Reliable LLM-as-a-judge Evaluation via Transferable Calibration

Kangjun Noh[1], **Ryunyi Lee**[1], **Heejin Jung**[1], **Young Jin Kim**[2], **Kyungwoo Song**[1†]

[1]Department of Applied Statistics & Data Science, Yonsei University
[2]Microsoft AI

Abstract

LLM-as-a-judge (LaaJ) has become a standard tool for scalable evaluation, but its reliability remains a central concern. Conformal inference (CI) can provide user-specified coverage guarantees for LaaJ outputs, but only under exchangeability. Shifts in evaluated models, topics, or response policies break this assumption and invalidate the coverage guarantee. Existing shift-aware CI methods theoretically suggest transferable calibration, but in LaaJ they face a covariate-nuisance bottleneck. We propose Robust Weighted Conformal Inference (RoWCI), which maps each example to a low-cardinality representation Z, capturing coarse, transferable internal evaluation states. Through Z-state shift correction, RoWCI reuses source calibration data with human scores, without target-domain human scores. We further show that, under covariate shift, RoWCI provides a provable coverage degradation bound in the target domain, even when calibration is transferred without target labels. Empirically, across shifts in evaluated model, response policy, and topic, RoWCI achieves coverage closer to the nominal level with lower variance and competitive interval size than post-hoc LaaJ score calibration baselines. To our knowledge, RoWCI is the first approach to design shift-aware CI for reliable LaaJ evaluation.

1. Introduction

LLM-as-a-judge (LaaJ) evaluation is increasingly needed for large-scale automated workflows (Liu et al., 2023; Zheng et al., 2023; Gu et al., 2025). However, LaaJ often disagrees with human judgments (Szymanski et al., 2024; Shi et al., 2025; Thakur et al., 2025; Bavaresco et al., 2025). This makes reliability a central problem for LaaJ.

Post-hoc calibration methods (Hashemi et al., 2024; Polo et al., 2025; Wang et al., 2026) improve LaaJ score alignment, but usually do not provide user-specified statistical guarantees. This has motivated recent work that applies conformal inference (CI) (Vovk et al., 2005) to LaaJ (Sheng et al., 2025), which provides rigorous statistical guarantees under exchangeability. CI estimates a cutoff from calibration errors, i.e., discrepancies between LaaJ and human-scores, and uses it to cover a human-score of target samples. However, in realistic LaaJ pipelines, evaluated models, prompt topics, system prompts, and response policies change over time (Wu et al., 2024; Richter et al., 2025; Li et al., 2024). These shifts break exchangeability: target judgment errors no longer follow the source calibration-error distribution, so source calibration may fail to achieve target-domain coverage. Thus, reliable LaaJ requires transferable calibration under such shifts. This is practically important because it allows existing human-scored source data, such as web-collected calibration sets, to be reused without expensively collecting target-domain human-scores.

General shift-aware CI methods (Tibshirani et al., 2020; Yang et al., 2022; Jin and Candès, 2023) provide a natural starting point: they reweight source calibration examples so target-like examples have larger influence and recover target-domain coverage. However, they are not directly suited to LaaJ. The covariate is free-form text, where text-space reweighting is unstable. Raw text can easily distinguish domains through nuisance cues such as formatting, response length, model-specific phrasing, or policy-specific style. These cues can dominate importance weights even when they are irrelevant to the judgment error that we want to calibrate. Thus, shift-aware CI for LaaJ needs a transport space that preserves evaluation semantics while suppressing nuisance domain cues. We propose Robust Weighted Conformal Inference (RoWCI) to address this problem. RoWCI turns a complex raw-text shift into a Z-state-frequency shift. Examples that share the same internal judgment situation are calibrated together, even when they come from different models, topics, or response policies. This enables transferable calibration under a common judgment axis, allowing existing human-scored source data to be reused while aiming to preserve the target-domain coverage guarantee, without new target-domain human-scores. We prove a coverage lower bound for the practical RoWCI procedure, making explicit how imperfect state transfer can degrade coverage. Empirically, across shifts in evaluated model, response pol-

icy, and topic, RoWCI achieves coverage closer to the nominal level with lower variance and competitive interval size than post-hoc LaaJ score calibration baselines.

Our contributions are threefold. First, we formulate CI under covariate shift for LaaJ as a problem of reusing human-scored calibration data. Second, we identify and analyze the covariate-nuisance bottleneck of existing shift-aware CI methods in LaaJ. Third, we propose RoWCI, a scalable framework that reweights in a low-cardinality judgment-state space and yields an approximate coverage lower bound. To our knowledge, RoWCI is the first shift-aware CI framework for LaaJ with an approximate target-domain coverage guarantee.

2. Related Work

LaaJ is widely used as a scalable method in open-ended generation evaluations (Zheng et al., 2023; Gu et al., 2025; Liu et al., 2023). Prior work shows that agreement with human annotations varies across tasks and domains, with larger gaps in specialized or difficult cases (Szymanski et al., 2024; Tan et al., 2025; Thakur et al., 2025; Bavaresco et al., 2025). Other studies document systematic bias and instability, including prompt-template sensitivity, position bias, and preference inconsistency (Chen et al., 2024; Shi et al., 2025; Wei et al., 2025; Zheng et al., 2023).

Post-hoc LaaJ score calibration methods aim to align judge scores with human evaluations. LLM-RUBRIC (Hashemi et al., 2024) decomposes evaluation into rubric-level dimensions and learns a calibration network from rubric responses to final human-scores. Bridge (Polo et al., 2025) models the human-LaaJ gap using judge scores and auxiliary covariates. TrustJudge (Wang et al., 2026) mitigates inconsistencies in LLM-as-a-judge outputs by refining the internal scoring behavior of LaaJ. SCI-BA (Sheng et al., 2025) applies conformal inference (Vovk et al., 2005; Angelopoulos and Bates, 2022) to LaaJ scores to construct prediction intervals for human-scores. However, these methods do not explicitly discuss shifts between the calibration and target data distributions.

Our setting is close to conformal inference under covariate shift. Weighted conformal prediction (Tibshirani et al., 2020) and doubly robust calibration (Yang et al., 2022) provide target-domain correction tools with source data that include human-scores and target covariates without human-scores. Related weighted conformal p-value methods (Jin and Candès, 2023) focus on selective inference under covariate shift. LR-QR (Joshi et al., 2025) further targets high-dimensional settings by avoiding explicit density-ratio estimation. These methods provide theoretically improved guarantees or more robust shift-correction models, but they do not explicitly explain how to handle text input, which creates a practical bottleneck for deployment to LaaJ settings.

3. Problem Setup

We view LaaJ evaluation as a scoring problem with a judge function $s_X : \mathcal{X} \to [0,1]$. Each data point is (X, Y), where X is an evaluation text and $Y \in [0,1]$ is a corresponding human-score. We write the judge score and error as $S := s_X(X)$, $E := |S - Y|$. Let $\mathbf{P}_0$ and $\mathbf{P}_1$ denote the source and target distributions. We study a covariate-shift setting in which the marginal distribution of X may change, but the human-scoring law is shared,

$$\mathbf{P}_0(Y \mid X) = \mathbf{P}_1(Y \mid X). \tag{1}$$

This captures shifts in topics, system prompts, evaluated models, and policies which frequently occur in practical deployments. We are given source data with human-scores $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbf{P}_0$ and target covariates without human-scores $\mathcal{D}^c = \{X_j^c\}_{j=1}^{n_c} \overset{\text{i.i.d.}}{\sim} \mathbf{P}_1^X$. We do not observe the target domain human-score Y from $\mathbf{P}_1$. Our goal is to construct a prediction interval $\widehat{\mathcal{I}}_\alpha(X) \subseteq [0,1]$ for $(X, Y) \sim \mathbf{P}_1$ such that

$$\mathbb{P}_{\mathbf{P}_1}\big(Y \in \widehat{\mathcal{I}}_\alpha(X)\big) \geq 1 - \alpha,$$

where $\alpha \in (0,1)$ is a user-specified error rate. Later, we introduce a structured state $Z = z(X)$ for shift-aware calibration. The ideal Z-space result (Section 4.2) still calibrates the raw error $E = |s_X(X) - Y|$, while a state-only scorer $s_Z : \mathcal{Z} \to [0,1]$ and its judgment error $E^Z = |s_Z(Z) - Y|$ are used in the practical RoWCI procedure (Section 4.3).

4. RoWCI: Robust Weighted Conformal Inference

Throughout this section, we work under equation (1). We first review WCI with raw covariate X and oracle weights, then explain why raw-text reweighting is unstable in LaaJ. Finally, we introduce RoWCI, which performs reweighting in the extracted Z-space.

4.1. Weighted Conformal Inference with Raw X

Weighted conformal inference (WCI) (Tibshirani et al., 2020) extends split conformal inference (Appendix D.1) to covariate-shift settings. Under shift, the unweighted source quantile may no longer reflect the target error distribution, so the cutoff can be miscalibrated. WCI corrects this by upweighting target-like calibration points. The target-likeness of X is measured by the importance weight $w(X) = \frac{d\mathbf{P}_1^X}{d\mathbf{P}_0^X}(X)$. Specifically, set $w_i = w(X_i)$ for each calibration point (X_i, Y_i) and $w_{n+1} = w(X_{n+1})$

for the test point, and normalize them into probabilities $Q_i = w_i/(\sum_{l=1}^{n} w_l + w_{n+1})$, so that target-like calibration points carry larger mass. The weighted empirical distribution $\sum_{i=1}^{n} Q_i \mathbf{1}\{x = E_i\} + Q_{n+1}\mathbf{1}\{x = \infty\}$ then places a mass Q_i on each observed error E_i, and a conservative placeholder $\mathbf{1}\{x = \infty\}$ at the test point whose error is unobserved. We define $\hat{t}_\alpha(X_{n+1})$ as the $(1-\alpha)$ quantile of this weighted distribution and set $\hat{\mathcal{I}}_\alpha(X_{n+1}) = [s_X(X_{n+1}) - \hat{t}_\alpha(X_{n+1}), s_X(X_{n+1}) + \hat{t}_\alpha(X_{n+1})] \cap [0,1]$.

Proposition 1 (WCI validity with the exact weights)**.** *Assume equation (1), $\mathbf{P}_1^X \ll \mathbf{P}_0^X$, and $0 < w(X) < \infty$ almost surely under $\mathbf{P}_1$. Let $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbf{P}_0$, and let $(X_{n+1}, Y_{n+1}) \sim \mathbf{P}_1$ be an independent test point. Then the interval $\widehat{\mathcal{I}}_\alpha(X_{n+1})$ defined above using the weighted empirical distribution $\sum_{i=1}^{n} Q_i(X_i)\mathbf{1}\{x = E_i\} + Q_{n+1}(X_{n+1})\mathbf{1}\{x = \infty\}$ with the exact weights $w(X) = \frac{d\mathbf{P}_1^X}{d\mathbf{P}_0^X}(X)$ satisfies*

$$\mathbb{P}_{\mathbf{P}_1}(Y \in \widehat{\mathcal{I}}_\alpha(X_{n+1})) \geq 1 - \alpha.$$

Proposition 1 gives validity in raw X-space. However, the practical difficulty is not validity, but estimating stable weights in text-space. We analyze this issue in the following section.

4.2. Transfer the Law of Judgment Error with Representation Z

As shown in Proposition 1, if the exact importance weight $w(X) = \frac{d\mathbf{P}_1^X}{d\mathbf{P}_0^X}(X)$ is available, WCI recovers target-domain coverage. Thus, the challenge is to estimate the exact weights. In practice, a density-ratio estimator (DRE) can distinguish source and target using many cues in raw text, such as topic, response length, formatting, model-specific phrasing, and so on. These cues are not directly related to the judgment error. Thus, weighting by the exact $w(X)$ then matches source and target not only on error-relevant states, but on residual nuisance variation. This residual matching is unnecessary for transferring the law of the judgment error and can increase finite-sample weight variance.

Our intuition is that target-domain coverage only requires transferring the part of X relevant to judgment error E, not the entire distribution of X. To formalize this, define a representation $Z = z(X)$ that sufficiently transfers the law of E.

$$\mathbf{P}_0(E \mid Z) = \mathbf{P}_1(E \mid Z). \tag{2}$$

This means that source and target share the same conditional distribution of E, given Z. Assume $\mathbf{P}_1^Z \ll \mathbf{P}_0^Z$ and let $w(X) = \frac{d\mathbf{P}_1^X}{d\mathbf{P}_0^X}(X)$, $w(Z) = \frac{d\mathbf{P}_1^Z}{d\mathbf{P}_0^Z}(Z)$ denote the raw importance weight and the Z-level importance weight.

Theorem 1 (Target validity from Z-space reweighting)**.** *Assume equation (2) is satisfied, where $E := |s_X(X) - Y|$. Let $\hat{t}_\alpha^Z(X)$ be the weighted conformal cutoff obtained from calibration scores $E_i = |s_X(X_i) - Y_i|$ by replacing the raw weights $w(X_i)$ and $w(X_{n+1})$ in Proposition 1 with $w(Z_i)$ and $w(Z_{n+1})$. Define $\widehat{\mathcal{I}}_\alpha^Z(X_{n+1}) = [s_X(X_{n+1}) - \hat{t}_\alpha^Z(X_{n+1}),\ s_X(X_{n+1}) + \hat{t}_\alpha^Z(X_{n+1})] \cap [0,1]$.*

Then

$$\mathbb{P}_{\mathbf{P}_1}\Big(Y \in \widehat{\mathcal{I}}_\alpha^Z(X_{n+1})\Big) \geq 1 - \alpha.$$

Theorem 1 shows that any Z and $w(Z)$ satisfying equation (2) are sufficient for target-domain coverage. This opens a route bypassing the information excess of X. To analyze the benefit of using $w(Z)$ instead of $w(X)$, let the residual part be $U = u(X)$ so X is almost surely equivalent to (Z, U).

