \mu_c > 0$ denote the accuracy gap and let $V_c(\pi) := \mathrm{Var}_\pi(\widehat{\mu}_b - \widehat{\mu}_c)$ denote the variance of the estimated gap. A larger gap makes the judges easier to distinguish, while a larger variance makes their comparison less reliable. A quadratic large-deviation approximation

summarizes this relationship as

$$G_c(\pi) = \frac{\Delta_c^2}{2V_c(\pi)}, \qquad -\log\{1 - \mathrm{PCS}(\pi)\} \approx \min_{c \neq b} G_c(\pi). \tag{2}$$

The quadratic form follows from a moderate-deviation approximation to the comparison errors (Merlevède & Peligrad, 2009). Since any competing judge can cause a selection error, the smallest $G_c(\pi)$ represents the comparison that is hardest to resolve. The variance $V_c(\pi)$ accounts for the fact that all judges are evaluated using the same human labels.

We therefore replace direct PCS maximization in (1) with the following surrogate problem:

$$\begin{aligned} \max_{\pi} \quad & \min_{c \neq b} \frac{\Delta_c^2}{2V_c(\pi)} \\ \text{subject to} \quad & \sum_{i=1}^{N} \pi_i \leq B, \qquad \varepsilon \leq \pi_i \leq 1, \quad i = 1, \ldots, N, \end{aligned} \tag{3}$$

where $0 < \varepsilon < B/N$. This objective allocates labels to improve the hardest comparison while respecting the expected budget. The lower bound ε gives every item a positive labeling probability. We use the convention $G_c(\pi) = \infty$ when $V_c(\pi) = 0$.

3.2 Oracle labeling probabilities

We assign higher labeling probabilities to items that help distinguish the best judge from its competitors. For binary labels, a human label distinguishes two judges only when their predictions disagree. Solving the oracle problem (3) gives weights $\alpha_c^\star \geq 0$ with $\sum_{c \neq b} \alpha_c^\star = 1$, which combine these disagreements into the information value of item i:

$$\sum_{c \neq b} \underbrace{\frac{\alpha_c^\star}{\Delta_c^2}}_{\text{comparison weight}} \underbrace{\mathbf{1}\{g_{ib} \neq g_{ic}\}}_{\text{disagreement on item } i}. \tag{4}$$

The comparison weight combines the difficulty of distinguishing the two judges with its importance for selection. The factor $1/\Delta_c^2$ emphasizes small accuracy gaps, while $\alpha_c^\star$ identifies the comparisons that limit correct selection. A disagreement contributes this weight, and agreement contributes zero. Contributions add when one item informs several comparisons, so a single human label can be especially valuable when it separates the best judge from several important competitors.

We translate these item values into labeling probabilities through a square-root rule:

$$\pi_i^* = \mathrm{clip}\left(\frac{1}{\lambda^\star} \sqrt{\sum_{c \neq b} \frac{\alpha_c^\star}{\Delta_c^2} \mathbf{1}\{g_{ib} \neq g_{ic}\}}; \varepsilon, 1 \right). \tag{5}$$

Here $\mathrm{clip}(x; \varepsilon, 1)$ truncates x to $[\varepsilon, 1]$, and $\lambda^\star$ scales the probabilities to satisfy the labeling budget. The square root reflects diminishing reductions in variance as an item's labeling probability increases. Both the labeling probabilities and the comparison weights are obtained by solving the equivalent convex formulation of (3).

4 Adaptive Labeling and Statistical Inference

In this section, we implement the oracle rule using a uniform pilot followed by optimized labeling. We then combine labels from both stages to identify the best judge and construct confidence intervals for judge accuracies. Figure 1 summarizes the full procedure.

Stage 1: Pilot labeling. We allocate $B_1 < B$ human labels to the pilot and select B_1 items uniformly without replacement. Denote this set by $\mathcal{P}$. After collecting their human labels, we estimate judge accuracies by $\widehat{\mu}_j^{(1)} = B_1^{-1} \sum_{i \in \mathcal{P}} \mathbf{1}\{g_{ij} = Y_i\}$ and select $\widehat{b}_1 = \arg\max_{1 \leq j \leq M} \widehat{\mu}_j^{(1)}$ as the pilot best judge. For each $c \neq \widehat{b}_1$, we use the estimated gap $\widehat{\Delta}_c = \max\{\widehat{\mu}_{\widehat{b}_1}^{(1)} - \widehat{\mu}_c^{(1)}, \kappa_{B_1}\}$, where the small positive threshold κ_{B_1} stabilizes the allocation when pilot accuracy estimates are close.

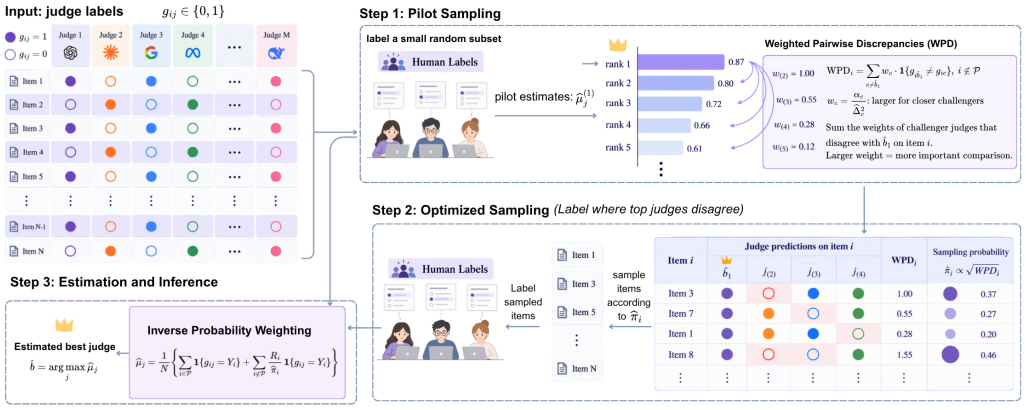

Figure 1: Overview of the two-stage WPD procedure.

Stage 2: Optimized labeling. We use the pilot estimates to allocate the remaining budget $B - B_1$ across unlabeled items by solving

$$\begin{aligned} \min_{\pi} \quad & \max_{c \neq \widehat{b}_1} \frac{1}{N^2 \widehat{\Delta}_c^2} \sum_{i \notin \mathcal{P}} \left(\frac{1}{\pi_i} - 1 \right) \mathbf{1}\{g_{i\widehat{b}_1} \neq g_{ic}\} \\ \text{subject to} \quad & \sum_{i \notin \mathcal{P}} \pi_i \leq B - B_1, \qquad \varepsilon \leq \pi_i \leq 1, \quad i \notin \mathcal{P}. \end{aligned} \tag{6}$$

We choose ε so that $\varepsilon(N - B_1) < B - B_1$ and denote a solution by $\widehat{\pi}$. Keeping these probabilities fixed, we independently collect each remaining item's human label with probability $\widehat{\pi}_i$, and let R_i indicate whether its label is collected. This allocation uses the pilot to identify which comparisons to prioritize and the observed judge disagreements to identify which items inform those comparisons.

Estimation and inference. We combine labels from both stages to estimate each judge's accuracy:

$$\widehat{\mu}_j = \frac{1}{N} \left\{ \sum_{i \in \mathcal{P}} \mathbf{1}\{g_{ij} = Y_i\} + \sum_{i \notin \mathcal{P}} \frac{R_i}{\widehat{\pi}_i} \mathbf{1}\{g_{ij} = Y_i\} \right\}. \tag{7}$$

Pilot labels enter directly, while inverse-probability weighting accounts for unequal labeling probabilities among the remaining items. We select $\widehat{b} = \arg\max_{1 \leq j \leq M} \widehat{\mu}_j$ and estimate the variance of each accuracy estimator by

$$\widehat{\mathrm{se}}_j^2 = \frac{1}{N^2} \sum_{i \notin \mathcal{P}} \frac{R_i(1 - \widehat{\pi}_i)}{\widehat{\pi}_i^2} \mathbf{1}\{g_{ij} = Y_i\}. \tag{8}$$

For a given judge j, a two-sided confidence interval at level $1 - \eta$ is $\widehat{\mu}_j \pm z_{1-\eta/2} \widehat{\mathrm{se}}_j$, where z_a is the a-quantile of the standard normal distribution. Theorem 6.3 establishes joint asymptotic normality of the accuracy estimators and asymptotic coverage of these intervals.

5 Extension to Non-Binary Labels

In this section, we extend WPD to categorical labels and numerical scores. Let $\ell(g, y) \in [0, 1]$ measure agreement with a human label $y \in \mathcal{Y}$, with larger values indicating better performance. The target becomes $\mu_j = N^{-1} \sum_i \ell(g_{ij}, Y_i)$, while b and Δ_c retain their definitions. Since squared score differences generally depend on the unknown human labels, we replace the binary disagreement indicator in (6) with one of two discrepancy coefficients.

The *conservative rule* values an item by how much its human label could distinguish two judges. It uses $\overline{a}_{i,jk} = \sup_{y \in \mathcal{Y}} \{\ell(g_{ij}, y) - \ell(g_{ik}, y)\}^2$. An item receives high value if some possible human label would produce a large score difference between the judges. The coefficient therefore captures

the item's maximum potential contribution to distinguishing the judges. For categorical accuracy, it equals the disagreement indicator.

The *imputed rule* instead values an item by how much its human label is expected to distinguish the judges. We fit a human-label distribution $\widehat{P}_i$ from pilot labels and available item information, and use $\widetilde{a}_{i,jk} = \int \{\ell(g_{ij}, y) - \ell(g_{ik}, y)\}^2 \, \mathrm{d}\widehat{P}_i(y)$. This rule averages squared score differences according to the predicted probabilities of human labels, giving less weight to differences associated with unlikely labels. Our experiments show that the conservative rule often performs better than the imputed rule.

We extend estimation and inference by replacing correctness indicators in the pilot and final estimators with evaluation scores. The variance estimators use squared scores for individual judges and squared score differences for judge comparisons, retaining the same IPW variance weights. Appendix A.4 provides the corresponding theoretical extension.

6 Theoretical Investigation

In this section, we characterize the oracle allocation, compare it with uniform labeling, and establish valid inference for the learned allocation. We consider binary labels on fixed item collections as $N \to \infty$ with M fixed. Appendix A.4 extends the results to general scores. For notational convenience, let $\mathcal{F}_1$ contain the pilot indices and labels, all judge predictions, and the randomness used to fit the allocation. Write $\mathbf{g}_i = (g_{i1}, \ldots, g_{iM})$. We impose the following assumptions.

Assumption 1 (Separation)**.** The best judge b is fixed and eventually unique. With $\Delta_{\min} = \min_{c \neq b} \Delta_c$, assume $B_1 \Delta_{\min}^2 \to \infty$, $\kappa_{B_1}/\Delta_{\min} \to 0$, and $\Delta_c/\Delta_{\min} \to \delta_c \in [1, \infty)$ for every $c \neq b$.

Assumption 2 (Adaptive labeling and positivity)**.** The pilot is sampled uniformly without replacement, with $B_1 \to \infty$ and $B_1/N \to \rho_1 \in [0, 1)$. The $\mathcal{F}_1$-measurable probabilities $\widehat{\pi}_i$ solve (6) and satisfy $\varepsilon \leq \widehat{\pi}_i \leq 1$ for a fixed $\varepsilon > 0$. Conditional on $\mathcal{F}_1$, the second-stage indicators are independent Bernoulli variables with these probabilities.

