WHEN RANDOM NUMBER CHOICES BREAK CONFIDENCE: RESTORING RELIABILITY IN LLM JUDGMENT GUARANTEE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used as automated evaluators to assess output quality and preference alignment. However, providing reliable guarantees that LLM judgments agree with human preferences remains challenging. Recent confidence-thresholding approaches offer such guarantees for pairwise comparisons, where the judge LLM selects one response from two candidates, relying on the assumption that higher estimated confidence implies lower disagreement risk with human judgments. In practice, however, the number of candidate responses often varies across inputs, requiring the judge to select one response from a variable number of candidates. Because confidence scores are derived from normalized judge model output preference probabilities, their scales are generally not comparable across different candidate counts. This mismatch can violate the monotonicity assumption underlying confidence-thresholding guarantees and thereby invalidate them. To address this issue, we introduce a kernel-based Gaussian cumulative distribution function (CDF) transformation that calibrates confidence scores across heterogeneous candidate sizes. We theoretically show that the proposed transformation approximately maps raw confidence scores onto a common $\mathrm{Uniform}(0, 1)$ scale, restoring comparability across candidate settings. Experiments across multiple datasets and judge LLMs demonstrate that our method re-builds monotonicity, improves guarantee success rates and coverage compared with existing baselines.

1 INTRODUCTION

LLMs are increasingly used as evaluators (judges) that score or rank candidate outputs to assess quality and preference alignment at scale (Zheng et al., 2023; Dubois et al., 2023; Park et al., 2025; Chiang & Lee, 2023). This paradigm can substantially reduce the cost and latency of human annotation, but it raises a central reliability question (Xiong et al., 2023; Ji et al., 2023): when an LLM judge expresses high confidence, when (and in what sense) should its decisions be trusted?

A growing body of work has begun to place LLM-as-a-judge systems on firmer statistical footing. For instance, Yadkori et al. (2024) study conformal abstention rules that selectively defer uncertain judgments, and Mohri & Hashimoto (2024) apply conformal prediction to obtain finite-sample correctness guarantees. Closest in spirit, Jung et al. (2025) derive exact disagreement-risk guarantees conditioned on a calibration set, but focus on the pairwise setting ($m = 2$). In many real deployments, however, the judge LLM must select a single human-preferred response from a variable number of candidates, where $m \geq 2$ can differ across inputs (e.g., medical question answering may generate and retain different numbers of candidate responses depending on retrieval, filtering, or system constraints). Reliability guarantees that do not explicitly accommodate variable-m decision problems are therefore insufficient for these settings, motivating methods that remain valid across heterogeneous candidate set sizes.

The previous confidence-thresholding approach implicitly relies on a monotonicity premise: higher estimated confidence should correspond to lower disagreement risk relative to human judgments. Yet this premise can break down when the number of candidates varies. Confidence scores derived from normalized preference probabilities are generally not comparable across different values of m, since the probability mass must be distributed over the available options. As a result, a fixed confidence

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

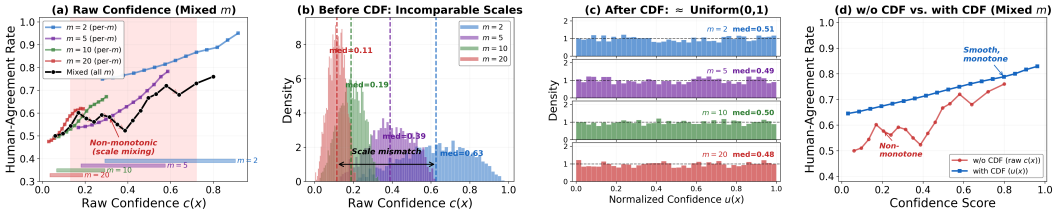

Figure 1: **Confidence–agreement relationship under mixed candidate counts** ($m \in \{2, 5, 10, 20\}$, uniformly sampled; TL;DR; Llama-3-70B). **(a)** Per-m confidence–agreement curves are monotone, but the pooled mixed-m curve is clearly non-monotone. **(b)** Raw confidence distributions differ sharply across m (medians: 0.63, 0.39, 0.19, 0.11 for $m = 2, 5, 10, 20$). **(c)** After CDF normalization, per-m distributions align to approximately $\mathrm{Uniform}(0, 1)$ (median $\approx$0.50). The y-axis is probability density; the dashed line at 1 marks the theoretical uniform level. **(d)** On pooled data, $u(x)$ restores monotonicity while raw $c(x)$ remains non-monotone. Additional examples are in Appendix D.

threshold may accept qualitatively different regimes of uncertainty across inputs with different candidate set sizes, undermining monotonicity and complicating the transfer of pairwise guarantees to the variable-candidate setting. As shown in Figure 1(a,b) and further discussed in Section 4, raw confidence estimates across varying m can exhibit markedly different distributions, thereby breaking the expected monotonic relationship between confidence and human–LLM agreement.