Theorem 2 (Nuisance-induced variance in raw importance weights)**.** *Suppose $X \equiv (Z, U)$ almost surely, where $U = u(X)$ is a residual coordinate. Assume $\mathbf{P}_1^Z \ll \mathbf{P}_0^Z$, $\mathbf{P}_1^{U|Z} \ll \mathbf{P}_0^{U|Z}$ and the involved weights have finite second moments under $\mathbf{P}_0$. Then*

$$\begin{aligned} w(X) &= w(Z)\, w(U \mid Z), \\ w(U \mid Z) &= \frac{d\mathbf{P}_1^{U|Z}}{d\mathbf{P}_0^{U|Z}}(U \mid Z), \\ w(Z) &= \mathbb{E}_{\mathbf{P}_0}[w(X) \mid Z]. \end{aligned}$$

In particular,

$$\begin{aligned} &\mathrm{Var}_{\mathbf{P}_0}(w(X)) - \mathrm{Var}_{\mathbf{P}_0}(w(Z)) \\ &\qquad = \mathbb{E}_{\mathbf{P}_0}\Big[w(Z)^2\ \mathrm{Var}_{\mathbf{P}_0}(w(U \mid Z) \mid Z)\Big]. \end{aligned}$$

Theorem 2 makes the nuisance issue precise. The residual factor $w(U \mid Z)$ averages to one within each Z-state, so it is not needed to recover the target law of E once equation (2) holds. This factor becomes problematic in practice, where $w(X)$ should be estimated by DRE. Z carries the complex semantic information related to judgment error and therefore has difficult patterns, whereas U consists of surface signals such as length, formatting, and model-specific phrasing that trivially separate target from source. Therefore, the estimator can learn $w(U \mid Z)$ rather than $w(Z)$, because easy nuisance signals dominate the learning objective.

Our goal is to design a representation extractor $z(X) : \mathcal{X} \to \mathcal{Z}$ that (i) automatically extracts Z from X, (ii) preserves the judgment error-relevant information required by equation (2) without observing Y or E from target, and (iii) excludes U.

4.3. Robust Weighted Conformal Inference

Recalling the problem setup in Section 3, our resources are the source data with human-scores $\{(X_i, Y_i)\} \sim \mathbf{P}_0$ and the

target covariates without human-scores $\{X_j^c\} \sim \mathbf{P}_1^X$; target-domain Y is unobserved. Training an extractor $z : \mathcal{X} \to \mathcal{Z}$ that simultaneously satisfies (ii) and (iii) is hard in our setup. We cannot identify which features come from the state Z or the residual U, since we cannot observe the target Y. Thus, training z requires either target Y observations or a strong assumption on the learned representation (Ben-David et al., 2010; Arjovsky et al., 2020).

Instead of learning Z using target-domain Y under impractical assumptions, we define Z as an internal judgment state. Source and target examples may differ in format, response style, or evaluated model, but under the same judgment axis they can share the same scoring situation. For example, a source-domain coding answer and a target-domain travel plan answer may be different in raw text, but both can belong to the same Z-state if they are in a state such as "partially performed while missing an important requirement." In other words, Z maps X from different domains into a common situation.

Here, Z is not designed to determine Y. The human-score Y may still vary within a Z-state because of fine-grained severity or annotator variation. RoWCI only needs the weaker stability condition that the remaining Y-distribution in a Z-state is similar across source and target domains. That is, $\mathbf{P}_0(Y \mid Z) \approx \mathbf{P}_1(Y \mid Z)$. A state Z satisfying this condition naturally connects to equation (2) when the base score is changed from $s_X(X)$ to $s_Z(Z)$. When z is fixed, the variation of $E^Z = |s_Z(Z) - Y|$ comes from the variation of $Y \mid Z = z$, so it leads to $\mathbf{P}_0(E^Z \mid Z) \approx \mathbf{P}_1(E^Z \mid Z)$. Therefore, after converting X to Z and computing $s_Z(Z)$, we obtain the coverage guarantee by calibrating the residual judgment error $E^Z = |s_Z(Z) - Y|$ using WCI with $w(Z)$. This structural design preserves the judgment-relevant information needed for calibration, satisfying (ii), while removing nuisance variables such as style and format before density-ratio estimation, satisfying (iii).

The next step is to implement an automated extractor $z(X)$ for (i). We define low-cardinality subquestions about X and collect their answers into a discrete vector Z. Each Z-cell represents one internal state. These subquestions are designed to retain the judgment information needed to keep the remaining Y-distribution comparable across source and target within the same Z-cell. Intuitively, once the major axis-specific reasons for score variation are fixed by the answer vector, the remaining variation within a cell is treated as residual severity or annotator variation rather than a domain-specific effect. Appendix A provides examples of such subquestions for judgment axes. The combination of the answers then serves as an observable proxy state for approximating the condition $\mathbf{P}_0(Y \mid Z) \approx \mathbf{P}_1(Y \mid Z)$. These answers need not be perfectly accurate. Because $z(X)$ maps X to an internal state rather than predicting Y, the requirement is consistency: source and target examples in the same internal state should map to the same Z-cell. Under this requirement, systematic but domain-stable bias in $z(X)$ can be tolerated. We further explain this in Theorem 3 and Appendix A.1.

Algorithm 1 shows the entire procedure of RoWCI. It has two steps: Z-extraction and weighted conformal inference (WCI). In the Z-extraction step, for each input X, we compute $\hat{Z} = \hat{z}(X)$ and the state-only score $S^{\hat{Z}} = s_Z(\hat{Z})$ (Appendix A.2). In the WCI step, we use the importance weights $\hat{w}(\hat{Z})$ estimated by a binary domain classifier and the weighted calibration errors $E_i^{\hat{Z}} := |s_Z(\hat{Z}_i) - Y_i|$ to construct the conformal cutoff $\hat{t}_\alpha(\hat{Z}_{n+1})$. In the inference step, for a new test input X_{n+1}, we compute $\hat{Z}_{n+1} = \hat{z}(X_{n+1})$ and output

$$\widehat{\mathcal{I}}_\alpha^{\hat{Z}}(X_{n+1}) = \big[s_Z(\hat{Z}_{n+1}) - \hat{t}_\alpha(\hat{Z}_{n+1}), \\ s_Z(\hat{Z}_{n+1}) + \hat{t}_\alpha(\hat{Z}_{n+1})\big] \cap [0,1].$$

We now theoretically analyze the target domain coverage of the practical RoWCI procedure. Let $\widehat{Z} = \widehat{z}(X)$ and let $w(\widehat{z}) := \frac{d\mathbf{P}_1^{\hat{Z}}}{d\mathbf{P}_0^{\hat{Z}}}(\widehat{z})$ denote the density-ratio in the extracted $\hat{Z}$-space. Let $\mathcal{D}_{\text{cal}}$ be the independent calibration split and set $n_{\text{cal}} := |\mathcal{D}_{\text{cal}}|$. Let $\mathcal{D}_{\text{dre}}$ be the source-target covariate set used for DRE; write m_0 and m_1 for the numbers of source and target covariates in $\mathcal{D}_{\text{dre}}$, and set $n_{\text{dre}} := m_0 + m_1$. Let $\eta(\hat{z}) := \mathbb{P}(A = 1 \mid \hat{Z} = \hat{z})$, where $A = 0$ denotes source and $A = 1$ denotes target. We estimate η by $\hat{\eta}$ and define the plug-in density ratio as $\hat{w}(\hat{z}) := \frac{m_0}{m_1}\frac{\hat{\eta}(\hat{z})}{1-\hat{\eta}(\hat{z})}$.

Theorem 3 (Coverage degradation bound under extracted guideline states)**.** *Assume equation (1), $\mathbf{P}_1^{\hat{Z}} \ll \mathbf{P}_0^{\hat{Z}}$, and the conditions of Lemma 2. Let $\Delta_{\text{stat}}(n_{\text{dre}}, n_{\text{cal}}, \delta)$ denote an upper bound on the sup-norm deviation between the ideal $\widehat{Z}$-weighted CDF of $E^{\widehat{Z}}$ and its practical estimated-weight empirical counterpart, as in Lemma 2. Define the representation gap by*

$$\varepsilon_{\text{rep}} := \int \mathrm{TV}\Big(\mathbf{P}_0(Y \mid \hat{Z} = \hat{z}), \mathbf{P}_1(Y \mid \hat{Z} = \hat{z})\Big) d\mathbf{P}_1^{\hat{Z}}(\hat{z}).$$

Then there exists a constant $C > 0$ such that, with probability at least $1 - 2\delta$,

$$\mathbb{P}_{\mathbf{P}_1}\Big(Y \in \widehat{\mathcal{I}}_\alpha^{\hat{Z}}(X_{n+1})\Big) \geq 1 - \alpha - \varepsilon_{\text{rep}} \\ - C\,\Delta_{\text{stat}}(n_{\text{dre}}, n_{\text{cal}}, \delta).$$

Theorem 3 separates the coverage loss into two terms. The representation gap ε_{rep} measures how much the source and target conditional laws of Y still differ within the induced $\hat{Z}$-cells. It is a cell-level condition, not a pointwise requirement for the correct subquestion answers. Even if the answers

that define a $\hat{Z}$-cell are systematically different from the oracle internal state, the bias is harmless for coverage as long as the induced cell preserves $\mathbf{P}_0(Y \mid \hat{Z}) \approx \mathbf{P}_1(Y \mid \hat{Z})$. The bias is harmful when it is domain-dependent, since the same $\hat{Z}$-cell may then contain source and target examples with different remaining Y-distributions. The second term $\Delta_{\mathrm{stat}}(n_{\mathrm{dre}}, n_{\mathrm{cal}}, \delta)$ captures DRE and finite-calibration error. Under Lemma 2, it decreases as the sample sizes $n_{\mathrm{dre}}, n_{\mathrm{cal}}$ grow.

5. Empirical Results

We empirically study whether real-world source calibration data with human-scores can be reused under covariate shift while keeping target-domain coverage close to the nominal level. Our primary criterion is coverage accuracy relative to the nominal level $1 - \alpha$; interval or set size is compared after coverage is controlled. We fix the target error rate $\alpha = 0.1$, so the nominal coverage level is 90%. Appendix C provides additional empirical results such as topic shifts and robustness against unnecessary shift correction.

Datasets and covariate shifts. We use HelpSteer 1 (HS1), HelpSteer 2 (HS2), and MentalBench (Wang et al., 2024a;b; Badawi et al., 2026). HS1 and HS2 provide multi-axis Likert-5 ratings for general-domain (e.g., code, math, common knowledge) user-instruction and LLM-response pairs, and HS2 keeps the same schema but includes more difficult instructions and stronger LLM responses. MentalBench provides Likert-5 ratings for mental-health support quality, including axes of empathy, guidance, and informativeness. We treat each axis as an independent task and use its human-score as Y. For HS, we formulate bidirectional version shifts (HS1 to HS2 and the reverse). For MentalBench, we use white-box LLM responses for the calibration set and black-box LLM responses as target sets.

Models and baselines. We compare two groups of baselines with RoWCI. First, post-hoc LaaJ score-alignment methods, including LLM-RUBRIC (Hashemi et al., 2024), Bridge (Polo et al., 2025), and TrustJudge (Wang et al., 2026), aim to align judge scores or score distributions with human-scores. Since they do not directly output conformal intervals, we convert their predicted score distributions into prediction sets under the common coverage-size protocol in Appendix A. We also include SCI-BA (Sheng et al., 2025), a CI baseline that calibrates LaaJ scores, targeting coverage under exchangeability. Second, we compare shift-aware CI baselines, including WCI (Tibshirani et al., 2020), DR (Yang et al., 2022), and LR-QR (Joshi et al., 2025). While these methods are not originally designed for LaaJ, we implement them to take raw text as input by first converting it into rich embeddings using a pre-trained embedding model (Zhang et al., 2025).

5.1. Comparison with Baselines

Comparison with Post-hoc LaaJ Score Calibration Baselines. Table 1 compares RoWCI with post-hoc LaaJ score calibration baselines. These methods improve score alignment or internal judge consistency, but they are not designed to make the prediction set cover the human-score with statistical guarantees. Thus their calibrated probabilities need not represent coverage, especially under distribution shifts. TrustJudge uses log-probabilities to form a distribution-sensitive continuous score and to reduce inconsistencies between single-score and pairwise judgments, but it does not calibrate residual errors against human-scores. As a result, it produces overly small sets and severe undercoverage in our experiments. Bridge fits a human-LaaJ gap model on the source split using the judge score distribution and auxiliary covariates such as length, sentiment, and formatting. When the target changes the relation between these covariates and human disagreement, or when the covariates mainly encode domain style, the learned gap correction can fail. LLM-RUBRIC is a close baseline because it also decomposes evaluation into rubric questions. However, it uses rubric responses to learn a source-supervised rubric-to-human-score aggregator, while RoWCI uses Z as a transport state for residual calibration. LLM-RUBRIC often preserves coverage only with large cumulative-mass sets, whereas RoWCI achieves target coverage more efficiently by calibrating residuals in Z-space. SCI-BA provides conformal intervals, but it calibrates on the source error distribution and relies on exchangeability. RoWCI instead uses Z not as a point predictor of Y, but as the transport state for residual calibration; once the Z-conditioned error law is stable, reweighting and conformal calibration yield stable coverage under shift.