Assumption 3 (Stable item frequencies and budget)**.** The proportion of items with $(\mathbf{g}_i, Y_i) = (g, y)$ converges to $p_{g,y}$ for every binary pattern (g, y). Writing $p_g = p_{g,0} + p_{g,1}$, assume $(B - B_1)/(N - B_1) \to \rho_2$, where $\varepsilon < \rho_2 < 1 - (1 - \varepsilon)(p_{\mathbf{0}} + p_{\mathbf{1}})$.

Let $\pi^*(g)$ minimize the limiting WPD criterion in (22), with frequencies p_g, relative gaps δ_c, and budget fraction ρ_2. Set $\pi_i^* = \pi^*(\mathbf{g}_i)$ and $\boldsymbol{I}(g, y) = (\mathbf{1}\{g_j = y\})_{j=1}^M$, and define

$$\Sigma = (1 - \rho_1) \sum_{g,y} p_{g,y} \{\pi^*(g)^{-1} - 1\} \boldsymbol{I}(g, y) \boldsymbol{I}(g, y)^\top. \tag{9}$$

Assumption 4 (Nondegenerate information)**.** The matrix Σ is positive definite.

Assumption 1 permits shrinking gaps and holds for fixed gaps when the pilot grows and regularization vanishes. Assumption 2 is enforced by design and keeps inverse-probability weights bounded. The frequency condition in Assumption 3 is natural when benchmarks grow from a common item distribution, while the budget restriction leaves some disagreement items probabilistically labeled. Assumption 4 holds when the correctness vectors of positive-frequency patterns labeled with probability below one span $\mathbb{R}^M$, supporting joint inference. We first show that the WPD rule solves the oracle allocation problem.

Lemma 6.1 (Oracle labeling rule)**.** *Suppose $0 < \varepsilon < B/N < 1$ and the best judge is unique. An oracle allocation exists. If some comparison variance remains positive at the optimum, every optimum follows* (5)*, with $\alpha_c^\star \geq 0$, $\sum_{c \neq b} \alpha_c^\star = 1$, and $\lambda^\star > 0$ enforcing the budget. Positive weights occur only for challengers maximizing $V_c(\pi^*)/\Delta_c^2$. If all comparison variances can be zero, an optimum labels every disagreement item with certainty.*

The optimization determines the comparison weights and labeling probabilities together. We next compare the resulting allocation with uniform labeling under the same budget. The following proposition shows that the oracle allocation achieves a deviation rate at least as large as that of uniform labeling.

Proposition 6.2 (Comparison with uniform labeling)**.** *Under Lemma 6.1, let $\pi_i^{\mathrm{u}} = B/N$. Then $\min_{c\neq b} G_c(\pi^*) \geq \min_{c\neq b} G_c(\pi^{\mathrm{u}})$, strictly if at least one item receives unanimous predictions. Moreover, if $\varepsilon > 0$ is fixed, $N\Delta_{\min}^2 \to \infty$, and, for every $c \neq b$, $\Delta_c \to 0$ and $V_c(\pi^*) \asymp V_c(\pi^{\mathrm{u}}) \asymp N^{-1}$, then*

$$\liminf_{N\to\infty} \frac{-\log\{1-\mathrm{PCS}(\pi^*)\}}{-\log\{1-\mathrm{PCS}(\pi^{\mathrm{u}})\}} \geq 1.$$

The rate improvement comes from reallocating labels from unanimous-agreement items to items that inform judge comparisons.

For inference, write $\mu = (\mu_1, \ldots, \mu_M)^\top$, $\widehat{\mu} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_M)^\top$, and $\boldsymbol{I}_i = \boldsymbol{I}(\mathbf{g}_i, Y_i)$. We estimate Σ by

$$\widehat{\Sigma}_N = \frac{1}{N} \sum_{i\notin\mathcal{P}} \frac{R_i(1-\widehat{\pi}_i)}{\widehat{\pi}_i^2} \boldsymbol{I}_i \boldsymbol{I}_i^\top. \tag{10}$$

The next theorem shows that the learned allocation converges to the limiting oracle allocation and that the procedure consistently identifies the best judge. It also establishes asymptotic normality of the accuracy estimates and consistency of the covariance estimator, providing the basis for confidence intervals for judge accuracies and their differences.

Theorem 6.3 (Learned allocation and valid inference)**.** *Under Assumptions 1–4, $\pi^*(g)$ is unique for $p_g > 0$, and*

$$\frac{1}{N} \sum_{i\notin\mathcal{P}} |\widehat{\pi}_i - \pi_i^*| \xrightarrow{p} 0, \qquad \Pr(\widehat{b} = b) \to 1. \tag{11}$$

The estimator (7) *satisfies*

$$\sqrt{N}(\widehat{\mu} - \mu) \xrightarrow{d} \mathcal{N}_M(0, \Sigma), \qquad \widehat{\Sigma}_N \xrightarrow{p} \Sigma. \tag{12}$$

Consequently, the marginal and pairwise Wald intervals in Section 4 and Appendix A.3 attain their nominal coverage asymptotically. The martingale proof in Appendix A.3 accounts for both the random pilot and the fitted allocation.

7 Experiments

7.1 Numerical Experiments

Experimental setup. We investigate whether weighting disagreements by their relevance to best-judge identification improves PCS. Each simulation contains $N = 10{,}000$ items and $M = 10$ judges, with binary or five-class labels. Binary settings contain two close leading judges with independent or shared errors; five-class settings additionally examine a group that often supplies the same incorrect label. We evaluate PCS over total budgets B from 500 to 1,500, including a uniform pilot of $B_1 = 100$, and examine $B_1 = 200$ in the appendix. Each fixed pool is evaluated over 500 labeling repetitions; the main binary curves average three separately generated pools. Appendix C.1 specifies the data-generating processes and labeling procedures.

Baselines. We compare WPD with **Uniform**, which allocates labels equally across items; **Hara** (Hara et al., 2024), which scores aggregate disagreement across judges; **CODA** (Kay et al., 2025), which models consensus and disagreement to reduce uncertainty about the best candidate; and **Model Selector** (Okanovic et al., 2025), which scores the expected information gain about the best candidate under a probabilistic model linking candidate predictions to human labels. For five-class labels, we compare **WPD conservative** with **WPD imputation**, using the item-level plurality label for hard imputation.

Results. WPD substantially improves correct selection when the leading judges are difficult to distinguish, as shown by the two binary settings in Figure 2. At $B = 1000$, WPD attains 87.9% PCS under independent binary errors, compared with 79.7% for Uniform, 79.6% for Hara, and 79.7% for CODA. Under shared binary errors, the corresponding values are 96.8%, 89.7%, 90.3%, and 91.9%. WPD also clearly improves on Model Selector in these settings. These gains demonstrate the value of allocating labels to resolve close accuracy comparisons. The advantage extends to non-binary labels. Under independent five-class errors, conservative WPD attains 93.8% PCS, compared with

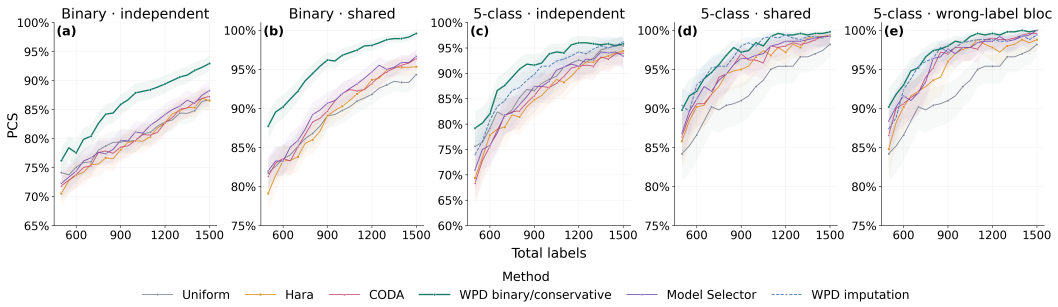

Figure 2: Simulation PCS with pilot $B_1 = 100$. Each pool uses 500 labeling repetitions, and binary panels average three pools. Shaded bands show pointwise 95% Monte Carlo intervals.

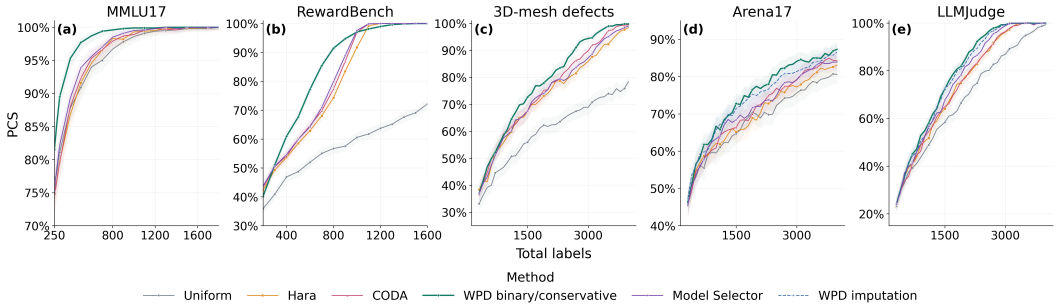

Figure 3: Real-data PCS with $B_1 = 100$ and pointwise 95% Monte Carlo intervals.

88.8% for Uniform and 87.2% for both Hara and CODA, while imputation attains 91.4%. Both WPD variants approach perfect selection under shared five-class errors and the wrong-label bloc, showing effective allocation even when judges often make the same mistake. Figures 6 and 7 provide further results across accuracy profiles and pilot sizes and show the resulting label savings. The coverage experiment in Appendix C.2 reports accuracy and pairwise-gap coverage close to the nominal 95% level, supporting the validity of the inference procedure.

7.2 Real-World Experiments

Datasets. We consider five complete item-by-judge panels spanning preference evaluation, relevance assessment, and visual defect detection. MMLU17 contains 28,084 MMLU-derived binary comparison items evaluated by 17 candidate models against standard reference labels (Hendrycks et al., 2020). RewardBench (Lambert et al., 2025) contains 2,305 response-comparison items with binary reference preferences and 12 judges. Arena17 (Zheng et al., 2023) contains 49,938 response comparisons with human preferences, including ties, and the same 17 candidate models as MMLU17. LLMJudge (Rahmani et al., 2025) contains 4,420 query–document pairs with four-level human relevance judgments and 33 judge configurations. The 3D defect dataset (Zhao et al., 2026) contains 7,858 evaluation 3D meshes and six vision-language judges, with a binary human reference label indicating whether a defect is present.

Experimental setup. We retain the candidate columns in each complete panel and define the best judge by full-panel exact agreement with the reference, including exact grade agreement for LLM-Judge. Reference labels are revealed to each labeling method only when queried; full-panel accuracies are used to score selection. The main comparison uses $B_1 = 100$, with $B_1 = 300$ as a sensitivity analysis, and the same baselines and WPD variants as the simulations. Budgets extend to 2,200 labels for RewardBench and 4,000 for the other panels. We use 3,000 labeling repetitions for RewardBench and 1,000 elsewhere. Label savings compare the budgets needed to attain the same empirical PCS, including pilot cost.