To address this issue, we introduce a kernel-based Gaussian CDF transformation to calibrate confidence scores across different candidate counts m. As shown in Figure 1(d), this transformation restores the monotonic relationship between confidence and disagreement risk. Moreover, we provide a theoretical analysis demonstrating that the proposed method approximately maps confidence scores obtained under different values of m onto a common $\mathrm{Uniform}(0, 1)$ scale. This normalization enables meaningful comparison of confidence estimates across heterogeneous candidate counts and provides a principled foundation for our hypothesis-testing–based human–LLM judgment guarantee framework.

Empirically, we demonstrate that, with the proposed kernel-based CDF transformation, the resulting confidence scores exhibit a substantially improved monotonic relationship between confidence and disagreement risk under diverse candidate-count settings. This improved calibration leads to higher success rates in achieving target human–LLM agreement levels across multiple datasets and judge models. A detailed discussion of related work is provided in Appendix B. An overview of the full framework is illustrated in Figure 2.

To summarize, the contributions of this work are as follows:

- ⋆ **Practical Extension.** We extend existing human–LLM judgment guarantee frameworks from the pairwise setting ($m = 2$) to the more practical scenario where the number of candidate generations varies across inputs (i.e., random m). We show that varying candidate counts can destroy the comparability of estimated confidence scores, thereby violating the monotonicity assumption that underlies confidence-thresholding guarantees.

- ⋆ **Comparable Confidence via CDF.** To address this issue, we propose a kernel-based CDF transformation to calibrate confidence scores across different candidate counts. This transformation restores the comparability of estimated confidence values and re-establishes the monotonic relationship between confidence and disagreement risk.

- ⋆ **Theoretical Analysis.** We provide theoretical analysis showing that the CDF transformation approximately maps confidence scores under different m onto a common $\mathrm{Uniform}(0, 1)$ scale, thereby restoring comparability across heterogeneous candidate settings.

- ⋆ **Experiments.** We conduct extensive experiments across multiple datasets (e.g., TL;DR, Chatbot Arena, HH-RLHF, HelpSteer3, and AlpacaEval) and judge LLMs (e.g., Gemma, MiMo, GPT-OSS, Qwen, Llama, and DeepSeek) under diverse candidate-count settings. The results demonstrate that our method effectively restores the monotonic relationship between confidence and disagreement risk and significantly improves the reliability of human–LLM agreement guarantees.

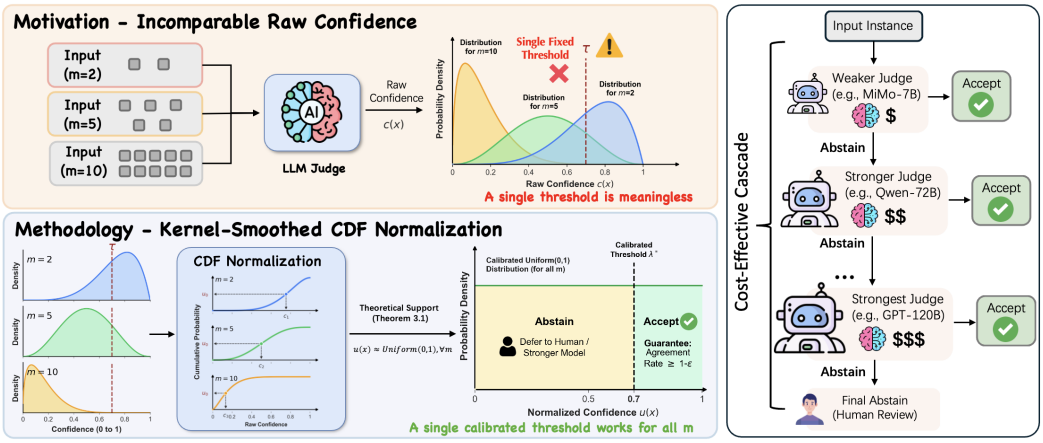

Figure 2: **Overview of the proposed framework. (Top)** Raw confidence distributions differ drastically across m, rendering a fixed threshold meaningless. **(Bottom)** A kernel-smoothed CDF maps each distribution onto $\mathrm{Uniform}(0,1)$ (Theorem 3.1); threshold λ^* partitions instances into Accept (agreement $\geq 1-\varepsilon$) or Abstain. **(Right)** Abstained instances are escalated through stronger judges until accepted or deferred to human review.

2 Preliminary

2.1 Problem Setup

Let $f_{\mathrm{LM}} : \mathcal{X} \to \mathcal{Y}$ denote an LLM judge. Each input $x \in \mathcal{X}$ consists of a query together with a variable number $m \in \mathbb{N}$ of candidate responses $\{g_1, \ldots, g_m\}$. The true label $y \in \mathcal{Y} = \{1, \ldots, m\}$ represents the human-preferred index among the m candidates. We write $z = (x, m, y)$ and denote the calibration dataset by $\mathcal{S} = \{(x_i, m_i, y_i)\}_{i=1}^n$.