Comparison with Shift-aware Conformal Inference Baselines. Table 1 first shows why shift-aware CI is needed: source-calibrated SCI-BA often under- or over-covers after shifts. Table 2 shows that standard shift-aware CI baselines share a LaaJ-specific failure mode. They take rich text embeddings as covariates, so the shift-correction model can be driven by nuisance cues such as topic, length, formatting, model-specific phrasing, or response policy. These cues separate source from target, but they do not necessarily determine the judgment error. Thus, the correction direction is not aligned with the target error shift. WCI estimates a density-ratio and upweights target-like source residuals. In text-space, this ratio often becomes a nuisance ratio, leading either to ineffective correction or high-variance cutoffs. DR is theoretically stronger because it combines the density-ratio with a model for the conditional error distribution and has a doubly robust structure, but in LaaJ both nuisance components are estimated from the same nuisance-rich text representation. LR-QR avoids explicit density-ratio estimation by learning an input-dependent threshold with

Table 1. Empirical coverage and size for post-hoc calibration baselines, SCI-BA, and RoWCI under covariate shift on HelpSteer and MentalBench across Qwen-3-30B and GPT-5.5 LaaJ backbones. Methods requiring internal log-probabilities are omitted for GPT-5.5. Each entry reports mean ± standard deviation over 20 runs. Coverage is reported on a 0-100% scale, size is normalized to $[0, 1]$. We prioritize coverage relative to the nominal 90% level; among methods that satisfy coverage, smaller size is preferred. Overall, RoWCI is the most stable coverage-oriented calibration method among the compared baselines.

HelpSteer

	Cal: HS1 | Target: HS2						Cal: HS2 | Target: HS1					
	Correctness		Helpfulness		Complexity		Correctness		Helpfulness		Complexity	
Qwen-3-30B	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**
LLM-RUBRIC	91%±0%	0.71±0.01	89%±0%	0.69±0.00	96%±0%	0.78±0.00	96%±0%	0.81±0.01	96%±0%	0.81±0.01	92%±1%	0.57±0.01
Bridge	82%±2%	0.57±0.03	80%±1%	0.56±0.00	86%±2%	0.58±0.01	99%±0%	0.79±0.01	99%±0%	0.84±0.01	94%±1%	0.64±0.00
TrustJudge	44%±1%	0.22±0.00	41%±1%	0.22±0.00	25%±1%	0.21±0.00	46%±1%	0.21±0.00	47%±2%	0.21±0.00	34%±1%	0.21±0.00
SCI-BA	84%±1%	0.49±0.01	81%±1%	0.45±0.02	78%±1%	0.33±0.01	95%±0%	0.61±0.01	97%±1%	0.52±0.02	96%±1%	0.67±0.02
RoWCI	**91%±1%**	**0.52±0.02**	**92%±1%**	**0.40±0.01**	**93%±1%**	**0.41±0.04**	**92%±1%**	**0.46±0.01**	**92%±1%**	**0.48±0.01**	88%±1%	0.38±0.01
GPT-5.5	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**
Bridge	84%±0%	0.54±0.00	82%±0%	0.54±0.00	96%±0%	0.70±0.00	97%±0%	0.72±0.00	97%±1%	0.75±0.03	95%±0%	0.65±0.00
SCI-BA	83%±2%	0.55±0.00	81%±1%	0.55±0.00	95%±0%	0.56±0.00	98%±0%	0.73±0.00	98%±0%	0.75±0.00	85%±1%	0.43±0.00
RoWCI	**90%±1%**	**0.32±0.01**	**91%±1%**	**0.29±0.01**	**91%±2%**	**0.28±0.05**	**92%±2%**	**0.34±0.01**	**95%±4%**	**0.46±0.01**	**90%±1%**	**0.33±0.01**

MentalBench

	Cal: White-box Light LLMs | Target: Claude-3.5-Haiku						Cal: White-box Light LLMs | Target: Gemini-2.0-Flash					
	Empathy		Guidance		Informativeness		Empathy		Guidance		Informativeness	
Qwen-3-30B	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**
LLM-RUBRIC	97%±0%	0.81±0.01	92%±0%	0.73±0.00	98%±0%	0.75±0.00	100%±0%	0.81±0.00	94%±0%	0.67±0.00	99%±1%	0.82±0.00
Bridge	89%±9%	0.55±0.07	95%±4%	0.74±0.06	97%±1%	0.72±0.00	91%±16%	0.60±0.06	98%±4%	0.76±0.05	100%±0%	0.65±0.03
TrustJudge	31%±2%	0.21±0.00	33%±1%	0.23±0.00	33%±1%	0.21±0.00	45%±1%	0.20±0.00	47%±2%	0.21±0.00	69%±1%	0.20±0.00
SCI-BA	81%±1%	0.33±0.02	93%±1%	0.55±0.01	92%±1%	0.45±0.03	87%±1%	0.34±0.02	**92%±1%**	**0.48±0.01**	97%±1%	0.46±0.01
RoWCI	**90%±1%**	**0.39±0.11**	**91%±1%**	**0.40±0.01**	**92%±1%**	**0.38±0.00**	**92%±1%**	**0.36±0.00**	93%±1%	**0.40±0.04**	**90%±1%**	**0.37±0.06**
GPT-5.5	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**
Bridge	97%±0%	0.60±0.00	97%±1%	0.76±0.01	98%±0%	0.76±0.00	100%±0%	0.80±0.00	98%±0%	0.64±0.07	99%±0%	0.63±0.00
SCI-BA	86%±1%	0.51±0.00	98%±0%	0.76±0.00	98%±1%	0.73±0.00	90%±0%	0.48±0.00	98%±0%	0.74±0.00	97%±0%	0.71±0.00
RoWCI	**92%±2%**	**0.32±0.02**	**90%±2%**	**0.33±0.05**	**91%±2%**	**0.37±0.03**	**91%±1%**	**0.35±0.01**	**91%±1%**	**0.31±0.02**	**92%±1%**	**0.32±0.01**

likelihood-ratio regularization, yet the text embedding still contains too many domain signals for the regularized linear model to isolate the error-relevant shift. RoWCI addresses this common failure mode before reweighting: Z is a low-cardinality judgment state constructed from scoring reasons, so it structurally removes surface nuisance while retaining the information needed for residual calibration. This is why RoWCI consistently moves coverage toward nominal with low variance, and it motivates the following qualitative analysis of $\hat{w}(X)$ versus $\hat{w}(\hat{Z})$.

5.2. Ablation Studies

Illustration of $w(X)$ instability and $w(Z)$ stability. Table 2 shows the stability difference clearly on the HelpSteer helpfulness transfer from HS1 to HS2. WCI moves in the wrong direction and becomes more variable, dropping from the source-calibrated CI baseline $81\% \pm 1\%$ to $76\% \pm 6\%$ instead of approaching $90\% \pm 0\%$. By contrast, RoWCI achieves the nominal coverage, reaching $92\% \pm 1\%$ at size 0.40 ± 0.00. Figure 1 explains this gap. Reweighting on raw X concentrates on a few nuisance-dominated samples and produces an overly optimistic local error CDF, whereas reweighting on Z stabilizes the weights and yields a cutoff closer to the target.

Empirical conditional invariance of $Y \mid \hat{Z}$. We examine whether the extracted $\hat{Z}$ empirically satisfies the Y-invariance condition. For the helpfulness axis of HelpSteer, we group examples by identical extracted states. Let $\hat{Z}^{(k)}$ denote the k-th most frequent extracted state, and let the corresponding cell be $C_k = \{\hat{Z} = \hat{z}^{(k)}\}$. We compare the empirical Y-distributions of HS1 and HS2 within each cell C_k. Figure 2 shows that for each cell C_k, the Y-distributions are aligned across domains. The two domains usually share the same dominant score region and similar shape. Since $S^{\hat{Z}}$ is a deterministic score computed from $\hat{Z}$, stability of $Y \mid \hat{Z}$ also implies stability of the error distribution $E^{\hat{Z}} \mid \hat{Z}$ up to the same representation gap.

Performance gain decomposition. Table 3 decomposes the roles of point prediction, source-calibrated CI, and shift-corrected CI. The state-only scorer $s_Z(Z)$ improves point accuracy over direct raw-text scoring, showing that $\hat{Z}$ pro-

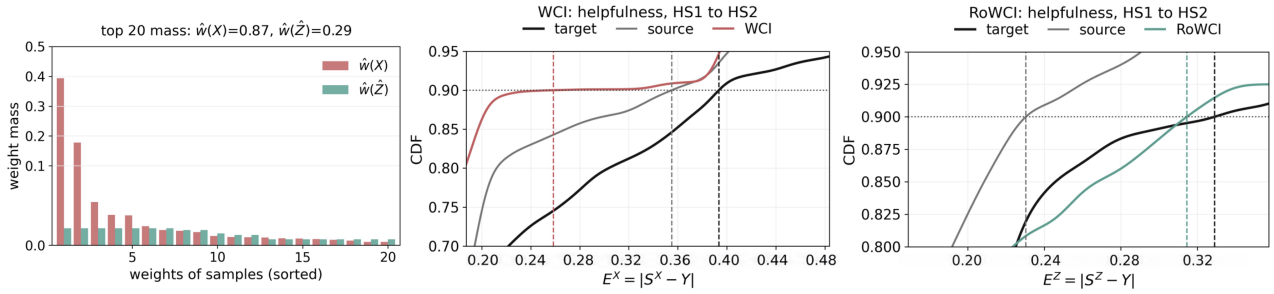
Table 2. Empirical coverage and size for Qwen-3-30B under covariate shift on HelpSteer and MentalBench. We compare baselines for conformal inference under covariate shift with RoWCI. Source-cal denotes unweighted conformal calibrated on the source split. Each entry reports mean ± standard deviation over 20 runs. Metrics are reported following the same rules as in Table 1.

HelpSteer

	Cal: HS1 | Target: HS2						Cal: HS2 | Target: HS1					
	Correctness		Helpfulness		Complexity		Correctness		Helpfulness		Complexity	
Method	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**
Source-cal	84%±1%	0.49±0.01	81%±1%	0.45±0.02	78%±1%	0.33±0.01	95%±0%	0.61±0.01	97%±1%	0.52±0.02	96%±1%	0.67±0.02
WCI	81%±3%	0.54±0.02	76%±6%	0.44±0.07	79%±3%	0.53±0.05	94%±2%	0.55±0.05	95%±3%	0.52±0.05	96%±2%	0.67±0.03
DR	80%±5%	0.43±0.07	80%±6%	0.43±0.06	84%±7%	0.58±0.06	89%±4%	0.50±0.05	93%±3%	0.49±0.04	93%±4%	0.62±0.07
LR-QR	71%±14%	0.46±0.13	74%±17%	0.54±0.14	71%±22%	0.54±0.18	74%±13%	0.49±0.12	62%±17%	0.42±0.14	79%±10%	0.58±0.08
RoWCI	**91%±1%**	**0.52±0.02**	**92%±1%**	**0.40±0.01**	**93%±1%**	**0.41±0.04**	**92%±1%**	**0.46±0.01**	**92%±1%**	**0.48±0.01**	88%±1%	0.38±0.01

MentalBench

	Cal: White-box Light LLMs | Target: Claude-3.5-Haiku						Cal: White-box Light LLMs | Target: Gemini-2.0-Flash					
	Empathy		Guidance		Informativeness		Empathy		Guidance		Informativeness	
Method	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**
Source-cal	81%±1%	0.33±0.02	93%±1%	0.55±0.01	91%±1%	0.45±0.03	87%±1%	0.34±0.02	92%±1%	0.48±0.01	97%±1%	0.46±0.01
WCI	88%±7%	0.46±0.01	91%±3%	0.60±0.02	95%±3%	0.45±0.01	90%±4%	0.42±0.01	91%±1%	0.57±0.01	97%±1%	0.46±0.01
DR	93%±4%	0.49±0.04	95%±2%	0.63±0.03	96%±2%	0.50±0.04	89%±4%	0.40±0.03	92%±2%	0.59±0.02	97%±1%	0.44±0.03
LR-QR	70%±7%	0.49±0.06	69%±5%	0.52±0.05	70%±6%	0.47±0.07	69%±14%	0.38±0.19	69%±24%	0.52±0.21	70%±19%	0.39±0.24
RoWCI	**90%±1%**	**0.39±0.11**	**91%±1%**	**0.40±0.01**	**92%±1%**	**0.38±0.00**	**92%±1%**	**0.36±0.00**	**93%±1%**	**0.40±0.04**	**90%±1%**	**0.37±0.06**



(a) Unstable $\hat{w}(X)$, stable $\hat{w}(Z)$ (b) Reweighted Local E-CDF with $\hat{w}(X)$ (c) Reweighted Local E-CDF with $\hat{w}(Z)$

Figure 1. **Why use Z instead of raw X.** The Local E-CDF means the empirical CDF of E in the neighborhood of the target quantile, which determines the cutoff. **(a)** The raw weights $\hat{w}(X)$ are highly concentrated on a few samples, whereas $\hat{w}(Z)$ are more stable, illustrating the variance gap predicted by Theorem 2. **(b)** Reweighting with $w(X)$ is driven by nuisance and performs correction in the opposite direction. **(c)** Reweighting with $w(Z)$ yields a Local E-CDF closer to the target and a cutoff nearer to the target cutoff. The improvement comes from removing nuisance-dominated weighting instability while retaining the cues that are relevant to judgement error.

Table 3. **Performance gain decomposition** on HelpSteer correctness and helpfulness using Qwen-3-30B. Calibration is performed on HS1 and evaluation is performed on HS2. The point scorers $s_X(X)$ and $s_Z(Z)$ report accuracy after predicting the nearest discrete score. CI denotes source-calibrated conformal inference, while WCI and RoWCI use importance weights estimated from X and Z, respectively. Overall, Z improves point prediction, but target-domain coverage is recovered only after residual calibration is reweighted in Z-space.

Axis	**Raw X variants**					**Z variants**				
	$s_X(X)$	CI		WCI		$s_Z(Z)$	CI		RoWCI	
	Acc.	Cov.	Size	Cov.	Size	Acc.	Cov.	Size	Cov.	Size
Correctness	34%	84%	0.49	81%	0.54	63%	86%	0.42	**91%**	0.52
Helpfulness	27%	81%	0.45	76%	0.44	71%	82%	0.35	**92%**	0.40

vides a useful signal for human-score prediction. However, $s_Z(Z)$ alone remains insufficient, and source-calibrated CI still undercovers under the shift. Raw-X WCI also fails to correct the shift reliably, as its weights are affected by nuisance signals. RoWCI instead performs residual calibration in Z-space and reaches nominal-or-better coverage on both axes, with 91% coverage at size 0.52 on correctness and 92% coverage at size 0.40 on helpfulness.