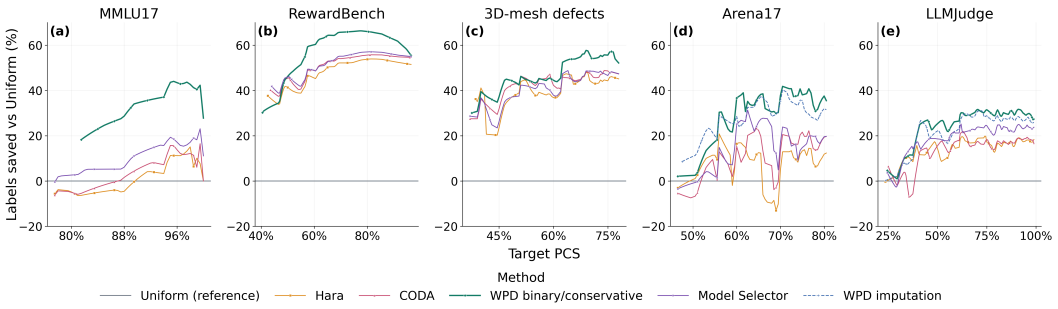

Figure 4: Labels saved relative to Uniform at matched target PCS, including the pilot. Curves show point estimates from isotonic PCS fits over the targets attained by both methods.

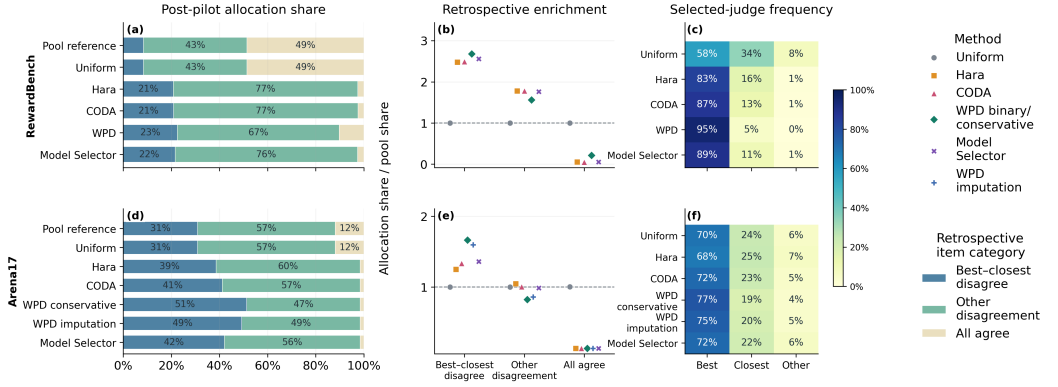

Figure 5: Post-pilot allocation shares, enrichment, and selection frequencies. Retrospective item categories use the full-panel best judge and closest competitor.

Selection performance. WPD increases the probability of selecting the best judge across the real-world tasks (Figure 3). On MMLU17 at $B = 500$, WPD reaches 97.7% PCS, versus 90.9% for Uniform, 91.6% for Hara, and 92.3% for CODA. On RewardBench at 900 expected labels, WPD reaches 94.7%, compared with 83.4% for Hara and 86.9% for CODA. On Arena17 at $B = 2000$, conservative WPD reaches 77.4%, versus 69.9%, 68.3%, and 72.1% for Uniform, Hara, and CODA, respectively, while imputation reaches 74.9%. Both WPD variants improve selection on LLMJudge, and WPD delivers a substantial gain over Uniform on defect detection. The curves also show gains over Model Selector on MMLU17, RewardBench, and Arena17 at modest labeling budgets.

Labeling efficiency. WPD reaches the same selection performance with substantially fewer human labels than Uniform (Figure 4). Around 90% target PCS, WPD saves approximately one-third of the labels on MMLU17 and three-fifths on RewardBench; on LLMJudge, savings are approximately 25–30% across much of the moderate-to-high PCS range. The defect panel shows savings exceeding 50% over part of the common attainable range below 80% PCS. Figure 8 further shows how the gains vary with pilot size.

Allocation and selection errors. WPD concentrates labels on disagreements between the leading judges and reduces selection of the closest competitor (Figure 5). On Arena17, the best–closest disagreement category contains approximately 31% of items but receives 51% of conservative WPD's second-stage labels, compared with 39% under Hara, 41% under CODA, and 42% under Model Selector. Correspondingly, selection of the closest competitor falls to approximately 19% under conservative WPD, versus 25%, 23%, and 22% under Hara, CODA, and Model Selector, respectively. On RewardBench, WPD allocates approximately 23% of second-stage labels to best–closest disagreements, compared with 21% for Hara and CODA and 22% for Model Selector. It selects the closest competitor about 5% of the time, compared with 11% for Model Selector.

AI USE STATEMENT

In this work, we used generative AI tools to assist with editing and polishing the manuscript for clarity, readability, and presentation, and to assist with writing and debugging experimental code. We did not rely on generative AI tools to independently formulate the research problem, establish the theoretical results, or make the final interpretation of the experimental findings. All AI-assisted text was reviewed and revised by the authors, and AI-assisted code was inspected, tested, and validated against the intended methodology and experimental procedures. The authors verified the mathematical statements, experimental results, and references reported in the paper. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REFERENCES

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *Proceedings of the 23rd Annual Conference on Learning Theory*, pp. 41–53, 2010.

Cheng-Han Chiang and Hung-yi Lee. A closer look into using large language models for automatic evaluation. In *Findings of the association for computational linguistics: EMNLP 2023*, pp. 8928–8942, 2023.

Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. In *International Conference on Learning Representations*, 2021.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 998–1027, 2016.

Peter W. Glynn and Sandeep Juneja. A large deviations perspective on ordinal optimization. In *Proceedings of the 2004 Winter Simulation Conference*, pp. 577–586, 2004.

Vitor Hadad, David A. Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, 118(15):e2014602118, 2021. doi: 10.1073/pnas.2014602118.

Peter Hall and Christopher C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, 1980.

Satoshi Hara, Mitsuru Matsuura, Junya Honda, and Shinji Ito. Active model selection: A variance minimization approach. *Machine Learning*, 113:8327–8345, 2024. doi: 10.1007/s10994-024-06603-1.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

Justin Kay, Grant Van Horn, Subhransu Maji, Daniel Sheldon, and Sara Beery. Consensus-driven active model selection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4594–4604, 2025.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.96.

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.

Florence Merlevède and Magda Peligrad. Functional moderate deviations for triangular arrays and applications. *ALEA. Latin American Journal of Probability and Mathematical Statistics*, 5:3–20, 2009.

Patrik Okanovic, Andreas Kirsch, Jannes Kasper, Torsten Hoefler, Andreas Krause, and Nezihe Merve Gürel. All models are wrong, some are useful: Model selection with limited labels. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2035–2043, 2025.

Hossein A. Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. Judging the judges: A collection of LLM-generated relevance judgements. *arXiv preprint arXiv:2502.13908*, 2025.

Po-An Wang, Ruo-Chun Tzeng, and Alexandre Proutiere. Best arm identification with fixed budget: A large deviation perspective. *Advances in Neural Information Processing Systems*, 36:16804–16815, 2023.

Kelly W. Zhang, Lucas Janson, and Susan A. Murphy. Inference for batched bandits. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9818–9829, 2020.

Zhenyu Zhao, Nanshan Jia, Jihyeon Je, Yifu Tang, Alvin Chan, Michael Spedden, Michael V Palleschi, Sui Huang, Jingshen Wang, and Zeyu Zheng. 3d-defectbench: A controlled factorial study of vision-language model evaluation pipelines for fine-grained 3d generation defects. *arXiv preprint arXiv:2607.10826*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023.

A PROOFS

We prove the three results in Section 6 and then extend them to general evaluation scores. All probabilities and expectations are with respect to the labeling design: the item collection, judge predictions, and human labels are fixed. Quantities may depend on N; we suppress this dependence except where needed for clarity. For binary labels, write $I_{ij} = \mathbf{1}\{g_{ij} = Y_i\}$, so that $\boldsymbol{I}_i = (I_{i1}, \ldots, I_{iM})^\top$. For a fixed design with independent $R_i \sim \text{Bernoulli}(\pi_i)$, independence and $(I_{ib} - I_{ic})^2 = \mathbf{1}\{g_{ib} \neq g_{ic}\}$ give

$$V_c(\pi) = \frac{1}{N^2}\sum_{i=1}^{N}(\pi_i^{-1} - 1)(I_{ib} - I_{ic})^2 = \frac{1}{N^2}\sum_{i=1}^{N}(\pi_i^{-1} - 1)\mathbf{1}\{g_{ib} \neq g_{ic}\}. \tag{13}$$

A.1 PROOF OF LEMMA 6.1

Proof. Since every gap is positive, the oracle objective satisfies

$$\min_{c\neq b} G_c(\pi) = \left\{2\max_{c\neq b}\frac{V_c(\pi)}{\Delta_c^2}\right\}^{-1},$$

with the value $+\infty$ when all variances vanish. Thus maximizing this objective is equivalent to minimizing the largest normalized variance. The feasible set is nonempty, since it contains $\pi_i = B/N$, and is compact. Moreover,

$$\frac{\partial^2}{\partial\pi_i^2}\frac{V_c(\pi)}{\Delta_c^2} = \frac{2\mathbf{1}\{g_{ib} \neq g_{ic}\}}{N^2\Delta_c^2\pi_i^3} \geq 0.$$

Each normalized variance is continuous and convex on the feasible set, as is their maximum. A minimizer therefore exists.

Introduce an epigraph variable t and constraints $V_c(\pi)/\Delta_c^2 \leq t$. Choose a constant r satisfying $\varepsilon < r < B/N < 1$, set $\pi_i = r$, and take $t > \max_c V_c(\pi)/\Delta_c^2$. All inequality constraints are then strict, so Slater's condition holds. The Karush–Kuhn–Tucker conditions are necessary and sufficient for optimality.

Let $\alpha_c \geq 0$ be the multiplier for comparison c and $\nu \geq 0$ the budget multiplier. Keeping the box constraints in the minimization domain, the Lagrangian is

$$\begin{aligned} L(\pi, t; \alpha, \nu) &= t\left(1 - \sum_{c\neq b}\alpha_c\right) + \frac{1}{N^2}\sum_i(\pi_i^{-1} - 1)w_i + \nu\left(\sum_i \pi_i - B\right), \\ w_i &= \sum_{c\neq b}\frac{\alpha_c}{\Delta_c^2}\mathbf{1}\{g_{ib} \neq g_{ic}\}. \end{aligned} \tag{14}$$

Stationarity in t and complementary slackness give

$$\begin{gathered} \sum_{c\neq b}\alpha_c = 1, \qquad \alpha_c\left\{\frac{V_c(\pi^*)}{\Delta_c^2} - t^\star\right\} = 0, \qquad \nu\left(\sum_i \pi_i^* - B\right) = 0, \\ \sum_{c\neq b}\alpha_c\frac{V_c(\pi^*)}{\Delta_c^2} = t^\star. \end{gathered} \tag{15}$$

In particular, positive comparison weights are confined to comparisons whose normalized variance equals the maximum $t^\star$.

Suppose $t^\star > 0$. If $\nu = 0$, minimizing (14) over the box requires $\pi_i^* = 1$ whenever $w_i > 0$, because w_i/π_i is strictly decreasing. Hence

$$\sum_{c\neq b}\alpha_c\frac{V_c(\pi^*)}{\Delta_c^2} = \frac{1}{N^2}\sum_i\{(\pi_i^*)^{-1} - 1\}w_i = 0,$$

contradicting (15). Therefore $\nu > 0$, and the budget constraint binds.