Given an LLM judge f_{LM}, let $\mathbb{S}(x, g_i)$ denote a scoring (raw confidence) function, e.g., the probability assigned to candidate g_i by f_{LM}, where larger values indicate stronger predicted preference for g_i (i.e., a higher likelihood of being human-best). Let $\hat{y}$ denote the predicted index by f_{LM},

$$\hat{y} = \underset{i \in \{1,\ldots,m\}}{\arg\max}\ \mathbb{S}(x, g_i).$$

In prior work (Jung et al., 2025; Xiong et al., 2023; Kadavath et al., 2022), the maximum score $\max_i \mathbb{S}(x, g_i)$ is commonly used as a confidence measure for the instance x. Other widely used measures build on this quantity, for example, the margin-based confidence (Bartlett & Wegkamp, 2008; Geifman & El-Yaniv, 2017; El-Yaniv et al., 2010) $\Delta(x) := \mathbb{S}(x, g_{\hat{y}}) - \max_{i \neq \hat{y}} \mathbb{S}(x, g_i)$, and its normalized variant

$$c(x) := \frac{\Delta(x)}{\sigma(x)}, \quad \sigma(x) = \mathrm{sd}(s_1, \ldots, s_m) + \epsilon, \tag{1}$$

where $\epsilon > 0$ is a small stabilizing constant and $s_i = \mathbb{S}(x, g_i)$.

Motivation. Although the normalized confidence measure $c(x)$ is scale-normalized, its distribution may still vary with the number of candidates m due to extreme-value effects and variations in prompt difficulty. As illustrated in Figure 1(b), changes in m can substantially distort the distribution of raw confidence scores, thereby destroying their comparability across inputs. In particular, this variability can violate the monotone confidence–agreement relationship, that is, the assumption that higher estimated confidence corresponds to lower disagreement risk with respect to human judgments, on which previous judgment guarantees rely. To address this issue and construct confidence scores that are comparable across different candidate counts m, we introduce a kernel-based CDF transformation for cross-m confidence calibration. We provide both theoretical analysis and empirical evidence to demonstrate the effectiveness of this approach.

2.2 Assumptions

Our theoretical analysis relies on the following standard and widely adopted assumptions.

Assumption 2.1 *Calibration examples $z_1, \ldots, z_n$ and a future test example z_{n+1} are exchangeable.*

Assumption 2.2 *For every x and $\{g_1, ..., g_m\}$, the human-best preference label y is unique.*

Assumption 2.3 *Conditional on m, the distribution of $c(x)$ has a continuous CDF (i.e., $F_m(t) = \mathbb{P}(c(x) \le t \mid m)$) and admits a density f_m satisfying $\sup_{t\in\mathbb{R}} f_m(t) \le f_{\max,m} < \infty$.*

3 Human-LLM Agreement Guarantee with CDF-Normalized Confidence

In this section, we first introduce the kernel-based CDF transformation that maps confidence scores obtained under different values of m onto a common $\mathrm{Uniform}(0,1)$ scale. We then incorporate the calibrated confidence into the guarantee framework developed by Jung et al. (2025).

3.1 Confidence Normalization via Kernel-Smoothed CDF

For calibration data $z_j = (x_j, m_j, y_j)$, let $l \in \{2, \ldots, L\}$, we define the set of calibration indices

$$\mathcal{I}_l = \{j \in \{1, \ldots, n\} : m_j = l\}, \qquad n_l = |\mathcal{I}_l|. \tag{2}$$

For each l, define the kernel-smoothed CDF estimator

$$\widetilde{F}_l(t) \;:=\; \frac{1}{n_l} \sum_{j\in\mathcal{I}_l} \Phi\!\left(\frac{t - c(x_j)}{h_{m_j}}\right), \tag{3}$$

where $\Phi(\cdot)$ is the standard Gaussian CDF and $h_{m_j} > 0$ is a bandwidth (defined in Section 4). This is a smooth approximation to the empirical CDF that reduces discretization effects when n_l is modest.

Then, for any test example (x, m, y), we define the normalized confidence as

$$u(x) \;:=\; \widetilde{F}_m\big(c(x)\big) \in [0, 1]. \tag{4}$$

Intuitively, $u(x)$ is the percentile of the raw confidence $c(x)$ relative to calibration prompts with similar candidate counts. This transformation promotes comparability across varying m. In the following theorem, we prove the approximate comparability of our kernel-smoothed CDF method.