Proxy diagnosis of error terms and coverage–size trade-off. We further diagnose Theorem 3 on the HS1-to-HS2 under the correctness-axis. Since the terms in theorem are conservative worst-case quantities, we use two empirical proxies with the same purpose as the two degra-

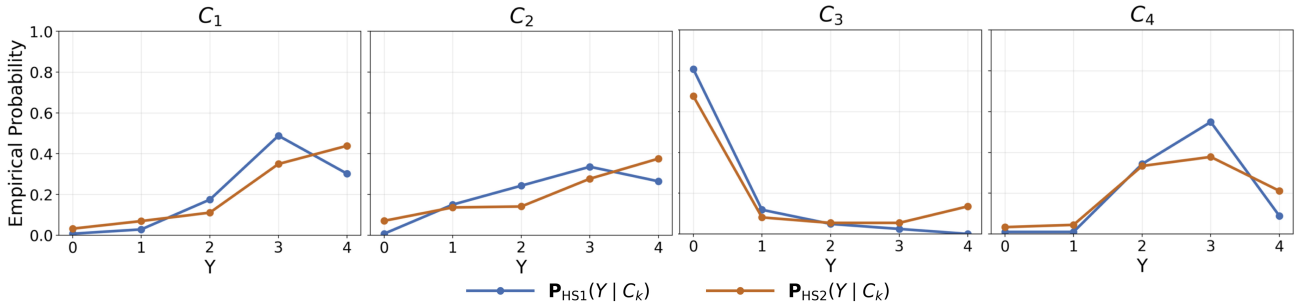
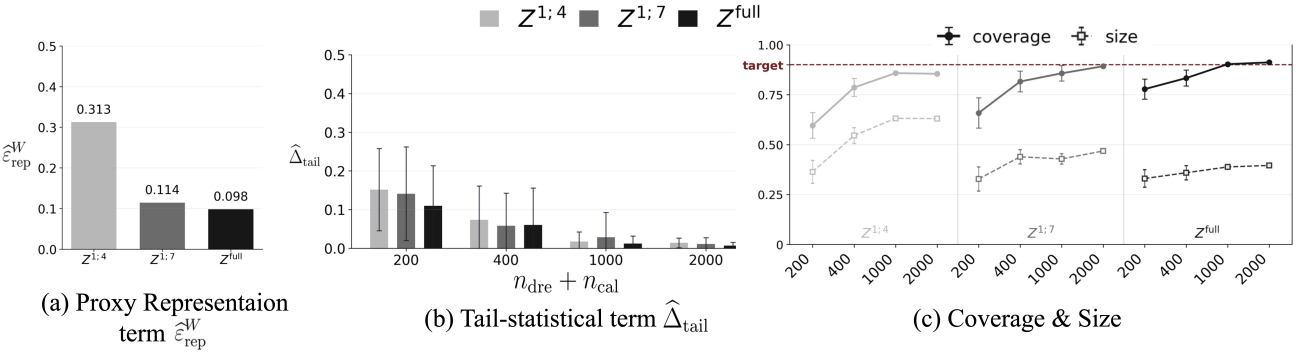


Figure 2. **Empirical conditional invariance of Y given extracted states.** Each panel corresponds to one of the top-K extracted-state cells C_k shared by HS1 and HS2. Within each cell, we plot the empirical helpfulness-score distribution of Y for HS1 and HS2. Conditioned on the same C_k, the two domains have similar overall shape, supporting $\mathbf{P}_{\mathrm{HS1}}(Y \mid \hat{Z}) \approx \mathbf{P}_{\mathrm{HS2}}(Y \mid \hat{Z})$.



(a) Proxy Representaion term $\widehat{\varepsilon}^{W}_{\mathrm{rep}}$ (b) Tail-statistical term $\widehat{\Delta}_{\mathrm{tail}}$ (c) Coverage & Size

Figure 3. **Post-hoc proxy diagnosis of representation and statistical terms.** (a) Extracted states that better separate broad correctness factors reduce the practical ordinal Y-invariance gap. (b) The statistical calibration diagnostic near the conformal cutoff becomes small once $n_{\mathrm{cal}} + n_{\mathrm{dre}} \geq 1000$. (c) These diagnostics align with the final coverage-size trade-off improvement: RoWCI stays near the nominal 90% coverage, and better Z-states produce smaller normalized intervals. Target labels are used only for post-hoc diagnostics and final evaluation.

dation terms. We define $\widehat{\varepsilon}^{W}_{\mathrm{rep}} = \sum_{\hat{z}} \widehat{\mathbf{P}}_1^{\hat{Z}}(\hat{z}) W_1(\widehat{\mathbf{P}}_0(Y \mid \hat{Z} = \hat{z}), \widehat{\mathbf{P}}_1(Y \mid \hat{Z} = \hat{z}))$, where W_1 is 1-Wasserstein distance, which measures ordinal Y-invariance within extracted states, and $\widehat{\Delta}_{\mathrm{stat}} = |\mathcal{U}|^{-1} \sum_{u \in \mathcal{U}} |\widehat{\mathbf{F}}_n(\widehat{q}_{\mathrm{full}}(u)) - u|$ with $\mathcal{U} = \{0.70, 0.75, \ldots, 0.95\}$, which measures cutoff-level calibration stability. Figure 3 first shows the representation side: the overly coarse state ($Z^{1;4}$) merges multiple internal situations and has a large Y-invariance gap, while the state (Z^{full}) that separates human-score-changing factors reduces this gap. This indicates that the resulting Z-cells better approximate a shared conditional human-scoring law. Second, the statistical diagnostic nearly vanishes once $n_{\mathrm{cal}} + n_{\mathrm{dre}} \geq 1000$, suggesting that finite-sample and DRE noise is not the main bottleneck. Finally, the outcome panel matches these diagnostics: once the human-scoring law is adequately separated by Z, RoWCI stays near nominal coverage, and the improved coverage-size trade-off appears. Comparing $Z^{1;7}$ with Z^{full} shows that more subquestions do not greatly change the invariance proxy, but they can improve the base estimator and thus the coverage–size trade-off. Since too many subquestions can make support sparse and unstable, we keep a moderate schema with at most ten subquestions.

6. Conclusion

When conformal inference is applied to LLM-as-a-judge, exchangeability is frequently violated and makes source calibration unreliable for the target domain. Although general shift-aware CI methods provide a theoretical route to correct such shifts, the text covariate introduces nuisance-driven weighting instability. RoWCI addresses this problem by mapping each example to a low-cardinality judgment-relevant representation Z and performing weighting in that space. The Z-space is constructed from the available evaluation specification, so that examples from different models, topics, or response policies can be calibrated together when they share the same internal judgment situation. This design reduces nuisance-driven weighting instability and enables more reliable reuse of existing source calibration data with human-scores, without target-domain human-scores. Across shifts in evaluated model, response policy, and topic, RoWCI moves coverage closer to the nominal level with lower variance and competitive interval size. RoWCI provides a practical route for more reliable LaaJ evaluation in realistic black-box settings where prompts, models, and judge policies keep changing.

Impact Statement

This paper aims to improve the reliability of automatic evaluation for open-ended generations. RoWCI can help make LLM-as-a-judge evaluation more transparent under distribution shift by reporting calibrated prediction intervals rather than relying only on a single judge score. RoWCI should not be interpreted as replacing human evaluation or expert review. The method inherits biases from human calibration labels, evaluation guidelines, and the LLM-based state extractor. Its guarantees also depend on the stability of the extracted state representation and representative unlabeled target covariates. In sensitive domains such as mental-health support evaluation, RoWCI should be used as a decision-support or auditing tool, with appropriate human oversight.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00457216). Also, This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) under the Leading Generative AI Human Resources Development(IITP-2026-RS-2026-25544647) grant funded by the Korea government(MSIT)

References

Anastasios N. Angelopoulos and Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification, 2022. *arXiv preprint arXiv:2107.07511*.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020. URL https://arxiv.org/abs/1907.02893.

Abeer Badawi, Elahe Rahimi, Md Tahmid Rahman Laskar, Sheri Grach, Lindsay Bertrand, Lames Danok, Prathiba Dhanesh, Jimmy Huang, Frank Rudzicz, and Elham Dolatabadi. When can we trust LLMs in mental health? large-scale benchmarks for reliable LLM evaluation. In Vera Demberg, Kentaro Inui, and Lluís Marquez, editors, *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3873–3896, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.180. URL https://aclanthology.org/2026.eacl-long.180/.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks, 2025. URL https://arxiv.org/abs/2406.18403.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or llms as the judge? a study on judgement biases, 2024. URL https://arxiv.org/abs/2402.10669.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794. ACM, August 2016. doi: 10.1145/2939672.2939785. URL http://dx.doi.org/10.1145/2939672.2939785.

G.B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2013. ISBN 9781118626399. URL https://books.google.co.kr/books?id=wI4fAwAAQBAJ.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 13806–13834. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.745. URL http://dx.doi.org/10.18653/v1/2024.acl-long.745.

Ying Jin and Emmanuel J. Candès. Model-free selective inference under covariate shift via weighted conformal p-values, 2023. URL https://arxiv.org/abs/2307.09291.

Sunay Joshi, Shayan Kiyani, George Pappas, Edgar Dobriban, and Hamed Hassani. Conformal inference under high-dimensional covariate shifts via likelihood-ratio regularization, 2025. URL https://arxiv.org/abs/2502.13030.

Kenneth Li, Tianle Liu, Naomi Bashkansky, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Measuring and controlling instruction (in)stability in language model dialogs, 2024. URL https://arxiv.org/abs/2402.10962.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment, 2023. URL https://arxiv.org/abs/2303.16634.

Felipe Maia Polo, Xinhe Wang, Mikhail Yurochkin, Gongjun Xu, Moulinath Banerjee, and Yuekai Sun. Bridging human and llm judgments: Understanding and narrowing the gap, 2025. URL https://arxiv.org/abs/2508.12792.

Leo Richter, Xuanli He, Pasquale Minervini, and Matt J. Kusner. An auditing test to detect behavioral shift in language models, 2025. URL https://arxiv.org/abs/2410.19406.

Huanxin Sheng, Xinyi Liu, Hangfeng He, Jieyu Zhao, and Jian Kang. Analyzing uncertainty of LLM-as-a-judge: Interval evaluations with conformal prediction. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11286–11328, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.569. URL https://aclanthology.org/2025.emnlp-main.569/.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in llm-as-a-judge, 2025. URL https://arxiv.org/abs/2406.07791.

Annalisa Szymanski, Noah Ziems, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks, 2024. URL https://arxiv.org/abs/2410.20266.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. Judgebench: A benchmark for evaluating llm-based judges, 2025. URL https://arxiv.org/abs/2410.12784.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges, 2025. URL https://arxiv.org/abs/2406.12624.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candes, and Aaditya Ramdas. Conformal prediction under covariate shift, 2020. URL https://arxiv.org/abs/1904.06019.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Vladimir Vovk, Ivan Petej, Paolo Toccaceli, and Alex Gammerman. Conformal calibrators, 2019. URL https://arxiv.org/abs/1902.06579.

Yidong Wang, Yunze Song, Tingyuan Zhu, Xuanwang Zhang, Zhuohao Yu, Hao Chen, Chiyu Song, Qiufeng Wang, Zhen Wu, Xinyu Dai, Yue Zhang, Cunxiang Wang, Wei Ye, and Shikun Zhang. Trustjudge: Inconsistencies of LLM-as-a-judge and how to alleviate them. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4uPyOCeN6U.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer2: Open-source dataset for training top-performing reward models, 2024a. URL https://arxiv.org/abs/2406.08673.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev. HelpSteer: Multi-attribute helpfulness dataset for SteerLM. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3371–3384, Mexico City, Mexico, June 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.185. URL https://aclanthology.org/2024.naacl-long.185/.

Hui Wei, Shenghua He, Tian Xia, Fei Liu, Andy Wong, Jingyang Lin, and Mei Han. Systematic evaluation of llm-as-a-judge in llm alignment tasks: Explainable metrics and diverse prompt templates, 2025. URL https://arxiv.org/abs/2408.13006.

Xuansheng Wu, Wenlin Yao, Jianshu Chen, Xiaoman Pan, Xiaoyang Wang, Ninghao Liu, and Dong Yu. From language modeling to instruction following: Understanding the behavior shift in llms after instruction tuning, 2024. URL https://arxiv.org/abs/2310.00492.

Yachong Yang, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. Doubly robust calibration of prediction sets under covariate shift, 2022. URL https://arxiv.org/abs/2203.01761.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models, 2025. URL https://arxiv.org/abs/2506.05176.

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, 2023. URL https://arxiv.org/abs/2306.05685.