For each coordinate, Lagrangian minimization reduces to minimizing $f_i(p) = w_i/(N^2 p) + \nu p$ over $[\varepsilon, 1]$. Its derivatives are

$$f_i'(p) = \nu - \frac{w_i}{N^2 p^2}, \qquad f_i''(p) = \frac{2w_i}{N^2 p^3} \geq 0.$$

If $w_i > 0$, the stationary point is $p = \sqrt{w_i}/(N\sqrt{\nu})$. The derivative is nonnegative throughout the interval when this point is at most ε, and nonpositive throughout when it is at least 1. Otherwise the stationary point is the unique minimum. If $w_i = 0$, f_i is strictly increasing and its minimum is ε. These cases give

$$\pi_i^* = \min\left\{1, \max\left\{\varepsilon, \frac{\sqrt{w_i}}{N\sqrt{\nu}}\right\}\right\}.$$

Taking $\lambda^\star = N\sqrt{\nu}$ yields (5). Since the KKT conditions are necessary for every minimizer, the characterization applies to every optimal allocation with $t^\star > 0$.

Finally, if $t^\star = 0$, nonnegativity of the summands in (13) implies

$$V_c(\pi) = 0 \text{ for all } c \neq b \quad \Longleftrightarrow \quad \pi_i = 1 \text{ whenever some } c \neq b \text{ has } g_{ib} \neq g_{ic}.$$

These are precisely the items on which the judges do not all agree. Whenever such an allocation is feasible, it attains zero comparison variance and is optimal. This proves the zero-variance case as well. □

A.2 Proof of Proposition 6.2

Proof. We first compare the leading terms of the log incorrect-selection probabilities under the oracle and uniform allocations. Work under the additional moderate-deviation conditions in the proposition, and fix $\pi \in \{\pi^*, \pi^{\mathrm{u}}\}$ and a challenger $c \neq b$. The comparison error can be written as

$$\begin{aligned} N(\widehat{\mu}_b - \widehat{\mu}_c - \Delta_c) &= \sum_i \xi_i, \qquad \xi_i = \left(\frac{R_i}{\pi_i} - 1\right)(I_{ib} - I_{ic}), \\ \mathbb{E}_\pi \xi_i = 0, \qquad \sum_i \mathbb{E}_\pi \xi_i^2 &= N^2 V_c(\pi), \qquad |\xi_i| \leq \varepsilon^{-1}. \end{aligned} \tag{16}$$

The summands are independent under either allocation. To apply the moderate-deviation result of Merlevède & Peligrad (2009, Corollary 1.5 and Comment 1.4), take its total variance to be $N^2 V_c(\pi)$ and its speed parameter to be $V_c(\pi)/\Delta_c^2$. The required conditions hold because

$$\begin{aligned} \frac{V_c(\pi)}{\Delta_c^2} &\asymp \frac{1}{N\Delta_c^2} \longrightarrow 0, \\ \frac{\max_i |\xi_i|}{NV_c(\pi)/\Delta_c} &\leq \frac{\Delta_c}{\varepsilon N V_c(\pi)} = O(\Delta_c) \longrightarrow 0. \end{aligned} \tag{17}$$

In particular, the second bound makes the indicator in their condition (1.7) identically zero for all sufficiently large N. The normalized contrast $(\widehat{\mu}_b - \widehat{\mu}_c - \Delta_c)/\Delta_c$ therefore has the moderate-deviation rate function $z^2/2$. Since $\inf_{z<-1} z^2/2 = \inf_{z\leq -1} z^2/2 = 1/2$, the open- and closed-set bounds give

$$\begin{aligned} \frac{\log \Pr_\pi(\widehat{\mu}_b - \widehat{\mu}_c < 0)}{G_c(\pi)} &\longrightarrow -1, \\ \frac{\log \Pr_\pi(\widehat{\mu}_b - \widehat{\mu}_c \leq 0)}{G_c(\pi)} &\longrightarrow -1. \end{aligned} \tag{18}$$

A selection error occurs whenever some challenger strictly exceeds b, and requires at least a tie with a challenger. Thus, for either design and any fixed tie-breaking rule,

$$\max_{c\neq b} \Pr_\pi(\widehat{\mu}_b - \widehat{\mu}_c < 0) \leq 1 - \mathrm{PCS}(\pi) \leq \sum_{c\neq b} \Pr_\pi(\widehat{\mu}_b - \widehat{\mu}_c \leq 0).$$

This step does not require independence across judge comparisons. Since M is fixed, the remainders in (18) are uniform over challengers. Taking logarithms yields

$$\begin{aligned} -\{1 + o(1)\} \min_{c\neq b} G_c(\pi) &\leq \log\{1 - \mathrm{PCS}(\pi)\} \\ &\leq \log(M-1) - \{1 + o(1)\} \min_{c\neq b} G_c(\pi). \end{aligned} \tag{19}$$

The additive term $\log(M-1)$ is negligible because $\min_c G_c(\pi) \asymp N \min_c \Delta_c^2 \to \infty$. Consequently, under the proposed oracle allocation and uniform labeling, respectively,

$$\frac{-\log\{1-\mathrm{PCS}(\pi^*)\}}{\min_{c\neq b} G_c(\pi^*)} \longrightarrow 1, \qquad \frac{-\log\{1-\mathrm{PCS}(\pi^{\mathrm{u}})\}}{\min_{c\neq b} G_c(\pi^{\mathrm{u}})} \longrightarrow 1. \tag{20}$$

Uniform sampling is feasible in (3), so optimality of π^* gives the deviation-rate inequality in Proposition 6.2 for every N, independently of the additional asymptotic conditions. Combining this criterion inequality with (20) gives

$$\frac{-\log\{1-\mathrm{PCS}(\pi^*)\}}{-\log\{1-\mathrm{PCS}(\pi^{\mathrm{u}})\}} = \{1+o(1)\}\frac{\min_{c\neq b} G_c(\pi^*)}{\min_{c\neq b} G_c(\pi^{\mathrm{u}})}.$$

The final ratio of criteria is at least one, proving the claimed limit-inferior inequality.

It remains to establish strict improvement of the finite-N criterion when an item i_0 receives the same prediction from every judge. Let $\rho = B/N$ and choose

$$0 < \delta < \min\{\rho-\varepsilon, (N-1)(1-\rho)\}, \qquad \pi_{i_0} = \rho - \delta, \qquad \pi_i = r := \rho + \frac{\delta}{N-1} \quad (i \neq i_0).$$

These probabilities lie in $(\varepsilon, 1)$ and sum to B. For every challenger, $N\Delta_c = \sum_i (I_{ib} - I_{ic}) \leq \sum_i \mathbf{1}\{g_{ib} \neq g_{ic}\}$. Since $\Delta_c > 0$, at least one item contributes to this comparison, and none of its disagreement items is i_0. Equation (13) therefore gives

$$V_c(\pi) = \frac{r^{-1}-1}{\rho^{-1}-1} V_c(\pi^{\mathrm{u}}), \qquad G_c(\pi) = \frac{\rho^{-1}-1}{r^{-1}-1} G_c(\pi^{\mathrm{u}}).$$

The multiplier of $G_c(\pi^{\mathrm{u}})$ is common to all challengers and strictly exceeds one. All uniform comparison variances are positive, so

$$\min_{c\neq b} G_c(\pi^*) \geq \min_{c\neq b} G_c(\pi) > \min_{c\neq b} G_c(\pi^{\mathrm{u}}).$$

This proves the strict criterion inequality. Its gain may vanish as N grows, so strictness of the asymptotic probability-rate comparison is not asserted. □

A.3 Proof of Theorem 6.3

A.3.1 Consistency

Proof. Pilot estimates. Simple random sampling gives

$$\mathbb{E}\Big[(\widehat{\mu}_j^{(1)} - \mu_j)^2\Big] = \frac{N-B_1}{B_1(N-1)}\mu_j(1-\mu_j) \leq \frac{1}{4B_1}.$$

Since M is fixed, Assumption 1 implies $\max_j |\widehat{\mu}_j^{(1)} - \mu_j| = O_p(B_1^{-1/2}) = o_p(\Delta_{\min})$. In particular,

$$\Pr(\widehat{b}_1 \neq b) \leq \Pr\left(\max_j |\widehat{\mu}_j^{(1)} - \mu_j| \geq \Delta_{\min}/2\right) \leq \frac{M}{B_1\Delta_{\min}^2} \to 0.$$

On an event with probability tending to one, the untruncated pilot gaps are at least $\Delta_{\min}/2$. Since $\kappa_{B_1}/\Delta_{\min} \to 0$, truncation is inactive on this event and

$$\max_{c\neq b}\left|\frac{\widehat{\Delta}_c}{\Delta_{\min}} - \frac{\Delta_c}{\Delta_{\min}}\right| \leq \frac{2\max_j |\widehat{\mu}_j^{(1)} - \mu_j|}{\Delta_{\min}} = o_p(1), \qquad \frac{\widehat{\Delta}_c}{\Delta_{\min}} \xrightarrow{p} \delta_c. \tag{21}$$

Let $p_{N,g} = N^{-1}\sum_i \mathbf{1}\{\mathbf{g}_i = g\}$ and $\widehat{p}_g = (N-B_1)^{-1}\sum_{i\notin\mathcal{P}} \mathbf{1}\{\mathbf{g}_i = g\}$. The remaining items form a simple random sample of size $N - B_1$, so

$$\mathbb{E}\widehat{p}_g = p_{N,g}, \qquad \mathbb{E}(\widehat{p}_g - p_{N,g})^2 = \frac{B_1}{(N-B_1)(N-1)} p_{N,g}(1-p_{N,g}) = O(N^{-1}).$$

Assumptions 2 and 3, and the finite number of prediction patterns, give $\max_g |\widehat{p}_g - p_g| \xrightarrow{p} 0$.

Limiting allocation. The normalized population problem is

$$\begin{aligned} \min_{\pi} \quad & F(\pi) := \max_{c \neq b} \frac{1}{\delta_c^2} \sum_g p_g \{\pi(g)^{-1} - 1\} \mathbf{1}\{g_b \neq g_c\} \\ \text{subject to} \quad & \sum_g p_g \pi(g) \leq \rho_2, \qquad \varepsilon \leq \pi(g) \leq 1. \end{aligned} \tag{22}$$

Its feasible set is compact and contains a strictly feasible constant allocation because $\rho_2 > \varepsilon$. The continuous objective therefore attains its minimum, denoted by t^*. If $t^* = 0$, every pattern with $p_g > 0$ on which two judges disagree must have $\pi(g) = 1$. Its labeling cost is then at least $1 - (1 - \varepsilon)(p_{\mathbf{0}} + p_{\mathbf{1}}) > \rho_2$, a contradiction. Thus $t^* > 0$.