Algorithm 1 Practical RoWCI

Input: source data with human-scores $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$, target covariates without human-scores $\mathcal{D}^c = \{X_j^c\}_{j=1}^m$, a new test input X_{n+1}, an extractor $\hat{z}$, a state scorer s_Z, and an error level $\alpha \in (0, 1)$.
Output: a prediction interval $\widehat{\mathcal{I}}_\alpha^{\hat{Z}}(X_{n+1})$.
— Z-extraction Step —
Split $\mathcal{D}$ into a calibration subset $\mathcal{D}_{\text{cal}}$ and a source subset $\mathcal{D}_{\text{src}}$ for density-ratio estimation.
Form $\mathcal{D}_{\text{dre}} = \{(X, A = 0) : X \in \mathcal{D}_{\text{src}}\} \cup \{(X, A = 1) : X \in \mathcal{D}^c\}$.
for each $(X_i, Y_i) \in \mathcal{D}_{\text{cal}}$ **do**
  Compute $\hat{Z}_i = \hat{z}(X_i)$, $S_i^{\hat{Z}} = s_Z(\hat{Z}_i)$, and $E_i^{\hat{Z}} = |S_i^{\hat{Z}} - Y_i|$.
end for
for each sample $X \in \mathcal{D}_{\text{dre}}$ **do**
  Compute $\hat{Z} = \hat{z}(X)$.
end for
Compute $\hat{Z}_{n+1} = \hat{z}(X_{n+1})$ and $S_{n+1}^{\hat{Z}} = s_Z(\hat{Z}_{n+1})$.
— WCI Step —
Fit a binary domain classifier to estimate the target-domain membership probability $\hat{\eta}(\hat{Z}) = \mathbb{P}(\hat{Z} \sim \mathbf{P}_1 \mid \hat{Z})$.
Define the estimated importance weight $\hat{w}(\hat{Z}) = \frac{m_0}{m_1} \frac{\hat{\eta}(\hat{Z})}{1-\hat{\eta}(\hat{Z})}$, where m_0, m_1 are the source and target sample counts in $\mathcal{D}_{\text{dre}}$.
Construct $\hat{t}_\alpha(\hat{Z}_{n+1})$ from $\{E_i^{\hat{Z}}\}_{(X_i,Y_i)\in\mathcal{D}_{\text{cal}}}$, $\{\hat{w}(\hat{Z}_i)\}_{(X_i,Y_i)\in\mathcal{D}_{\text{cal}}}$, and the test weight $\hat{w}(\hat{Z}_{n+1})$.
— Inference Step —
Construct the prediction interval

$$\widehat{\mathcal{I}}_\alpha^{\hat{Z}}(X_{n+1}) = \left[S_{n+1}^{\hat{Z}} - \hat{t}_\alpha(\hat{Z}_{n+1}), S_{n+1}^{\hat{Z}} + \hat{t}_\alpha(\hat{Z}_{n+1})\right] \cap [0, 1].$$

Return $\widehat{\mathcal{I}}_\alpha^{\hat{Z}}(X_{n+1})$.

A. Implementation Details

A.1. Constructing the subquestions for Z-space

Available evaluation specification. We construct the Z-space from the available evaluation specification for each judgment axis. When the benchmark provides a detailed written rubric, we use that rubric directly. When the benchmark provides only a short axis description, we use the axis name and benchmark description as a sparse specification and apply the same conservative rule below. In all cases, the schema is fixed without using target-domain human-scores, target-domain coverage, or target residuals.

We illustrate the construction using the helpfulness axis of HelpSteer-style annotations (Wang et al., 2024a;b). The human guideline for helpfulness is a five-level rubric:

0. Completely Unhelpful The response is not useful or helpful at all. The response completely misses the essence of what the user wanted.
1. Mostly Unhelpful The response is borderline unhelpful and mostly does not capture what the user was looking for, but is still usable and helpful in a small way.
2. Partially Helpful The response is partially helpful but misses the overall goal of the user's query/input in some way. The response did not fully satisfy what the user was looking for.
3. Mostly Helpful The response is mostly helpful and mainly aligned with what the user was looking for, but there is still some room for improvement.
4. Extremely Helpful The response is extremely helpful and completely aligned with the spirit of what the prompt was asking for.

From guideline to internal-state subquestions. The schema is induced by a single target-outcome-blind rule: keep the coarsest judgment distinctions that can move examples between broad score regions under the evaluation specification, and encode each distinction as a low-cardinality categorical subquestion. The goal is not to recover all information in the raw response, nor to maximize source-domain score prediction. The goal is to define cells in which the remaining human-score distribution is expected to be comparable across source and target.

For helpfulness, the rubric above repeatedly contrasts whether the response misses the user's request, partially satisfies it, mostly aligns with it, or completely aligns with its spirit. We turn these score contrasts into observable state questions. Missing the essence of the request motivates questions about whether the response is on the requested task and whether the requested output is present. Partial satisfaction motivates questions about how many main parts are addressed and whether explicit requirements are followed. The distinction between merely usable and strongly helpful motivates questions about concrete details, usable advice or directly usable output, and the amount of useful information. The phrase "aligned with the spirit of what the prompt was asking for" motivates questions about whether the content fits the user's goal and how the turn is handled when the request is limited, protocol-like, or under-specified. The task-shape question is included only as a coarse routing state: it records whether the same helpfulness checks should be interpreted for an ordinary interaction, a structured/artifact request, or a formal-reference/external-information request.

The same rule is applied to other axes. For correctness, the relevant score-changing distinctions concern whether the response attempts the task, whether the requested output is present, whether the core request is covered, whether a visible factual or reasoning problem exists, and whether special verification or exact-output constraints are needed. For guidance, the relevant distinctions concern how much actionable guidance is given, whether concrete next steps are present, whether a care or resource path is provided, whether the advice is structured or tailored, and whether a follow-up path is offered. Table 4 shows the resulting low-cardinality schemas for representative axes.

Design principle and auditability. A useful Z-schema should be neither maximally informative nor arbitrarily coarse. If the schema is too rich, density-ratio estimation in Z-space approaches raw-text or embedding-space matching, and source-target differences unrelated to calibration can dominate the weights. If the schema is too coarse, many different scoring situations are merged into the same cell, leaving more residual uncertainty for conformal calibration. We therefore use low-cardinality subquestions: each answer records a broad judgment distinction, not a detailed description of the example.

The construction is auditable in four ways. First, each subquestion must be traceable to the available evaluation specification, either directly from a rubric phrase or conservatively from the axis meaning in sparse specifications. Second, each answer set must be low-cardinality so that the induced Z-space remains a transport space rather than a raw-text proxy. Third, the answer should have the same interpretation in source and target domains. Fourth, the schema is frozen before calibration and target-domain evaluation. Thus, the table should be read as an explicit invariance hypothesis: after fixing the major score-changing distinctions for the axis, examples in the same Z-cell should have comparable remaining Y-distributions across domains. When labels are available only for analysis, we audit this hypothesis post hoc by comparing cell-wise empirical Y-distributions and by checking support overlap and weight concentration in Z-space; these diagnostics do not choose or tune the schema.

Why these questions need not be sufficient for Y in practical RoWCI. The extracted questions are not required to be sufficient for the human score Y in the sense of determining a unique rating. Even after conditioning on $\hat{Z}$, different human annotators may reasonably give different scores, and examples in the same $\hat{Z}$-cell may still have nonzero variation in Y. This is not a failure of the construction. The remaining within-cell uncertainty is precisely what the conformal calibration step is designed to absorb. The condition targeted by RoWCI is weaker and more appropriate for transfer. For the same extracted state, the conditional law of the human score should be approximately stable across source and target domains, i.e.,

$$\mathbf{P}_0(Y \mid \hat{Z} = \hat{z}) \approx \mathbf{P}_1(Y \mid \hat{Z} = \hat{z}).$$

Thus, the goal is not to make $\hat{Z}$ maximally predictive of Y, but to make the residual uncertainty comparable across domains. A stable extraction bias from the LLM is also tolerable. For example, if the extractor consistently labels borderline specificity as mixed rather than specific in both source and target domains, this simply redefines the cells of $\hat{Z}$. It affects efficiency only through coarsening. What harms coverage is a domain-dependent extraction bias that changes the conditional law of Y given the extracted state. This is exactly the representation gap captured by ε_{rep} in Theorem 3.

A.2. Computing the Score $s_Z(\hat{Z})$

After extracting the guideline state $\hat{Z} = \hat{z}(X)$, RoWCI requires a base score $S^{\hat{Z}} = s_Z(\hat{Z})$. The important structural requirement is that this score is a function of the extracted state rather than a direct function of the raw text. We consider two practical ways to construct this state-only score.

LLM scorer. The first option is to ask an LLM-as-a-judge to map the extracted state $\hat{Z}$ back to a final score. The judge is given the original guideline and the extracted state values, but not the raw response text. This enforces the structural form

$$S^{\hat{Z}} = s_{Z,\mathrm{LLM}}(\hat{Z}).$$

We consider zero-shot, five-shot, and ten-shot versions. In the zero-shot version, the prompt contains only the guideline and the extracted state. In the few-shot versions, the prompt additionally contains five or ten state-score demonstrations. When demonstrations are used, they are drawn from a non-calibration source split and fixed before computing conformal residuals. We do not use target-domain human-scores for these demonstrations.

Traditional machine-learning-based scorer. The second option is to learn a lightweight state-to-score model on the source subset with human-scores that is not used for conformal calibration. This is the implementation used in our experiments. Recall that the source dataset $\mathcal{D}$ is split into a held-out calibration set $\mathcal{D}_{\mathrm{cal}}$ and a remaining source subset used to fit the density-ratio estimator. We use this remaining source subset with human-scores to train the state scorer. Thus, no additional human-scores are required beyond the source calibration human-scores already available in the source calibration data.

Let

$$\mathcal{D}_{\mathrm{score}} = \{(X_j^s, Y_j^s)\}_{j=1}^{n_{\mathrm{score}}}$$

denote the source subset with human-scores used for fitting the state scorer, disjoint from $\mathcal{D}_{\mathrm{cal}}$. For each $X_j^s \in \mathcal{D}_{\mathrm{score}}$, we extract

$$\hat{Z}_j^s = \hat{z}(X_j^s).$$

We then one-hot encode the categorical state $\hat{Z}_j^s$ into $\phi(\hat{Z}_j^s)$, and train an XGBoost regressor or a multiclass classifier (Chen and Guestrin, 2016) to predict the human-score from the extracted state:

$$\left\{(\phi(\hat{Z}_j^s), Y_j^s)\right\}_{j=1}^{n_{\mathrm{score}}}.$$

In our implementation, we use a multiclass XGBoost classifier over the five Likert labels. Let

$$\hat{p}_\theta(k \mid \hat{Z}) = \mathbb{P}_\theta(Y = k \mid \hat{Z}), \qquad k \in \{0, 1, \cdots, K\},$$

be the predicted class probabilities. We define the normalized state-only score by the probability-weighted average

$$s_{Z,\mathrm{XGB}}(\hat{Z}) = \sum_{i=0}^{R} \frac{i}{R}\, \hat{p}_\theta(i \mid \hat{Z}) \in [0, 1].$$

This design has several practical advantages. First, compared with asking an LLM-as-a-judge to score every extracted state, the additional resource cost is almost negligible. Once the states are extracted, fitting and evaluating XGBoost on a low-dimensional categorical table is inexpensive, deterministic, and easy to reproduce. Second, the learned mapping is restricted to the guideline state $\hat{Z}$, rather than the raw text X. Hence the scorer cannot exploit domain-specific nuisance variables such as response length, formatting, model-specific phrasing, or topic markers. Third, the learning problem is deliberately simple. It is a low-dimensional categorical prediction problem from $\hat{Z}$ to Y. This makes the score function less sensitive to raw-text distribution shift than a scorer trained directly on text embeddings. Importantly, this scorer is not intended to solve the full human-evaluation problem by supervised learning. It only provides a base prediction around which conformal calibration builds an interval. The role of XGBoost is to learn a stable state-to-score map on the source domain, while the subsequent weighted conformal step corrects the residual error distribution under source-target shift in the same $\hat{Z}$-space.

B. Details of Experiment Settings

We report the following metrics:

$$\mathrm{Cov.} := \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}\left\{Y_j \in \widehat{\mathcal{I}}_\alpha(X_j)\right\},$$

$$\text{Size} := \frac{1}{Rm}\sum_{j=1}^{m}\left|\widehat{\mathcal{I}}_\alpha(X_j)\right|.$$

Here, m is the number of target-domain test samples. The factor R rescales normalized interval width or set size to $[0, 1]$.

B.1. Randomness and error bars.

Each reported value is computed over 20 runs with different global random seeds. The seed controls the source/calibration split, the target evaluation subset when subsampling is used, and the randomized components of the state scorer and density-ratio estimator. LLM-based state extraction is run with fixed decoding settings and cached before conformal calibration. We report sample standard deviation, not standard error, as the error bar.

B.2. Detailed Implementations of Baselines

B.2.1. Post-hoc LaaJ score calibration baselines

We provide implementation details for the non-conformal post-hoc calibration baselines used in Table 1. Since these methods do not directly output conformal intervals, we convert their predicted score distributions into prediction sets in a common way. For each example, a method produces an empirical probability mass function $\hat{p}(y \mid X)$ over the discrete score classes $\mathcal{Y}$. We sort the classes by decreasing probability and add them one by one until the cumulative probability mass reaches at least $1-\alpha$. This gives a prediction set

$$\widehat{\mathcal{C}}_\alpha(X) = \min\left\{C \subseteq \mathcal{Y} : \sum_{y\in C}\hat{p}(y \mid X) \geq 1-\alpha\right\},$$

where the minimum is implemented by the greedy probability ranking. We report coverage as

$$\frac{1}{m}\sum_{j=1}^{m}\mathbf{1}\{Y_j \in \widehat{\mathcal{C}}_\alpha(X_j)\},$$

and normalized size as

$$\frac{1}{m}\sum_{j=1}^{m}\frac{|\widehat{\mathcal{C}}_\alpha(X_j)|}{|\mathcal{Y}|}.$$

This conversion is used only to make the non-conformal baselines comparable under the same coverage-size evaluation protocol. It does not give these baselines a conformal coverage guarantee.

LLM-RUBRIC (Hashemi et al., 2024). We implement the full LLM-RUBRIC pipeline. LLM-RUBRIC is a training-heavy baseline because, for each new evaluation rubric, it requires supervision for multiple rubric-level auxiliary questions in addition to the final human-score. Our datasets provide final human-scores but do not provide ground-truth labels for the auxiliary LLM-RUBRIC questions. Therefore, we keep the original LLM-RUBRIC auxiliary-question structure and obtain hard labels for the auxiliary questions using GPT-5.5 reasoning mode. These auxiliary labels are used as rubric-level supervision, while final human-scores from the source split are used to train the final score predictor. We do not use target-domain human-scores for training. After training, the model outputs a probability distribution over the classes of final scores, which is converted into a prediction set by the cumulative-mass rule above.