Slater's condition and the argument in Appendix A.1 give optimal multipliers $\alpha_c^* \geq 0$, $\sum_{c \neq b} \alpha_c^* = 1$, and $\nu^* > 0$. Define

$$w_g = \sum_{c \neq b} \frac{\alpha_c^*}{\delta_c^2} \mathbf{1}\{g_b \neq g_c\}, \qquad f_g(p) = w_g(p^{-1} - 1) + \nu^* p.$$

For $p_g > 0$, every primal optimum minimizes this same optimal Lagrangian coordinatewise. If $w_g > 0$, f_g is strictly convex; if $w_g = 0$, it is strictly increasing. Consequently, its minimizer is unique and equals

$$\pi^*(g) = \min\left\{1, \max\left\{\varepsilon, \sqrt{w_g/\nu^*}\right\}\right\}.$$

Use the same formula to define $\pi^*(g)$ on patterns with $p_g = 0$. Complementary slackness gives $\sum_g p_g \pi^*(g) = \rho_2$ and $\sum_g p_g f_g(\pi^*(g)) = t^* + \nu^* \rho_2$. Moreover, for any $p \in [\varepsilon, 1]$,

$$\begin{aligned} f_g(p) - f_g(\pi^*(g)) &= (p - \pi^*(g)) \left\{\nu^* - \frac{w_g}{p \pi^*(g)}\right\} \\ &\geq \frac{\nu^*}{p} \{p - \pi^*(g)\}^2 \geq \nu^* \{p - \pi^*(g)\}^2. \end{aligned} \tag{23}$$

The first inequality is an equality for an interior minimizer, where $w_g = \nu^* \pi^*(g)^2$. At the lower and upper endpoints it follows from $w_g \leq \nu^* \varepsilon^2$ and $w_g \geq \nu^*$, respectively.

Consistency of the optimized allocation. On $\{\widehat{b}_1 = b\}$, multiplying the objective in (6) by $N^2 \Delta_{\min}^2 / (N - B_1)$ gives

$$\widehat{F}_N(\pi) = \max_{c \neq b} \frac{\Delta_{\min}^2}{\widehat{\Delta}_c^2} \frac{1}{N - B_1} \sum_{i \notin \mathcal{P}} (\pi_i^{-1} - 1) \mathbf{1}\{g_{ib} \neq g_{ic}\}.$$

The fitted allocation minimizes this criterion subject to the average budget $\rho_{2,N} = (B - B_1)/(N - B_1)$. For fixed $\theta \in (0, 1)$, set $\pi_\theta(g) = (1 - \theta)\pi^*(g) + \theta\varepsilon$. Then

$$\sum_g \widehat{p}_g \pi_\theta(g) \xrightarrow{p} (1 - \theta)\rho_2 + \theta\varepsilon < \rho_2, \qquad \widehat{F}_N(\pi_\theta) \xrightarrow{p} F(\pi_\theta).$$

Here the second limit uses (21); π_θ in the empirical criterion denotes its values at $\mathbf{g}_i$. Since $\rho_{2,N} \to \rho_2$, this allocation is feasible with probability tending to one. Optimality therefore implies $\widehat{F}_N(\widehat{\pi}) \leq F(\pi_\theta) + o_p(1)$.

The bound $0 \leq \pi_i^{-1} - 1 \leq \varepsilon^{-1} - 1$ and (21) allow replacing $\Delta_{\min}^2 / \widehat{\Delta}_c^2$ by δ_c^{-2} uniformly over all feasible allocations. Combining the budget constraint with (23) yields

$$\begin{aligned} \frac{\nu^*}{N - B_1} \sum_{i \notin \mathcal{P}} (\widehat{\pi}_i - \pi_i^*)^2 &\leq \frac{1}{N - B_1} \sum_{i \notin \mathcal{P}} \{f_{\mathbf{g}_i}(\widehat{\pi}_i) - f_{\mathbf{g}_i}(\pi_i^*)\} \\ &\leq \widehat{F}_N(\widehat{\pi}) + \nu^* \rho_{2,N} - \sum_g \widehat{p}_g f_g(\pi^*(g)) + o_p(1) \\ &= \widehat{F}_N(\widehat{\pi}) - t^* + o_p(1) \\ &\leq F(\pi_\theta) - t^* + o_p(1). \end{aligned}$$

Continuity gives $F(\pi_\theta) \to t^*$ as $\theta \downarrow 0$. Since $\nu^* > 0$, the average squared allocation error tends to zero in probability. Cauchy–Schwarz then gives

$$\frac{1}{N}\sum_{i\notin\mathcal{P}}|\widehat{\pi}_i - \pi_i^*| \le \frac{N-B_1}{N}\left\{\frac{1}{N-B_1}\sum_{i\notin\mathcal{P}}(\widehat{\pi}_i-\pi_i^*)^2\right\}^{1/2} \xrightarrow{p} 0,$$

proving (11).

Accuracy estimation and selection. The pilot items are observed exactly, so subtracting the full-collection target from (7) gives

$$\sqrt{N}(\widehat{\mu}-\mu) = \frac{1}{\sqrt{N}}\sum_{i\notin\mathcal{P}}\left(\frac{R_i}{\widehat{\pi}_i}-1\right)\boldsymbol{I}_i = \sum_{i=1}^{N} D_i, \tag{24}$$

where D_i is the displayed vector summand for $i \notin \mathcal{P}$ and zero otherwise. Define

$$\mathcal{H}_0 = \mathcal{F}_1, \qquad \mathcal{H}_i = \sigma(\mathcal{F}_1, \{R_k, R_kY_k : k \le i,\ k \notin \mathcal{P}\}), \quad i = 1,\dots,N.$$

The correctness vectors are constants under the fixed-collection sampling law, and the pilot set and fitted probabilities are $\mathcal{H}_0$-measurable. For each remaining item, Assumption 2 gives

$$\mathbb{E}\left[\frac{R_i}{\widehat{\pi}_i}-1\,\middle|\,\mathcal{H}_{i-1}\right] = 0,$$
$$\mathbb{E}\left[\left(\frac{R_i}{\widehat{\pi}_i}-1\right)^2\,\middle|\,\mathcal{H}_{i-1}\right] = \widehat{\pi}_i\left(\frac{1}{\widehat{\pi}_i}-1\right)^2 + (1-\widehat{\pi}_i) = \widehat{\pi}_i^{-1}-1.$$

Thus $(D_i, \mathcal{H}_i)_{i=1}^N$ is a martingale difference array. Denote its random predictable covariance by

$$\widetilde{\Sigma}_N := \sum_{i=1}^{N}\mathbb{E}(D_iD_i^\top \mid \mathcal{H}_{i-1}) = \frac{1}{N}\sum_{i\notin\mathcal{P}}(\widehat{\pi}_i^{-1}-1)\boldsymbol{I}_i\boldsymbol{I}_i^\top. \tag{25}$$

For $i < k$, the tower property gives

$$\mathbb{E}(D_iD_k^\top) = \mathbb{E}\big[D_i\mathbb{E}(D_k^\top \mid \mathcal{H}_{k-1})\big] = 0.$$

Using (24), positivity, and $I_{ij}^2 = I_{ij} \le 1$, we obtain

$$\mathbb{E}\big[N\|\widehat{\mu}-\mu\|_2^2\big] = \sum_i \mathbb{E}\|D_i\|_2^2 = \mathbb{E}\,\mathrm{tr}(\widetilde{\Sigma}_N) \le M(\varepsilon^{-1}-1). \tag{26}$$

Markov's inequality gives $\sqrt{N}\|\widehat{\mu}-\mu\|_\infty = O_p(1)$. If every coordinate error is smaller than $\Delta_{\min}/2$, then $\widehat{\mu}_b - \widehat{\mu}_c \ge \Delta_c - |\widehat{\mu}_b-\mu_b| - |\widehat{\mu}_c-\mu_c| > 0$ for every challenger. Hence

$$\Pr(\widehat{b}\neq b) \le \Pr(\|\widehat{\mu}-\mu\|_\infty \ge \Delta_{\min}/2) \le \frac{4M(\varepsilon^{-1}-1)}{N\Delta_{\min}^2} \to 0,$$

since $N \ge B_1$ and Assumption 1 implies $N\Delta_{\min}^2 \to \infty$. □

A.3.2 Asymptotic normality and variance estimation

Proof. Convergence of the predictable covariance. Introduce the deterministic finite-collection matrix

$$\Sigma_N = \left(1-\frac{B_1}{N}\right)\frac{1}{N}\sum_{i=1}^{N}\big\{(\pi_i^*)^{-1}-1\big\}\boldsymbol{I}_i\boldsymbol{I}_i^\top.$$

The item-frequency limits in Assumption 3 and $B_1/N \to \rho_1$ give $\Sigma_N \to \Sigma$ in (9). Decompose the difference from (25) as

$$\begin{aligned}\widetilde{\Sigma}_N - \Sigma_N = {}& \frac{1}{N}\sum_{i\notin\mathcal{P}}\big\{\widehat{\pi}_i^{-1}-(\pi_i^*)^{-1}\big\}\boldsymbol{I}_i\boldsymbol{I}_i^\top \\ & + \frac{1}{N}\sum_{i=1}^{N}\left\{\mathbf{1}\{i\notin\mathcal{P}\} - \left(1-\frac{B_1}{N}\right)\right\}\big\{(\pi_i^*)^{-1}-1\big\}\boldsymbol{I}_i\boldsymbol{I}_i^\top.\end{aligned} \tag{27}$$

For the first term, $\|\boldsymbol{I}_i\boldsymbol{I}_i^\top\|_{\mathrm{op}} = \|\boldsymbol{I}_i\|_2^2 \le M$ and positivity give

$$\left\| \frac{1}{N} \sum_{i \notin \mathcal{P}} \{\widehat{\pi}_i^{-1} - (\pi_i^*)^{-1}\} \boldsymbol{I}_i \boldsymbol{I}_i^\top \right\|_{\mathrm{op}} \le \frac{M}{\varepsilon^2 N} \sum_{i \notin \mathcal{P}} |\widehat{\pi}_i - \pi_i^*| = o_p(1)$$

by (11).

The second term accounts for random pilot membership. For each entry, write $w_{i,jk} = \{(\pi_i^*)^{-1} - 1\} I_{ij} I_{ik}$ and $\overline{w}_{jk} = N^{-1} \sum_i w_{i,jk}$. These are deterministic and $0 \le w_{i,jk} \le \varepsilon^{-1} - 1$. Simple random sampling gives $\mathbb{E}\mathbf{1}\{i \in \mathcal{P}\} = B_1/N$ and, for distinct items $i \ne h$,

$$\mathrm{Cov}(\mathbf{1}\{i \in \mathcal{P}\}, \mathbf{1}\{h \in \mathcal{P}\}) = \frac{B_1(B_1-1)}{N(N-1)} - \frac{B_1^2}{N^2} = -\frac{B_1(N-B_1)}{N^2(N-1)}.$$

Expanding the squared sum using these inclusion moments yields

$$\begin{aligned} &\mathbb{E}\left[\left\{ \frac{1}{N} \sum_i \left(\mathbf{1}\{i \notin \mathcal{P}\} - 1 + \frac{B_1}{N} \right) w_{i,jk} \right\}^2 \right] \\ &\quad = \frac{B_1(1 - B_1/N)}{N^2(N-1)} \sum_i (w_{i,jk} - \overline{w}_{jk})^2 = O(N^{-1}). \end{aligned}$$

Since M is fixed, the second matrix in (27) is $O_p(N^{-1/2})$ in operator norm. Combining the two bounds gives

$$\begin{aligned} \|\widetilde{\Sigma}_N - \Sigma\|_{\mathrm{op}} \le \frac{M}{\varepsilon^2 N} \sum_{i \notin \mathcal{P}} |\widehat{\pi}_i - \pi_i^*| \\ + O_p(N^{-1/2}) + \|\Sigma_N - \Sigma\|_{\mathrm{op}} = o_p(1). \end{aligned} \tag{28}$$

Asymptotic normality. For a fixed $u \in \mathbb{R}^M$, set $X_{N,i} = u^\top D_i$. These are martingale differences, and (25)–(28) imply

$$\sum_i \mathbb{E}(X_{N,i}^2 \mid \mathcal{H}_{i-1}) = u^\top \widetilde{\Sigma}_N u \xrightarrow{p} u^\top \Sigma u. \tag{29}$$

Moreover, the increments satisfy the deterministic bound

$$|X_{N,i}| \le \|u\|_2 \|D_i\|_2 \le \frac{\|u\|_2 \sqrt{M}}{\varepsilon \sqrt{N}}. \tag{30}$$

For every $a > 0$, the bound is smaller than a for all sufficiently large N, so the conditional Lindeberg sum is then exactly zero:

$$\sum_i \mathbb{E}\big[X_{N,i}^2 \mathbf{1}\{|X_{N,i}| > a\} \,\big|\, \mathcal{H}_{i-1} \big] = 0.$$

The martingale central limit theorem (Hall & Heyde, 1980, Chapter 3) therefore gives

$$\sqrt{N}\, u^\top (\widehat{\mu} - \mu) = \sum_i X_{N,i} \xrightarrow{d} \mathcal{N}(0, u^\top \Sigma u).$$

The Cramér–Wold theorem proves the multivariate limit in (12).