TrustJudge (Wang et al., 2026). We implement a TrustJudge-style log-probability baseline using the score distribution produced by the judge. Specifically, for each example, we obtain token log-probabilities over the candidate score labels, normalize them into a probability mass function and apply the TrustJudge softmax transformation to produce a sharpened score distribution. This baseline does not use target-domain human-scores and does not perform conformal calibration or source-target reweighting. As with Bridge, TrustJudge requires access to score-label log-probabilities, so we evaluate it only for Qwen. The resulting score distribution is converted into a prediction set by accumulating classes in decreasing probability order until the mass reaches $1-\alpha$.

Bridge (Polo et al., 2025). We implement Bridge as a supervised post-hoc calibration model that maps LLM score distributions and lightweight response covariates to human-scores. Following the Bridge setup, we first obtain token log-probabilities (Qwen-3-30B, GPT-4o-mini) or multi-sample approximations of probabilities (GPT-5.5) for the candidate score labels and normalize them into a score distribution over $\mathcal{Y}$. We then use this distribution as the LLM-side judgment signal and combine it with lightweight textual covariates such as length, readability, formatting, sentiment, and discourse-marker features. The Bridge model is trained only on the source split and evaluated on the shifted target split. The fitted Bridge model outputs a calibrated score distribution, which we convert into a prediction set using the same cumulative-mass rule.

B.2.2. Shift-aware Conformal Inference Baselines

For all covariate-shift baselines, we use the same fixed embedding representation as the algorithmic input. Each evaluation instance consists of a user instruction u and a model response v. We first serialize the pair into a single text string using a fixed template that explicitly marks the two fields:

$$\text{text}(u, v) = \texttt{User instruction is: \{u\}. Model response is: \{v\}.}$$

We then encode this serialized text with Qwen3-Embedding-8B (Zhang et al., 2025) and denote the resulting embedding by

$$X = \phi(\text{text}(u, v)) \in \mathbb{R}^d.$$

In the following descriptions, X always denotes this embedding vector, not the raw text. All domain classifiers, nuisance estimators, and threshold functions take X as their input. Let $g(X) \in [0, 1]$ be the uncalibrated LLM-as-a-judge score and let $Y \in [0, 1]$ be the human-score. For labeled source examples, we use the absolute judgment error

$$R(X, Y) = |g(X) - Y|$$

as the nonconformity score.

WCI (Tibshirani et al., 2020). We implement weighted conformal inference using the weighted conformal prediction construction for covariate shift in Section 2.1 and Corollary 1, Eqs. (6)–(7), of Tibshirani et al. (2020); this construction is the covariate-shift specialization of the weighted-exchangeability result in Theorem 2. The original paper presents this method through the weighted conformal band rather than through a numbered algorithm. Let P_0^X and P_1^X denote the source and target distributions over the embedding X, respectively. Under covariate shift, WCI reweights source calibration scores by the source-to-target covariate likelihood ratio

$$w(x) = \frac{dP_1^X}{dP_0^X}(x).$$

Since this ratio is unknown in our setting, we estimate it from embeddings using a binary domain classifier trained to distinguish source examples from target examples. Specifically, with domain label $D = 0$ for source and $D = 1$ for target, we train a Random Forest classifier and write

$$\hat{a}(x) = \widehat{\mathbb{P}}(D = 1 \mid X = x).$$

We use $\hat{w}(x) \propto \frac{\hat{a}(x)}{1-\hat{a}(x)}$, where the proportionality constant is immaterial for the normalized weighted quantile. For a target embedding x, WCI forms the weighted empirical distribution of source calibration scores $\{R_i\}$ with weights proportional to $\hat{w}(X_i)$, together with the test-point mass used in the weighted conformal construction. The conformal cutoff is the corresponding $(1 - \alpha)$-quantile, denoted $\hat{q}_\alpha(x)$. The final prediction interval is

$$\widehat{\mathcal{I}}_\alpha^{\text{WCI}}(x) = \big[g(x) - \hat{q}_\alpha(x),\, g(x) + \hat{q}_\alpha(x)\big] \cap [0, 1].$$

DR (Yang et al., 2022). We implement the doubly robust calibration baseline following Algorithm 1, "Split doubly robust prediction," of Yang et al. (2022). The algorithm is designed for covariate shift with labeled source data and unlabeled target covariates; in our implementation, the covariate is the embedding X. Following the paper's missing-data notation, we set $T = 0$ for labeled source examples and $T = 1$ for target examples. For source examples, the observed score is $R = |g(X) - Y|$. Algorithm 1 estimates two nuisance functions: the covariate-shift odds

$$\pi(x) = \frac{\mathbb{P}(T = 1 \mid X = x)}{\mathbb{P}(T = 0 \mid X = x)}$$

and the conditional score distribution

$$m(\theta, x) = \mathbb{E}[\mathbf{1}\{R \le \theta\} \mid X = x].$$

We estimate π from embeddings using the same Random Forest domain-classification approach as in WCI. Given $\hat{a}(x) = \widehat{\mathbb{P}}(T = 1 \mid X = x)$, we set

$$\hat{\pi}(x) = \frac{\hat{a}(x)}{1 - \hat{a}(x)}.$$

Given nuisance estimates $\hat{\pi}$ and $\hat{m}$, Algorithm 1 selects the smallest cutoff $\hat{\theta}$ satisfying the empirical doubly robust estimating equation based on the influence-function form

$$\widehat{\mathbb{P}}[\mathbf{1}\{T = 0\}\hat{\pi}(X)\{\mathbf{1}(R \le \theta) - \hat{m}(\theta, X)\} + \mathbf{1}\{T = 1\}\{\hat{m}(\theta, X) - (1 - \alpha)\}] \ge 0.$$

The returned prediction set is $\{y : R(x, y) \le \hat{\theta}\}$. With our absolute-error score, this becomes the interval

$$\widehat{\mathcal{I}}_\alpha^{\mathrm{DR}}(x) = \left[g(x) - \hat{\theta},\ g(x) + \hat{\theta}\right] \cap [0, 1].$$

LR-QR (Joshi et al., 2025). We implement likelihood-ratio regularized quantile regression following Algorithm 1, "Likelihood-ratio regularized quantile regression," of Joshi et al. (2025). The algorithm takes n_1 labeled source examples, n_2 unlabeled target examples, and n_3 unlabeled source examples. In our implementation, all three groups are represented by the embedding X. For each labeled source example, we compute the score

$$S_i = S(X_i, Y_i) = |g(X_i) - Y_i|.$$

LR-QR then learns an input-dependent threshold function $h \in \mathcal{H}$ without directly estimating the full likelihood-ratio function. Let ℓ_α denote the pinball loss,

$$\ell_\alpha(c, s) = (1 - \alpha)(s - c)\mathbf{1}\{s \ge c\} + \alpha(c - s)\mathbf{1}\{s < c\}.$$

Following Algorithm 1, we solve the empirical LR-QR objective

$$(\hat{h}, \hat{\beta}) \in \arg\min_{h \in \mathcal{H},\, \beta \in \mathbb{R}} \widehat{\mathbb{E}}_1\left[\ell_\alpha\big(h(X), S(X, Y)\big)\right] + \lambda \widehat{\mathbb{E}}_3\left[\beta^2 h(X)^2\right] - \lambda \widehat{\mathbb{E}}_2[2\beta h(X)].$$

Here, $\widehat{\mathbb{E}}_1$, $\widehat{\mathbb{E}}_2$, and $\widehat{\mathbb{E}}_3$ denote empirical expectations over labeled source examples, unlabeled target examples, and unlabeled source examples, respectively. The scalar β is an auxiliary normalization variable optimized jointly with h; the final prediction set returned by Algorithm 1 uses $\hat{h}$, not $\hat{\beta}$, as the threshold. Thus, for a target embedding x, LR-QR returns

$$\widehat{\mathcal{I}}_\alpha^{\text{LR-QR}}(x) = \{y : S(x, y) \le \hat{h}(x)\}.$$

With our absolute-error score, this is

$$\widehat{\mathcal{I}}_\alpha^{\text{LR-QR}}(x) = \left[g(x) - \hat{h}(x),\ g(x) + \hat{h}(x)\right] \cap [0, 1].$$

C. Additional Empirical Results

RoWCI under Domain Shift. Table 5 evaluates a transfer setting where calibration and target differ by topic domain. SCI-BA again undercovers across all three transfers and requires large prediction sets. RoWCI consistently moves coverage toward nominal coverage while reducing size. The improvement is especially clear on medical $\rightarrow$ math, where RoWCI reaches the nominal 90% level almost exactly while also producing the smallest size. This suggests that transport through Z remains useful because Z is designed to encode states that are directly relevant to the judgment axis, such as correctness, rather than domain-specific surface features.

Robustness of RoWCI to unnecessary shift correction. When both calibration and target come from HS2, where exchangeability holds, RoWCI remains close to the Target-cal benchmark across all axes. This indicates that the representation-and-reweighting pipeline does not noticeably distort the conformal cutoff when additional shift correction is unnecessary. Table 6 shows that this robustness is not limited to coverage alone. The RoWCI intervals remain nearly identical to the

Target-cal intervals, with size changing only from 0.39 to 0.40 on complexity, from 0.57 to 0.60 on correctness, and from 0.40 to 0.41 on helpfulness.

Real examples of $\hat{Z}$. Table 7 shows two examples from HS1 and HS2. The examples have different raw surface forms. The HS1 example is a paragraph-style cyber-safety answer, while the HS2 example is a numbered operating-system instruction. Nevertheless, after applying the helpfulness-state extractor, both examples map to the same high-helpfulness state.

HS1 example.

Instruction: What is a common scheme used by online scammers, and what can you do to avoid becoming a victim? Please provide your answer in paragraph form.

Assistant: Online scams come in many forms, but one common scheme is the advance fee fraud. This occurs when a scammer first establishes contact with their victim through email, text, social media, or other online platform. They may initially make contact by offering a prize or lottery win, or by claiming to be a long-lost relative or friend. Once the scammer has established a relationship with the victim, they will then offer a large sum of money or other valuable asset in exchange for a fee or other payment. The scammer will often ask for the fee to be paid through wire transfer or other non-reversible payment method, and will then disappear without providing the promised prize or asset.

To avoid becoming a victim of an advance fee fraud, it is important to be cautious of any unexpected or unsolicited offers of money or prizes. It is also important to never pay a fee or other payment in exchange for a prize or asset, especially if the payment is requested through non-reversible methods such as wire transfer. If you are unsure about the legitimacy of an offer or request, it is best to seek the advice of a trusted friend, family member, or financial advisor before making any payments or providing any personal information.

Human helpfulness score: $Y = 4$.

HS2 example.

Instruction: How do I change the short cut on mac osx to toggle the voiceover feature?

Assistant: To change the shortcut on Mac OS X to toggle the VoiceOver feature, follow these steps:

1. Open System Preferences by clicking on the Apple icon in the top-left corner of the screen and selecting "System Preferences" from the drop-down menu.

2. Click on "Accessibility" in the list of preferences.

3. Select "VoiceOver" from the left-hand menu.

4. Click on the "Keyboard Shortcuts" tab.

5. Select "VoiceOver" from the "Shortcut" drop-down menu.

6. Press the key combination that you want to use to toggle VoiceOver on and off.

7. Click on the "Add" button to save your new shortcut.

That's it! Now you can use the key combination you specified to toggle VoiceOver on and off.

Human helpfulness score: $Y = 4$.

These examples illustrate the desired invariance of the RoWCI state. In raw-text space, the two inputs are easy to separate: one contains cyber-safety vocabulary and paragraph-form advice, while the other contains Mac OS X interface terms and numbered procedural steps. These differences are nuisance coordinates U for helpfulness calibration. A raw-text or embedding-space density-ratio estimator can easily encode such domain and format differences, even though they are not the reason the human helpfulness score is high.

By contrast, the extracted state keeps the guideline-relevant facts that matter for helpfulness. In both examples, the response addresses the requested task, provides the requested output type, belongs to an ordinary answerable interaction, covers the main parts of the request, includes concrete task-relevant details, gives specific usable information, fits the user's goal, satisfies any explicit constraint or has none to violate, contains substantial useful information, and is handled as ordinary answerable work. Thus, conditioning on $\hat{Z}$ makes the two examples comparable for the helpfulness judgment, supporting the working approximation

$$\mathbf{P}_{\mathrm{HS1}}(Y \mid \hat{Z} = \hat{z}) \approx \mathbf{P}_{\mathrm{HS2}}(Y \mid \hat{Z} = \hat{z}),$$

while avoiding the nuisance-driven matching problem that arises from estimating weights directly on X.

D. Details of Theoretical Results

D.1. Conformal Inference for LLM-as-a-judge

Conformal Inference (Vovk et al., 2005; 2019) is a distribution-free procedure that calibrates a prediction interval using a held-out calibration set. In the LaaJ setting, (Sheng et al., 2025) apply conformal inference to construct an interval around the raw judge score $S = s_X(X)$. Without loss of generality, we set the score range to $[0, 1]$. Given a cutoff $\hat{t}_\alpha \in [0, 1]$, they define

$$\widehat{\mathcal{I}}_\alpha(X) \;=\; [\, s_X(X) - \hat{t}_\alpha,\; s_X(X) + \hat{t}_\alpha \,] \cap [0, 1]. \tag{3}$$

The interval covers the human-score Y if and only if the absolute error is below the cutoff: $Y \in \widehat{\mathcal{I}}_\alpha(X) \iff |s_X(X) - Y| \le \hat{t}_\alpha$. For each $i \in [n]$, define $S_i := s_X(X_i)$ and calibration nonconformity scores $E_i := |S_i - Y_i|$ for $(X_i, Y_i) \in \mathcal{D}$. To choose $\hat{t}_\alpha$, they take an empirical $(1-\alpha)$ quantile of $\{E_i\}_{i=1}^n$ with the standard finite-sample correction. Concretely, letting $E_{(1)} \le \cdots \le E_{(n)}$ be the sorted scores and setting $E_{(n+1)} := 1$, they set $\hat{t}_\alpha = E_{(k)}$ where $k = \lceil (1-\alpha)(n+1) \rceil$.

Proposition 2 (Conformal validity under i.i.d., following (Sheng et al., 2025))**.** *Let $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbf{P}_0$ and $(X_{n+1}, Y_{n+1}) \sim \mathbf{P}_0$ be an independent test point. Define $\widehat{\mathcal{I}}_\alpha$ by equation (3) with $\hat{t}_\alpha$ chosen as above. Then*

$$\mathbb{P}_{\mathbf{P}_0}\Big(Y \in \widehat{\mathcal{I}}_\alpha(X_{n+1})\Big) \;\ge\; 1 - \alpha.$$

Proof of Proposition 2. Let $(X_{n+1}, Y_{n+1}) \sim \mathbf{P}_0$ be an independent test point, and let $S_i = s_X(X_i)$ and $E_i = |S_i - Y_i|$ for $i \in [n+1]$. By equation (3), the coverage event is

$$Y_{n+1} \in \widehat{\mathcal{I}}_\alpha(X_{n+1}) \iff E_{n+1} \le \hat{t}_\alpha.$$

Since $\{(X_i, Y_i)\}_{i=1}^{n+1} \overset{\text{i.i.d.}}{\sim} \mathbf{P}_0$, the scores $E_1, \ldots, E_n, E_{n+1}$ are exchangeable. Let $E_{(1)} \le \cdots \le E_{(n)}$ be the order statistics of the calibration scores, set $E_{(n+1)} := 1$, and let $k = \lceil (1-\alpha)(n+1) \rceil$. The cutoff is chosen as the k-th calibration order statistic, so $\hat{t}_\alpha = E_{(k)}$, and therefore

$$\{E_{n+1} \le \hat{t}_\alpha\} = \{E_{n+1} \le E_{(k)}\}.$$

By exchangeability, the rank of E_{n+1} among $E_1, \ldots, E_n, E_{n+1}$ is uniform over $\{1, \ldots, n+1\}$. If ties occur at the cutoff, the event $\{E_{n+1} \le E_{(k)}\}$ becomes only more likely. Hence

$$\mathbb{P}_{\mathbf{P}_0}(E_{n+1} \le E_{(k)}) \ge k/(n+1) \ge 1 - \alpha.$$

This gives

$$\mathbb{P}_{\mathbf{P}_0}(Y_{n+1} \in \widehat{\mathcal{I}}_\alpha(X_{n+1})) \ge 1 - \alpha.$$

□

Proof of Proposition 1. Let $(X_{n+1}, Y_{n+1}) \sim \mathbf{P}_1$ be an independent test point, and let $S_i = s_X(X_i)$ and $E_i = |S_i - Y_i|$ for $i \in [n+1]$. By construction of the interval, the coverage event is equivalent to

$$E_{n+1} \le \hat{t}_\alpha(X_{n+1}).$$

Under equation (1) and $\mathbf{P}_1^X \ll \mathbf{P}_0^X$, $w(x) = d\mathbf{P}_1^X / d\mathbf{P}_0^X(x)$ is well-defined. Since $\hat{t}_\alpha(x)$ is defined above by the standard weighted conformal rule using calibration weights together with the test weight $w(x)$, the resulting procedure is exactly the weighted conformal procedure under covariate shift (Tibshirani et al., 2020). Therefore, by the standard validity result for weighted conformal prediction,

$$\mathbb{P}_{\mathbf{P}_1}(E_{n+1} \le \hat{t}_\alpha(X_{n+1})) \ge 1 - \alpha.$$

Equivalently, $\mathbb{P}_{\mathbf{P}_1}(Y_{n+1} \in \widehat{\mathcal{I}}_\alpha(X_{n+1})) \ge 1 - \alpha$. □

Lemma 1 (Sufficiency of Representation Weights)**.** *Assume equation (2) holds, i.e., $\mathbf{P}_0(E \mid Z) = \mathbf{P}_1(E \mid Z)$ and assume $\mathbf{P}_1^Z \ll \mathbf{P}_0^Z$. Let $w(Z) := \frac{d\mathbf{P}_1^Z}{d\mathbf{P}_0^Z}(Z)$ be the density-ratio in the Z-space. For any bounded measurable function f, the expectation of $f(E)$ under the target distribution $\mathbf{P}_1$ can be exactly recovered by reweighting the source distribution $\mathbf{P}_0$ with $w(Z)$:*

$$\mathbb{E}_{\mathbf{P}_1}[f(E)] \;=\; \mathbb{E}_{\mathbf{P}_0}\big[w(Z) f(E)\big].$$

Proof of Lemma 1. By the law of total expectation under the target distribution $\mathbf{P}_1$, we can condition on the representation Z:

$$\mathbb{E}_{\mathbf{P}_1}[f(E)] \;=\; \mathbb{E}_{\mathbf{P}_1^Z}\Big[\mathbb{E}_{\mathbf{P}_1}\big[f(E) \mid Z\big]\Big].$$

Under the assumption $\mathbf{P}_0(E \mid Z) = \mathbf{P}_1(E \mid Z)$, the conditional expectation of $f(E)$ given Z is invariant across the two policies. Therefore, we can substitute the inner expectation:

$$\mathbb{E}_{\mathbf{P}_1}[f(E)] \;=\; \mathbb{E}_{\mathbf{P}_1^Z}\Big[\mathbb{E}_{\mathbf{P}_0}\big[f(E) \mid Z\big]\Big].$$

Next, we perform a change of measure from the target marginal $\mathbf{P}_1^Z$ to the source marginal $\mathbf{P}_0^Z$ using the representation density-ratio $w(Z) = \frac{d\mathbf{P}_1^Z}{d\mathbf{P}_0^Z}(Z)$. This yields:

$$\mathbb{E}_{\mathbf{P}_1}[f(E)] \;=\; \mathbb{E}_{\mathbf{P}_0^Z}\Big[w(Z)\,\mathbb{E}_{\mathbf{P}_0}\big[f(E) \mid Z\big]\Big].$$

Since the weight $w(Z)$ is purely a function of the conditioning variable Z, it can be absorbed into the inner conditional expectation:

$$\mathbb{E}_{\mathbf{P}_1}[f(E)] \;=\; \mathbb{E}_{\mathbf{P}_0^Z}\Big[\mathbb{E}_{\mathbf{P}_0}\big[w(Z)f(E) \mid Z\big]\Big].$$

Finally, applying the law of total expectation under $\mathbf{P}_0$ outward, we obtain

$$\mathbb{E}_{\mathbf{P}_1}[f(E)] \;=\; \mathbb{E}_{\mathbf{P}_0}\big[w(Z)f(E)\big].$$

□

Proof of Theorem 1. Let $(X_{n+1}, Y_{n+1}) \sim \mathbf{P}_1$ be an independent test point, and let $Z_{n+1} = z(X_{n+1})$ and $E_{n+1} = |s_X(X_{n+1}) - Y_{n+1}|$. Define the raw-score error by $E := |s_X(X) - Y|$. For any bounded measurable function f, the law of total expectation under the target distribution gives

$$\mathbb{E}_{\mathbf{P}_1}[f(E)] = \mathbb{E}_{\mathbf{P}_1^Z}\Big[\mathbb{E}_{\mathbf{P}_1}[f(E) \mid Z]\Big].$$

By equation (2), the conditional law of E given Z is the same under $\mathbf{P}_0$ and $\mathbf{P}_1$, so

$$\mathbb{E}_{\mathbf{P}_1}[f(E)] = \mathbb{E}_{\mathbf{P}_1^Z}\Big[\mathbb{E}_{\mathbf{P}_0}[f(E) \mid Z]\Big].$$

Changing measure from $\mathbf{P}_1^Z$ to $\mathbf{P}_0^Z$ with the density-ratio $w(Z) = d\mathbf{P}_1^Z/d\mathbf{P}_0^Z(Z)$, we obtain

$$\mathbb{E}_{\mathbf{P}_1}[f(E)] = \mathbb{E}_{\mathbf{P}_0^Z}\Big[w(Z)\,\mathbb{E}_{\mathbf{P}_0}[f(E) \mid Z]\Big].$$

Since $w(Z)$ is measurable with respect to Z, it can be absorbed into the inner conditional expectation, giving

$$\mathbb{E}_{\mathbf{P}_1}[f(E)] = \mathbb{E}_{\mathbf{P}_0^Z}\Big[\mathbb{E}_{\mathbf{P}_0}[w(Z)f(E) \mid Z]\Big] = \mathbb{E}_{\mathbf{P}_0}[w(Z)f(E)].$$

Thus $w(Z)$ is the exact importance weight for the target law of E. In particular, for every threshold t, taking $f(e) = \mathbf{1}\{e \le t\}$ yields

$$\mathbb{P}_{\mathbf{P}_1}(E \le t) = \mathbb{E}_{\mathbf{P}_0}\big[w(Z)\mathbf{1}\{E \le t\}\big].$$

Therefore, applying the weighted conformal validity argument of Proposition 1 to calibration judgment errors $E_i = |s_X(X_i) - Y_i|$, calibration weights $w(Z_i)$, and test weight $w(Z_{n+1})$ yields

$$\mathbb{P}_{\mathbf{P}_1}\big(E_{n+1} \le \hat{t}_\alpha^Z(X_{n+1})\big) \ge 1 - \alpha.$$

Since

$$\{Y_{n+1} \in \widehat{\mathcal{I}}_\alpha^Z(X_{n+1})\} \iff \{E_{n+1} \le \hat{t}_\alpha^Z(X_{n+1})\},$$

we conclude that

$$\mathbb{P}_{\mathbf{P}_1}\Big(Y_{n+1} \in \widehat{\mathcal{I}}_\alpha^Z(X_{n+1})\Big) \ge 1 - \alpha.$$

□

Proof of Theorem 2. Under the stated absolute continuity conditions, $w(z) = d\mathbf{P}_1^Z/d\mathbf{P}_0^Z(z)$ and $w(u \mid Z) = d\mathbf{P}_1^{U|Z}/d\mathbf{P}_0^{U|Z}(u \mid Z)$ are well-defined. Since $X \equiv (Z, U)$ almost surely, the chain rule for Radon–Nikodym (Folland, 2013) derivatives gives

$$\frac{d\mathbf{P}_1^X}{d\mathbf{P}_0^X}(X) = \frac{d\mathbf{P}_1^Z}{d\mathbf{P}_0^Z}(Z)\,\frac{d\mathbf{P}_1^{U|Z}}{d\mathbf{P}_0^{U|Z}}(U \mid Z),$$

which is exactly

$$w(X) = w(Z)\,w(U \mid Z).$$

Conditioning on Z under $\mathbf{P}_0$, we then have

$$\mathbb{E}_{\mathbf{P}_0}[w(X) \mid Z] = w(Z)\,\mathbb{E}_{\mathbf{P}_0}[w(U \mid Z) \mid Z].$$

By the defining property of the conditional density-ratio,

$$\mathbb{E}_{\mathbf{P}_0}[w(U \mid Z) \mid Z] = \int \frac{d\mathbf{P}_1^{U|Z}}{d\mathbf{P}_0^{U|Z}}(u \mid Z)\,d\mathbf{P}_0^{U|Z}(u \mid Z) = 1,$$

so

$$\mathbb{E}_{\mathbf{P}_0}[w(X) \mid Z] = w(Z).$$

Thus $w(Z)$ is the conditional expectation of the raw weight given Z. Applying the law of total variance,

$$\mathrm{Var}_{\mathbf{P}_0}(w(X)) = \mathrm{Var}_{\mathbf{P}_0}\big(\mathbb{E}_{\mathbf{P}_0}[w(X) \mid Z]\big) + \mathbb{E}_{\mathbf{P}_0}\big[\,\mathrm{Var}_{\mathbf{P}_0}(w(X) \mid Z)\big].$$

Substituting $\mathbb{E}_{\mathbf{P}_0}[w(X) \mid Z] = w(Z)$ gives

$$\mathrm{Var}_{\mathbf{P}_0}(w(X)) - \mathrm{Var}_{\mathbf{P}_0}(w(Z)) = \mathbb{E}_{\mathbf{P}_0}\big[\,\mathrm{Var}_{\mathbf{P}_0}(w(X) \mid Z)\big].$$

Finally, since $w(X) = w(Z)\,w(U \mid Z)$ and $w(Z)$ is fixed given Z,

$$\mathrm{Var}_{\mathbf{P}_0}(w(X) \mid Z) = \mathrm{Var}_{\mathbf{P}_0}\big(w(Z)w(U \mid Z) \mid Z\big) = w(Z)^2\,\mathrm{Var}_{\mathbf{P}_0}(w(U \mid Z) \mid Z).$$

Therefore,

$$\mathrm{Var}_{\mathbf{P}_0}(w(X)) - \mathrm{Var}_{\mathbf{P}_0}(w(Z)) = \mathbb{E}_{\mathbf{P}_0}\Big[w(Z)^2\,\mathrm{Var}_{\mathbf{P}_0}(w(U \mid Z) \mid Z)\Big].$$