Variance estimation and studentization. For every entry (j,k), subtracting the predictable covariance gives

$$\widehat{\Sigma}_{N,jk} - \widetilde{\Sigma}_{N,jk} = \frac{1}{N} \sum_{i \notin \mathcal{P}} \left(\frac{R_i}{\widehat{\pi}_i} - 1 \right) (\widehat{\pi}_i^{-1} - 1) I_{ij} I_{ik}. \tag{31}$$

The summands have zero conditional means and cross moments given $\mathcal{F}_1$. Since $(I_{ij}I_{ik})^2 = I_{ij}I_{ik}$,

$$\begin{aligned} \mathbb{E}\Big[(\widehat{\Sigma}_{N,jk} - \widetilde{\Sigma}_{N,jk})^2 \,\Big|\, \mathcal{F}_1 \Big] &= \frac{1}{N^2} \sum_{i \notin \mathcal{P}} (\widehat{\pi}_i^{-1} - 1)^3 I_{ij} I_{ik} \\ &= \frac{1}{N^2} \sum_{i \notin \mathcal{P}} \frac{(1 - \widehat{\pi}_i)^3}{\widehat{\pi}_i^3} I_{ij} I_{ik} \le \frac{1}{N \varepsilon^3}. \end{aligned}$$

Taking expectations over the pilot and summing the M^2 entries yields

$$\mathbb{E}\|\widehat{\Sigma}_N - \widetilde{\Sigma}_N\|_{\mathrm{op}}^2 \le \mathbb{E}\|\widehat{\Sigma}_N - \widetilde{\Sigma}_N\|_{\mathrm{F}}^2 \le \frac{M^2}{N\varepsilon^3}.$$

Thus $\|\widehat{\Sigma}_N - \widetilde{\Sigma}_N\|_{\mathrm{op}} = O_p(N^{-1/2})$. Together with (28), this proves $\widehat{\Sigma}_N \xrightarrow{p} \Sigma$.

Let $v = \lambda_{\min}(\Sigma) > 0$. Weyl's inequality gives

$$\lambda_{\min}(\widehat{\Sigma}_N) \ge v - \|\widehat{\Sigma}_N - \Sigma\|_{\mathrm{op}},$$

so $\widehat{\Sigma}_N$ is positive definite with probability tending to one. Continuity of the inverse square root on positive-definite matrices implies $\widehat{\Sigma}_N^{-1/2} \xrightarrow{p} \Sigma^{-1/2}$. Slutsky's theorem applied to (12) gives

$$\widehat{\Sigma}_N^{-1/2}\sqrt{N}(\widehat{\mu} - \mu) \xrightarrow{d} \mathcal{N}_M(0, \mathrm{Id}_M), \tag{32}$$

where Id_M is the $M \times M$ identity matrix; the inverse square root can be defined arbitrarily on the event that $\widehat{\Sigma}_N$ is singular.

Marginal and pairwise intervals. For any fixed contrast u with $u^\top \Sigma u > 0$, the scalar limit and variance consistency give

$$\frac{\sqrt{N}\, u^\top(\widehat{\mu} - \mu)}{\sqrt{u^\top \widehat{\Sigma}_N u}} \xrightarrow{d} \mathcal{N}(0,1).$$

The scalar martingale proof only needs this positive contrast variance, even if the full covariance is singular. For $u = e_j$, the variance estimate is $\widehat{\Sigma}_{N,jj}/N = \widehat{\mathrm{se}}_j^2$ in (8). For $u = e_j - e_k$, binary labels imply $(I_{ij} - I_{ik})^2 = \mathbf{1}\{g_{ij} \ne g_{ik}\}$, and thus

$$N\mathrm{se}_{jk}^2 = \frac{1}{N}\sum_{i \notin \mathcal{P}} (\widehat{\pi}_i^{-1} - 1)\mathbf{1}\{g_{ij} \ne g_{ik}\} = u^\top \widetilde{\Sigma}_N u \xrightarrow{p} u^\top \Sigma u.$$

Hence the pairwise interval $(\widehat{\mu}_j - \widehat{\mu}_k) \pm z_{1-\eta/2}\mathrm{se}_{jk}$ also uses a consistent estimate of its deterministic asymptotic variance. Convergence to a standard normal and continuity at $\pm z_{1-\eta/2}$ give coverage tending to $1 - \eta$ for both intervals. □

A.4 Extension to Non-Binary Labels

Write $\ell_{ij} = \ell(g_{ij}, Y_i) \in [0,1]$, $\boldsymbol{\ell}_i = (\ell_{i1}, \ldots, \ell_{iM})^\top$, and $\mu_j = N^{-1}\sum_i \ell_{ij}$. The definitions of b and Δ_c apply to these scores.

Comparison variances and allocation. Independence and $\mathbb{E}[(R_i/\pi_i - 1)^2] = \pi_i^{-1} - 1$ give

$$V_c(\pi) = \frac{1}{N^2}\sum_i (\pi_i^{-1} - 1) a_{i,bc}, \qquad a_{i,jk} = (\ell_{ij} - \ell_{ik})^2. \tag{33}$$

For categorical accuracy, $a_{i,jk} = \mathbf{1}\{g_{ij} \ne g_{ik}\}\mathbf{1}\{Y_i \in \{g_{ij}, g_{ik}\}\}$, whereas $\overline{a}_{i,jk} = \mathbf{1}\{g_{ij} \ne g_{ik}\}$. In general, $a_{i,jk} \le \overline{a}_{i,jk}$ implies $V_c(\pi) \le \overline{V}_c(\pi)$; the imputed coefficient need not provide an upper bound. With discrepancy coefficients A_{ic}, stationarity becomes $\nu = (N^2\pi_i^2)^{-1}\sum_{c \ne b} \alpha_c A_{ic}/\Delta_c^2$, yielding the clipped WPD rule. Uniform feasibility gives improvement in the corresponding criterion. For the true coefficients, the increments $(R_i/\pi_i - 1)(\ell_{ib} - \ell_{ic})$ are bounded by ε^{-1}, so the moderate-deviation calculation in (17)–(20) also applies.

Allocation consistency. Retain the pilot and gap conditions of Assumptions 1–2, using score-based coefficients in the design. Suppose the empirical law of the deterministic vectors $(A_i, \boldsymbol{\ell}_i)$ converges weakly to a law H, where $A_i = (A_{ic})_{c \ne b} \in [0,1]^{M-1}$, and $\varepsilon < \rho_2 < 1 - (1-\varepsilon)H(A = 0)$. For fitted coefficients, require $N^{-1}\sum_{i \notin \mathcal{P}} \max_c |\widehat{A}_{ic} - A_{ic}| = o_p(1)$; this is automatic for the conservative rule. The limiting design has dual objective

$$\mathcal{D}(\alpha, \nu) = \mathbb{E}_H \min_{\varepsilon \le p \le 1} \{w(A)(p^{-1} - 1) + \nu p\} - \nu\rho_2, \qquad w(A) = \sum_{c \ne b} \frac{\alpha_c A_c}{\delta_c^2},$$

with α in the probability simplex and $\nu \geq 0$. Continuity and $\mathcal{D} \leq \varepsilon^{-1} - 1 - \nu(\rho_2 - \varepsilon)$ give a maximizer. Since $\mathcal{D}(\alpha, 0) = 0$ and, at uniform α, its right derivative at $\nu = 0$ is $1 - (1-\varepsilon)H(A = 0) - \rho_2 > 0$, any maximizer has $\nu^* > 0$. Its pointwise minimizer is $\pi^*(A) = \min\{1, \max\{\varepsilon, \sqrt{w^*(A)/\nu^*}\}\}$. Differentiating the dual at (α^*, ν^*) gives

$$\partial_\nu \mathcal{D} = \mathbb{E}_H \pi^*(A) - \rho_2, \qquad \partial_{\alpha_c} \mathcal{D} = \delta_c^{-2} \mathbb{E}_H[A_c\{\pi^*(A)^{-1} - 1\}].$$

The maximizing conditions imply that the budget binds and every comparison is at most their α^*-weighted mean, with equality for positive α_c^*. Thus the primal objective equals the dual value, and (23) gives uniqueness H-almost surely. Fitted coefficients change the normalized criterion uniformly by at most $(\varepsilon^{-1} - 1) \max_c(\Delta_{\min}^2/\widehat{\Delta}_c^2)(N - B_1)^{-1} \sum_{i \notin \mathcal{P}} \max_c |\widehat{A}_{ic} - A_{ic}| = o_p(1)$. Weak convergence and simple random sampling give convergence of empirical averages at each continuous rule $\pi_\theta = (1-\theta)\pi^* + \theta\varepsilon$. The feasibility and quadratic-growth argument in Appendix A.3 therefore proves (11), with $\pi_i^* = \pi^*(A_i)$.

Estimation and inference. Replace correctness vectors by score vectors in (7) and (10). The covariance limit is

$$\Sigma = (1 - \rho_1)\mathbb{E}_H[\{\pi^*(A)^{-1} - 1\}\boldsymbol{\ell}\boldsymbol{\ell}^\top]. \tag{34}$$

The increments $D_i = N^{-1/2}(R_i/\widehat{\pi}_i - 1)\boldsymbol{\ell}_i$ off the pilot, and zero on it, have conditional mean zero and norm at most $\sqrt{M}/(\varepsilon\sqrt{N})$. Hence $\mathbb{E}[N\|\widehat{\mu} - \mu\|_2^2] \leq M(\varepsilon^{-1} - 1)$. The decomposition (27) and allocation consistency give $\widetilde{\Sigma}_N \xrightarrow{p} \Sigma$. Continuity of $\pi^*(A)$ justifies the deterministic second-moment limit; the increment bound verifies Lindeberg. Finally,

$$\mathbb{E}\Big[(\widehat{\Sigma}_{N,jk} - \widetilde{\Sigma}_{N,jk})^2 \,\Big|\, \mathcal{F}_1\Big] = \frac{1}{N^2} \sum_{i \notin \mathcal{P}} \frac{(1 - \widehat{\pi}_i)^3}{\widehat{\pi}_i^3} \ell_{ij}^2 \ell_{ik}^2 \leq \frac{1}{N\varepsilon^3}. \tag{35}$$

Thus (12) holds, and $\Sigma \succ 0$ gives (32) and Wald coverage. Pairwise variance estimates use $(\ell_{ij} - \ell_{ik})^2$ in the IPW sum. Imputation may converge to a misspecified model; inference uses the observed scores.