□

Lemma 2 (Statistical term bound)**.** *Assume $\{\eta(\hat{Z}), \hat{\eta}(\hat{Z})\} \in [\tau, 1-\tau]$ almost surely for some $\tau \in (0, \frac{1}{2})$, and suppose*

$$\|\hat{\eta} - \eta\|_{L_1(\mathbf{P}_0^{\hat{Z}})} \le \Gamma_{\mathrm{dre}}(n_{\mathrm{dre}}, \delta)$$

with probability at least $1-\delta$. Then there exists a constant $C > 0$, depending only on τ and m_0/m_1, such that with probability at least $1-2\delta$,

$$\Delta_{\mathrm{stat}}(n_{\mathrm{dre}}, n_{\mathrm{cal}}, \delta) \le C\left(\Gamma_{\mathrm{dre}}(n_{\mathrm{dre}}, \delta) + \sqrt{\frac{\log(1/\delta)}{n_{\mathrm{cal}}}}\right).$$

In particular, if $\Gamma_{\mathrm{dre}}(n_{\mathrm{dre}}, \delta) \to 0$ and $n_{\mathrm{cal}} \to \infty$, then

$$\Delta_{\mathrm{stat}}(n_{\mathrm{dre}}, n_{\mathrm{cal}}, \delta) \to 0.$$

Proof of Lemma 2. Let

$$F_w^{\hat{Z}}(t) := \mathbb{E}_{\mathbf{P}_0}\left[w(\hat{Z})\mathbf{1}\{E^{\hat{Z}} \le t\}\right], \qquad F_{\hat{w}}^{\hat{Z}}(t) := \mathbb{E}_{\mathbf{P}_0}\left[\hat{w}(\hat{Z})\mathbf{1}\{E^{\hat{Z}} \le t\}\right],$$

and let $\hat{F}_{n_{\mathrm{cal}}}^{\hat{Z}}(t)$ denote the weighted empirical CDF constructed from the calibration sample. By definition, $\Delta_{\mathrm{stat}}(n_{\mathrm{dre}}, n_{\mathrm{cal}}, \delta)$ upper bounds the discrepancy between the ideal reweighted CDF and its practical approximation, so it is enough to control

$$\sup_{t\in[0,1]} \left|F_w^{\hat{Z}}(t) - \hat{F}_{n_{\mathrm{cal}}}^{\hat{Z}}(t)\right|.$$

We decompose this term as

$$\sup_{t\in[0,1]} \left|F_w^{\hat{Z}}(t) - \hat{F}_{n_{\mathrm{cal}}}^{\hat{Z}}(t)\right| \le A_1 + A_2,$$

where

$$A_1 := \sup_{t\in[0,1]} \left|F_w^{\hat{Z}}(t) - F_{\hat{w}}^{\hat{Z}}(t)\right|, \qquad A_2 := \sup_{t\in[0,1]} \left|F_{\hat{w}}^{\hat{Z}}(t) - \hat{F}_{n_{\mathrm{cal}}}^{\hat{Z}}(t)\right|.$$

We first bound A_1. Since $\mathbf{1}\{E^{\hat{Z}} \le t\} \le 1$,

$$A_1 \le \mathbb{E}_{\mathbf{P}_0}\left[|w(\hat{Z}) - \hat{w}(\hat{Z})|\right].$$

Write

$$w(\hat{Z}) = g(\eta(\hat{Z})), \qquad \hat{w}(\hat{Z}) = g(\hat{\eta}(\hat{Z})), \qquad g(u) := \frac{m_0}{m_1}\frac{u}{1-u}.$$

Under the clipping assumption $\eta(\hat{Z}), \hat{\eta}(\hat{Z}) \in [\tau, 1-\tau]$, the map g is Lipschitz on $[\tau, 1-\tau]$, with Lipschitz constant bounded by a quantity depending only on τ and m_0/m_1. Therefore,

$$A_1 \le C\,\|\hat{\eta} - \eta\|_{L_1(\mathbf{P}_0^{\hat{Z}})} \le C\,\Gamma_{\mathrm{dre}}(n_{\mathrm{dre}}, \delta).$$

Next, we bound A_2. Under the same clipping assumption, the estimated weight $\hat{w}(\hat{Z})$ is uniformly bounded by a constant depending only on τ and m_0/m_1. Hence the class of functions

$$\left\{(\hat{z}, e) \mapsto \hat{w}(\hat{z})\mathbf{1}\{e \le t\} : t \in [0,1]\right\}$$

is a bounded threshold class. By a standard concentration inequality for bounded empirical processes, there exists a constant $C > 0$ such that, with probability at least $1-\delta$,

$$A_2 \le C\sqrt{\frac{\log(1/\delta)}{n_{\mathrm{cal}}}}.$$

Combining the bounds for A_1 and A_2, and applying a union bound, we obtain that with probability at least $1-2\delta$,

$$\Delta_{\mathrm{stat}}(n_{\mathrm{dre}}, n_{\mathrm{cal}}, \delta) \le C\left(\Gamma_{\mathrm{dre}}(n_{\mathrm{dre}}, \delta) + \sqrt{\frac{\log(1/\delta)}{n_{\mathrm{cal}}}}\right),$$

where $C > 0$ depends only on τ and m_0/m_1. □

Proof of Theorem 3. Let $(X_{n+1}, Y_{n+1}) \sim \mathbf{P}_1$ be an independent test point, and let $\hat{Z}_{n+1} = \hat{z}(X_{n+1})$. Condition on the fitted extractor $\hat{z}$, the state scorer s_Z, and the data used to fit them.

Define an auxiliary target distribution $\widetilde{\mathbf{P}}_1$ over $(\hat{Z}, Y)$ by keeping the target marginal of the extracted state and replacing the conditional human-score law with the source conditional law:

$$\widetilde{\mathbf{P}}_1^{\hat{Z}} = \mathbf{P}_1^{\hat{Z}}, \qquad \widetilde{\mathbf{P}}_1(Y \mid \hat{Z}) = \mathbf{P}_0(Y \mid \hat{Z}).$$

Since $s_Z(\hat{Z})$ is a deterministic function of $\hat{Z}$, the conditional law of

$$E^{\hat{Z}} = |s_Z(\hat{Z}) - Y|$$

given $\hat{Z}$ under $\widetilde{\mathbf{P}}_1$ is the same as under $\mathbf{P}_0$. Therefore, under $\widetilde{\mathbf{P}}_1$, reweighting source calibration examples by

$$w(\hat{Z}) = \frac{d\mathbf{P}_1^{\hat{Z}}}{d\mathbf{P}_0^{\hat{Z}}}(\hat{Z})$$

is the exact representation-space covariate-shift correction for $E^{\hat{Z}}$.

On the event controlled by Lemma 2, the practical weighted empirical CDF constructed with $\hat{w}$ and n_{cal} calibration samples is within $C\Delta_{\mathrm{stat}}(n_{\mathrm{dre}}, n_{\mathrm{cal}}, \delta)$ in supremum norm of the ideal $\hat{Z}$-weighted CDF. Hence the usual quantile implication for conformal calibration gives

$$\mathbb{P}_{\widetilde{\mathbf{P}}_1}\Big(Y_{n+1} \in \widehat{\mathcal{I}}_\alpha^{\hat{Z}}(X_{n+1})\Big) \geq 1 - \alpha - C\,\Delta_{\mathrm{stat}}(n_{\mathrm{dre}}, n_{\mathrm{cal}}, \delta).$$

It remains to compare the auxiliary target distribution $\widetilde{\mathbf{P}}_1$ with the true target distribution $\mathbf{P}_1$. Fix the calibration and density-ratio estimation data. The event

$$A := \Big\{Y_{n+1} \in \widehat{\mathcal{I}}_\alpha^{\hat{Z}}(X_{n+1})\Big\}$$

depends on the test point only through $(\hat{Z}_{n+1}, Y_{n+1})$. For each state value $\hat{z}$, define

$$A_{\hat{z}} := \left\{y : y \in [s_Z(\hat{z}) - \hat{t}_\alpha(\hat{z}), s_Z(\hat{z}) + \hat{t}_\alpha(\hat{z})] \cap [0,1]\right\}.$$

Since $\mathbf{P}_1$ and $\widetilde{\mathbf{P}}_1$ have the same marginal distribution of $\hat{Z}$, we have

$$\begin{aligned}
&\Big|\mathbb{P}_{\mathbf{P}_1}(A) - \mathbb{P}_{\widetilde{\mathbf{P}}_1}(A)\Big| \\
&\quad = \left|\int \Big[\mathbf{P}_1(Y \in A_{\hat{z}} \mid \hat{Z} = \hat{z}) - \mathbf{P}_0(Y \in A_{\hat{z}} \mid \hat{Z} = \hat{z})\Big] d\mathbf{P}_1^{\hat{Z}}(\hat{z})\right| \\
&\quad \leq \int \mathrm{TV}\Big(\mathbf{P}_1(Y \mid \hat{Z} = \hat{z}), \mathbf{P}_0(Y \mid \hat{Z} = \hat{z})\Big) d\mathbf{P}_1^{\hat{Z}}(\hat{z}) = \varepsilon_{\mathrm{rep}}.
\end{aligned}$$

Combining the two bounds gives

$$\mathbb{P}_{\mathbf{P}_1}\Big(Y_{n+1} \in \widehat{\mathcal{I}}_\alpha^{\hat{Z}}(X_{n+1})\Big) \geq 1 - \alpha - \varepsilon_{\mathrm{rep}} - C\,\Delta_{\mathrm{stat}}(n_{\mathrm{dre}}, n_{\mathrm{cal}}, \delta).$$

The event from Lemma 2 holds with probability at least $1 - 2\delta$, which proves the theorem. □

Table 4. Guideline states for the helpfulness, correctness, and guidance axes.

Coordinate	State question	Allowed classes
	Helpfulness	
Z^1	Is the response about the requested task?	`wrong`, `related`, `requested`
Z^2	Does it provide the requested output type?	`missing`, `partial`, `present`
Z^3	What simple task shape matters for helpfulness?	`ordinary_or_interaction`, `structured_or_artifact`, `formal_reference_or_external`
Z^4	How many main parts are addressed?	`many_missing`, `some_missing`, `covered`
Z^5	Are there concrete task-relevant details?	`generic`, `some`, `concrete`
Z^6	Is there usable advice, instruction, or directly usable output?	`none`, `generic`, `specific`
Z^7	Does the content fit the user goal?	`poor`, `adequate_or_protocol`, `direct`
Z^8	Are explicit constraints followed?	`ignored`, `partial`, `followed_or_none`
Z^9	How much useful information is present?	`tiny`, `some`, `substantial`
Z^{10}	Is the turn handled as answerable work, limited/protocol work, or unhelpful missing output?	`unhelpful_missing`, `protocol_or_limited`, `ordinary_answer`
	Correctness	
Z^1	Does the response attempt the requested task?	`no`, `partial`, `yes`
Z^2	Is the requested output present?	`missing_or_wrong`, `partial`, `present`
Z^3	How much of the core request is covered?	`low`, `partial`, `covered`
Z^4	Can correctness be checked as ordinary prompt-following, or does it need a special check?	`ordinary_check`, `special_check`
Z^5	Is a correctness or task-following problem visible?	`none`, `uncertain_or_check_needed`, `major_problem`
Z^6	Is the visible answer sufficient without extra guessing?	`insufficient`, `partial_or_needs_check`, `sufficient`
Z^7	Are explicit format, count, reference, code, artifact, or exact constraints satisfied?	`none_or_soft`, `partial`, `satisfied`
Z^8	For special-check tasks, is the answer grounded or verifiable enough?	`not_needed`, `underverified`, `grounded_or_exact`
Z^9	Does the task require an exact or artifact-like output?	`ordinary_or_multi_answer`, `exact_or_artifact_output`
Z^{10}	Overall, using only visible correctness evidence, is the answer usable?	`unusable_or_wrong`, `partly_valid`, `apparently_valid`
	Guidance	
Z^1	How much actual guidance appears beyond validation?	`sparse`, `basic`, `developed`
Z^2	What executable step profile appears?	`low`, `concrete_steps`, `care_or_script_actions`
Z^3	What care/resource path is provided?	`none`, `general_care`, `specific_care`
Z^4	What sequence or tailoring structure appears?	`brief`, `structured`, `conditional`
Z^5	What follow-up path is offered?	`none_or_general`, `self_monitoring`, `care_path`

Table 5. Domain-transfer results on the correctness axis using combined HS1 and HS2 data under topic-level shifts.

Transfer	**SCI-BA Cov. ↑**	**SCI-BA Size ↓**	**RoWCI Cov. ↑**	**RoWCI Size ↓**
General → Math	83%±1%	0.63±0.01	**92%±2%**	**0.43±0.01**
Medical → Math	83%±0%	0.62±0.02	**90%±1%**	**0.36±0.01**
Medical → Coding	85%±1%	0.62±0.02	**92%±2%**	**0.42±0.01**

Table 6. Target-cal and RoWCI on HS2 under exchangeability. Target-cal denotes oracle conformal calibrated on the target split.

	Target-cal		**RoWCI**	
Axis	**Cov. ↑**	**Size ↓**	**Cov. ↑**	**Size ↓**
Complexity	90%±0%	0.39±0.00	90%±0%	0.40±0.00
Correctness	90%±0%	0.57±0.02	90%±1%	0.60±0.02
Helpfulness	90%±0%	0.40±0.01	91%±1%	0.41±0.01

Table 7. Example helpfulness states for one HS1 example and one HS2 example under the ten-coordinate helpfulness schema.

Coord.	State question	Extracted state for both examples
Z^1	Is the response about the requested task?	`requested`
Z^2	Does it provide the requested output type?	`present`
Z^3	What simple task shape matters for helpfulness?	`ordinary_or_interaction`
Z^4	How many main parts are addressed?	`covered`
Z^5	Are there concrete task-relevant details?	`concrete`
Z^6	Is there usable advice, instruction, or directly usable output?	`specific`
Z^7	Does the content fit the user goal?	`direct`
Z^8	Are explicit constraints followed?	`followed_or_none`
Z^9	How much useful information is present?	`substantial`
Z^{10}	Is the turn handled as answerable work, limited/protocol work, or unhelpful missing output?	`ordinary_answer`