B Top-K Identification

For $1 \leq K < M$, order the accuracies as $\mu_{(1)} \geq \cdots \geq \mu_{(M)}$ and assume $\mu_{(K)} > \mu_{(K+1)}$. The true top-K set is $\mathcal{T}_K = \{j : \mu_j \geq \mu_{(K)}\}$, and its estimate $\widehat{\mathcal{T}}_K$ contains the K judges with the largest estimated accuracies, using the same tie-breaking rule. Correct selection means $\widehat{\mathcal{T}}_K = \mathcal{T}_K$. The relevant comparisons cross the selection boundary: for $j \in \mathcal{T}_K$ and $k \notin \mathcal{T}_K$, define $\Delta_{jk} = \mu_j - \mu_k > 0$ and $V_{jk}(\pi) = \mathrm{Var}_\pi(\widehat{\mu}_j - \widehat{\mu}_k)$. For binary labels, $V_{jk}(\pi)$ is given by (13) with (b, c) replaced by (j, k).

Applying the quadratic large-deviation criterion in (2) to these cross-boundary comparisons gives the oracle objective

$$\begin{aligned} \max_{\pi} \quad & \min_{\substack{j \in \mathcal{T}_K \\ k \notin \mathcal{T}_K}} \frac{\Delta_{jk}^2}{2V_{jk}(\pi)} \\ \text{subject to} \quad & \sum_{i=1}^N \pi_i \leq B, \qquad \varepsilon \leq \pi_i \leq 1, \quad i = 1, \ldots, N. \end{aligned} \tag{36}$$

The same WPD principle follows: the item score becomes $\sum_{j \in \mathcal{T}_K,\, k \notin \mathcal{T}_K} (\alpha_{jk}^\star/\Delta_{jk}^2)\mathbf{1}\{g_{ij} \neq g_{ik}\}$, with nonnegative dual weights summing to one over these pairs, and the oracle uses the clipped square-root allocation in (5). The two-stage implementation substitutes the pilot top-K set and estimated cross-boundary gaps; estimation and the non-binary discrepancy choices remain unchanged.

C Additional Experiments

C.1 Additional Experimental Details

Simulation data-generating processes. All pools contain $N = 10{,}000$ items and $M = 10$ judges. For $K \in \{2, 5\}$ classes, independently draw Y_i uniformly from $\{0, \ldots, K-1\}$ and standard normal variables $U_i, \epsilon_{i1}, \ldots, \epsilon_{iM}$. Given target accuracies a_j, judge j is correct on item i if

$$\Phi^{-1}(a_j) + \sqrt{\rho}\, U_i + \sqrt{1-\rho}\, \epsilon_{ij} > 0, \tag{37}$$

where Φ is the standard normal distribution function. Correct predictions equal Y_i; incorrect binary predictions equal $1 - Y_i$. The main binary settings use $(a_1, a_2) = (0.850, 0.835)$ and eight equally spaced targets from $a_3 = 0.720$ to $a_{10} = 0.550$, with $\rho \in \{0, 0.30\}$. The original binary settings and all five-class settings use $(0.780, 0.755, 0.740, 0.720, 0.700, 0.680, 0.660, 0.640, 0.620, 0.600)$. Original binary errors have $\rho \in \{0, 0.62\}$. Independent five-class errors have $\rho = 0$ and choose each incorrect prediction uniformly among the four incorrect classes. In the shared-error five-class setting, $\rho = 0.62$ and each incorrect judge independently chooses an item-specific shared wrong label with probability 0.10, otherwise an independent uniformly drawn wrong label. The wrong-label-bloc setting preserves the same truth and correctness arrays but increases this probability to 0.95 for judges $j = 4, \ldots, 10$, retaining 0.10 for the first three judges. Shared wrong labels are themselves uniform over the four incorrect classes, and independent wrong draws may coincide with them.

Simulation configurations. Numerical experiments use total budgets from 500 to 1,500 and pilot sizes $B_1 \in \{100, 200\}$. The original binary and five-class configurations each use one fixed pool and 500 labeling repetitions per pilot. The main binary accuracy profile is evaluated on three pools, each with 500 repetitions per pilot and correlation setting. Reported PCS averages these three pools equally. Plurality ties are broken independently of reference labels and held fixed across labeling repetitions.

WPD implementation. A shared uniform pilot identifies the provisional best judge and estimates its accuracy gaps from the challengers; these quantities remain fixed during the second stage. Conservative WPD uses the pairwise disagreement indicator, while imputation WPD evaluates the squared correctness difference at the plurality prediction. These are the conservative and point-mass imputed discrepancies in Section 5. Final accuracy estimates combine the observed pilot contribution with inverse-probability-weighted second-stage reference labels, with variance estimates matched to the sampling design.

Baseline acquisition rules. Uniform gives each item the same labeling probability. **Hara** (Hara et al., 2024) prioritizes aggregate disagreement across judges. **CODA** (Kay et al., 2025) uses a probabilistic model of judge consensus and disagreement to score reductions in uncertainty about the best candidate. **Model Selector** (Okanovic et al., 2025) scores the expected reduction in the entropy of the best candidate's identity, averaging over possible human labels. All comparisons include the pilot in the total labeling budget.

Real-data panels and reference construction. Table 1 summarizes the complete item-by-judge panels. MMLU17 uses standard binary reference comparisons derived from MMLU; Arena17 uses three-way human response preferences, including ties; RewardBench uses binary reference preferences; and LLMJudge uses exact agreement with four-level human relevance judgments. Candidate identities are fixed across repetitions. MMLU17 and Arena17 share 17 LLMs, RewardBench contains 12 judge configurations, and LLMJudge contains 33 submitted configurations, with distinct prompts or procedures treated as distinct candidates. Full-panel accuracies define the target best judge, while labeling methods access reference labels only for queried items.

Real-data budgets and repetitions. The main pilot is 100 and the sensitivity pilot is 300. Total labeling budgets extend to 2,200 for RewardBench and 4,000 for the other datasets. Each real-data setting has 1,000 repetitions, except RewardBench with 3,000.

PCS and Monte Carlo uncertainty. Empirical PCS is the fraction of repetitions selecting the judge with maximal full-panel accuracy. Shaded bands are pointwise 95% Wilson intervals for a single pool. For the three-pool binary validation, we average the pool-specific PCS estimates and

Table 1: Retained real-data panels. The gap is the difference between the two highest full-panel accuracies, in percentage points.

Panel	Items N	Judges M	Classes K	Gap (pp)
MMLU17	28,084	17	2	2.923
RewardBench	2,305	12	2	0.260
Arena17	49,938	17	3	0.693
LLMJudge	4,420	33	4	0.498
Defect detection	7,858	6	2	0.305

use a normal interval with standard error $\sqrt{\sum_{s=1}^{3} \widehat{\mathrm{se}}_s^2}/3$, where $\widehat{\mathrm{se}}_s$ is the within-pool Monte Carlo standard error.

Matched-PCS label savings. We fit each method's mean PCS curve by nondecreasing isotonic regression and obtain its inverse $B_m(p)$ by linear interpolation. Figures 4 and 7 report $100\{1 - B_m(p)/B_{\mathrm{Uniform}}(p)\}$, including pilot cost, over targets attained by both methods. These point estimates directly compare labeling efficiency at matched selection performance; negative values indicate higher label cost than Uniform.

Allocation and accuracy diagnostics. Figure 5 uses pilot 100 with budget 900 on RewardBench and 2,000 on Arena17, over 3,000 and 1,000 repetitions, respectively. Its disjoint retrospective categories are: "Best–closest disagree," where the full-panel best and runner-up predict differently; "Other disagreement," where those two agree but another judge differs; and "All agree," where predictions are unanimous. Allocation shares count post-pilot acquisitions, and enrichment divides allocation share by pool share. Figure 9 shows one repetition with pilot 100 and budgets 500 for MMLU17, 1,500 for RewardBench, and 2,000 elsewhere. It displays individual 95% Wald accuracy intervals, with full-panel accuracies determining judge order and reference diamonds.

Confidence-interval coverage. We assess nominal 95% Wald intervals for judge accuracies and pairwise accuracy gaps in the five main numerical settings. Empirical coverage records how often the intervals contain the corresponding full-pool accuracies or gaps over repeated labeling runs. Marginal intervals use (8), and pairwise intervals use the covariance estimator in (10). The comparison includes Uniform and the applicable WPD variants across the displayed labeling budgets.

C.2 Additional Experimental Results

Accuracy profiles and pilot sensitivity in simulations. Figure 6 shows that WPD's PCS advantage persists with pilot 200 and with the original binary accuracy profile, which has more competitive judges than the two-leader configuration. At pilot 100 and $B = 1000$, original independent-binary PCS is 92.0% for WPD, versus 87.4%, 86.2%, and 86.4% for Uniform, Hara, and CODA; under shared errors, the corresponding values are 98.4%, 92.4%, 96.4%, and 97.0%. The independent five-class curves start lower with pilot 200 than with pilot 100, illustrating the trade-off between pilot estimation and second-stage allocation at a fixed total budget.

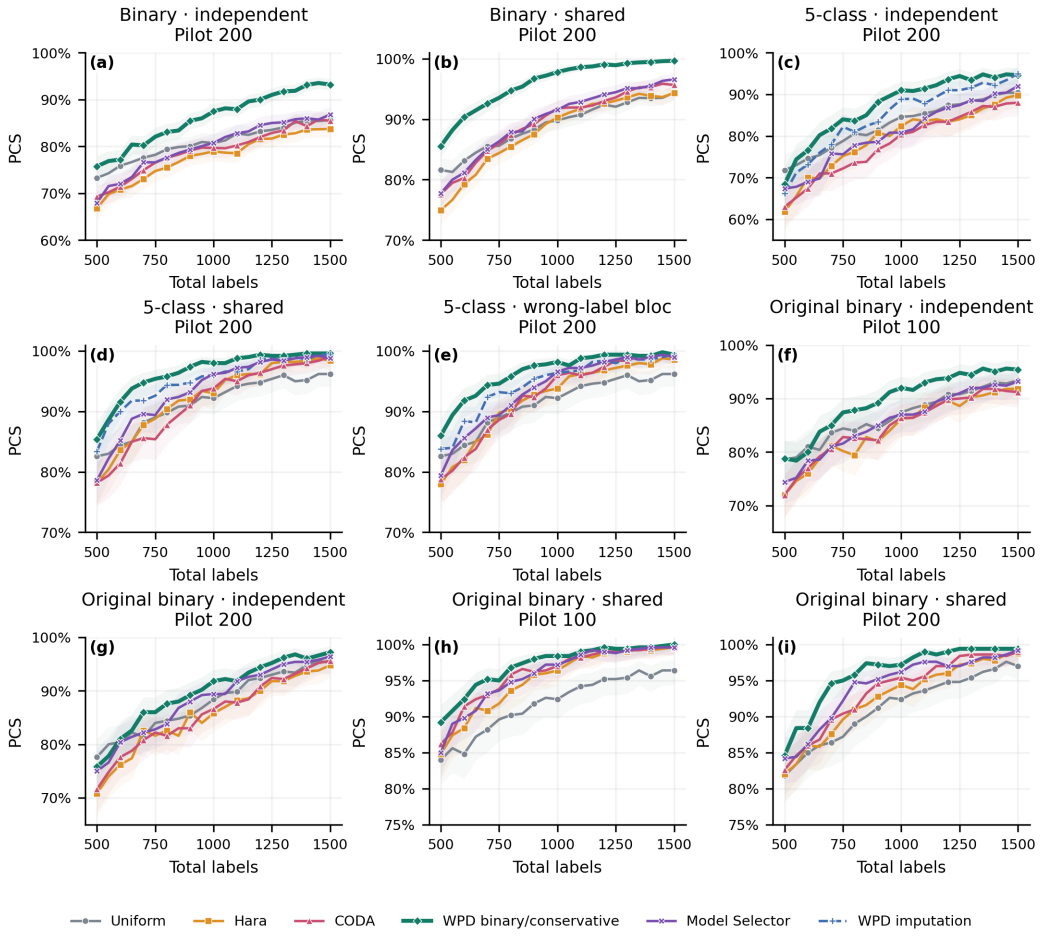

Figure 6: Additional simulation PCS for pilot 200 and the original binary accuracy profile. Bands show pointwise 95% Monte Carlo intervals, as in Figure 2.

Simulation label savings. Figure 7 shows that PCS improvements translate into lower matched-target budgets. In the shared-error binary validation, WPD saves roughly 40–50% relative to Uniform over the displayed common target range. Under independent five-class errors, conservative WPD typically saves more labels than imputation, while Hara and CODA require more labels than Uniform across much of the attainable range. Under shared five-class errors, the two WPD variants achieve similar label savings, and all adaptive methods reduce label cost relative to Uniform.



Figure 7: Simulation label savings relative to Uniform with pilot 100. Values use isotonic mean-PCS inversion over targets attained by both methods; negative values indicate greater label cost than Uniform.

Real-data savings across target levels. Figure 4 shows label savings across target PCS levels for each dataset. WPD's advantage is sustained across broad target ranges on RewardBench and LLMJudge. On Arena17 the size of the savings varies more with the target, while defect detection exhibits substantial savings over the common attainable range below 80% PCS.

Real-data pilot sensitivity. Figure 8 shows higher PCS for WPD with pilot 300. On LLMJudge at $B = 2000$, conservative WPD reaches 88.6% PCS and imputation reaches 85.3%, compared with 67.9% for Uniform, 75.4% for Hara, and 76.9% for CODA. On RewardBench at 700 expected labels, WPD reaches 78.6%, versus 51.2%, 62.1%, and 65.6%, respectively. Conservative WPD also separates more visibly from imputation on Arena17 at larger budgets, whereas CODA and WPD are close on defect detection.

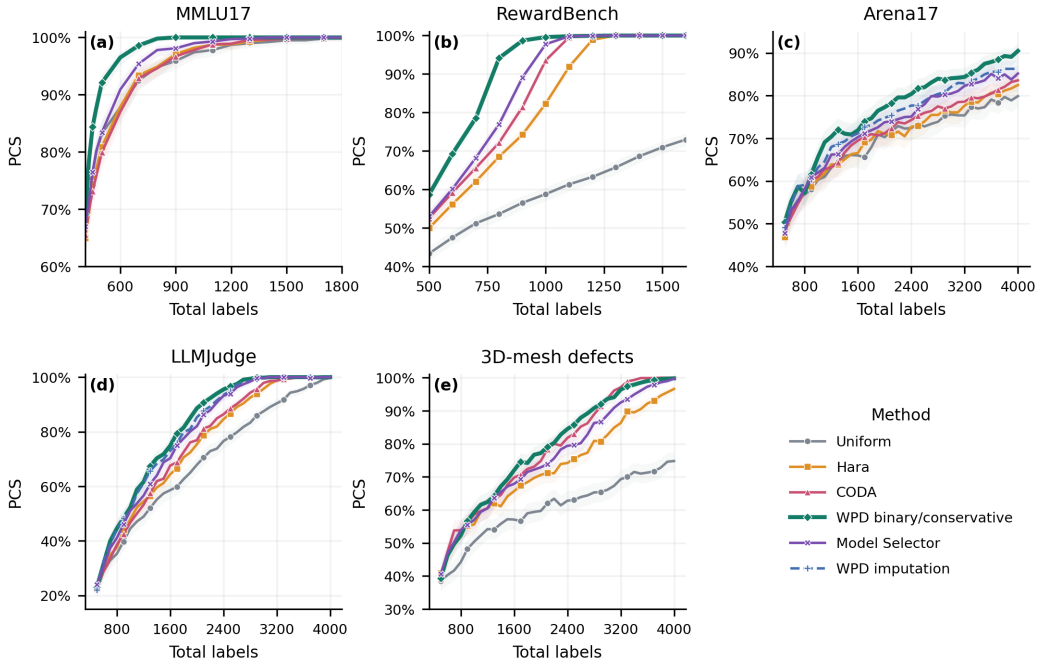

Figure 8: Real-data PCS with pilot 300. The same five datasets and pointwise 95% Monte Carlo intervals as Figure 3 are used.

Individual judge-accuracy estimates. Figure 9 shows overlapping individual accuracy intervals, particularly among the leading LLMJudge configurations and the closely ranked defect judges; several Arena17 estimates also move together below their full-panel reference values in this repetition. These patterns highlight the importance of the joint accuracy estimates: shared item labels induce covariance across judges, which enters the variance of each pairwise comparison.

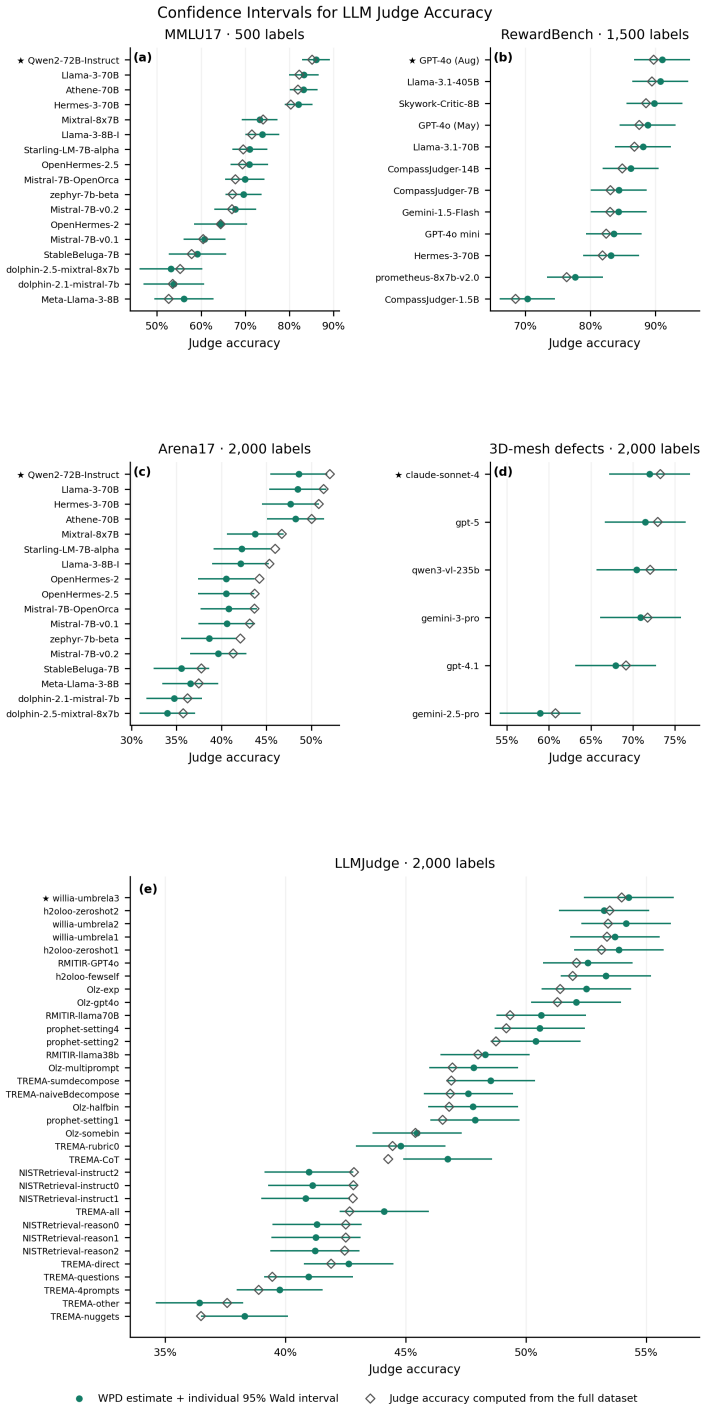

Figure 9: Individual 95% Wald accuracy intervals from one labeling repetition, with full-panel accuracies shown as open diamonds. The display uses WPD for binary panels and conservative WPD otherwise.

Confidence-interval coverage. Figure 10 shows empirical coverage close to the nominal 95% level for both judge accuracies and pairwise accuracy gaps across the numerical settings. The WPD curves remain close to the Uniform reference as the labeling budget changes. These results support the accuracy of the variance estimates under adaptive labeling and the use of confidence intervals to assess performance differences between judges, alongside best-judge identification.

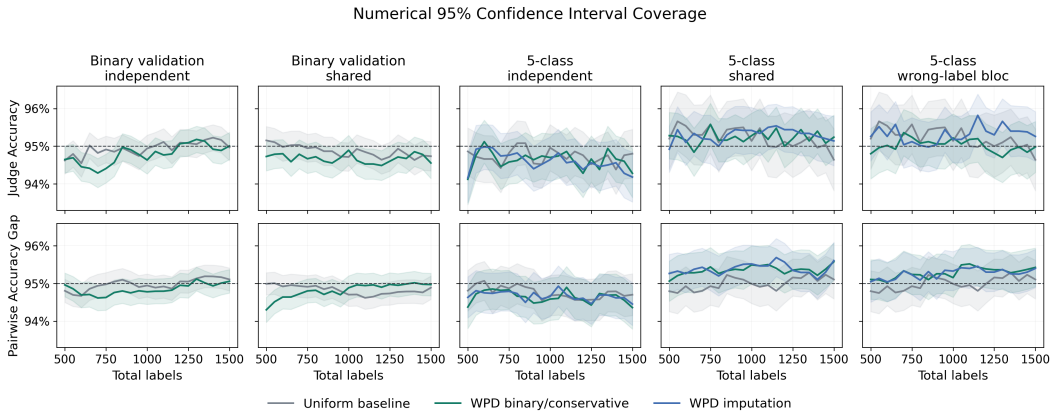

Figure 10: Empirical coverage of nominal 95% confidence intervals in the five main numerical settings. The top row evaluates judge accuracies and the bottom row evaluates pairwise accuracy gaps. The dashed horizontal line marks nominal coverage. Uniform is compared with binary or conservative WPD and, for five-class labels, imputation WPD.