Theorem 3.1 (Approximate Comparability) *For a fresh test example (x, m, y) with the same conditional law, define the CDF-normalized score as $u(x) = \widetilde{F}_m(c(x))$. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the calibration draw,*

$$\sup_{t\in[0,1]} \big|\mathbb{P}\big(u(x) \le t \mid m\big) - t\big| \;\le\; 2\Big(\eta_m + \sqrt{\frac{2}{\pi}}\, f_{\max,m}\, h_m\Big),$$

where

$$\eta_m = \sqrt{\frac{\log(2/\delta)}{2n_m}}.$$

In particular, conditional on a specific m, the distribution of $u(x)$ is $\mathcal{O}(n_m^{-1/2} + h_m)$-close to $\mathrm{Uniform}(0, 1)$ in Kolmogorov distance.

Proof. See Appendix C. □

Remark. *The key idea of the theorem is straightforward: although the raw confidence score $c(x)$ may exhibit different scales or distributions across candidate counts m, mapping it through its conditional CDF transforms it into an approximately uniform random variable. If the true conditional CDF F_m were known, then $F_m(c(x))$ would follow a $\mathrm{Uniform}(0, 1)$ distribution by the probability integral transform. In practice, however, F_m is unknown and must be estimated from calibration data. The Dvoretzky–Kiefer–Wolfowitz (DKW) inequality guarantees that the empirical CDF uniformly approximates the true CDF, implying that the normalized score $u(x)$ is close to*

$\mathrm{Uniform}(0,1)$ *up to an error that shrinks as* $\mathcal{O}(n_m^{-1/2})$ *plus an* $\mathcal{O}(h_m)$ *kernel smoothing bias. Intuitively, the CDF transformation converts heterogeneous confidence scales into comparable percentiles within each candidate-count slice. After this transformation, a threshold such as* $u(x) \geq \lambda$ *has the same statistical meaning across different* m*, restoring cross-pool comparability.*

3.2 Human-LLM Agreement Guarantee

Following Jung et al. (2025), we build the human-LLM agreement guarantee based on the fixed-sequence testing method. For a threshold λ, define the accepted-and-covered index set

$$A(\lambda) \;=\; \{j : u(x_j) \geq \lambda \text{ and } (x_j, m_j, y_j) \in \mathcal{S}\}, \qquad N(\lambda) = |A(\lambda)|,$$

and the number of within-area errors $Y(\lambda) = \sum_{j \in A(\lambda)} \mathbf{1}\{\hat{y}_j \neq y_j\}$. When $N(\lambda) \geq 1$, define the empirical within-area error rate $\widehat{R}(\lambda) = \frac{Y(\lambda)}{N(\lambda)}$.

Since $Y(\lambda) \mid N(\lambda) \sim \mathrm{Bin}\big(N(\lambda), R(\lambda)\big)$ under exchangeability, we define the exact $(1-\delta)$ upper confidence bound (UCB):

$$\widehat{R}^{+}(\lambda) \;:=\; \sup\Big\{r \in [0,1] : \mathbb{P}\big(\mathrm{Bin}(N(\lambda), r) \leq Y(\lambda)\big) \;\geq\; \delta\Big\}. \tag{5}$$

This matches the risk upper bound construction in selective evaluation.

Fixed-sequence threshold selection. Empirically, lowering the confidence threshold λ tends to increase the selective risk, since smaller thresholds admit more ambiguous instances. Consequently, the estimated upper bound $\widehat{R}^{+}(\lambda)$ typically increases as λ decreases. This near-monotonic behavior enables the use of fixed-sequence testing (Bauer, 1991) for threshold selection.

Specifically, we examine thresholds in descending order, starting from the largest value of λ (e.g., 0.999) and gradually decreasing it. We select the largest acceptance set whose estimated risk remains below the target level ε, namely

$$\lambda^{*} \;=\; \inf\Big\{\lambda : \widehat{R}^{+}(\lambda') \leq \varepsilon \;\text{ for all }\; \lambda' \geq \lambda\Big\}. \tag{6}$$

This procedure corresponds to the fixed-sequence testing rule commonly used in selective evaluation, which avoids the conservatism of Bonferroni-type corrections when testing over large hypothesis spaces. The resulting threshold λ^{*} leads to the following guarantee from Jung et al. (2025).

Theorem 3.2 *Let* λ^{*} *be selected by the fixed-sequence rule* (6) *using the UCB* (5)*. Then, with probability at least* $1-\delta$ *over the draw of the calibration split* $\mathcal{S}$*, we have*

$$\mathbb{P}(\hat{y} \neq y \mid u(x) \geq \lambda^{*}) \;\leq\; \varepsilon. \tag{7}$$

Proof. See Appendix C. □

Remark. *When the number of candidate responses varies across inputs, raw confidence scores derived from normalized probabilities are not directly comparable across different candidate sizes, which may violate the monotonicity assumption required by Theorem 3.2. Our CDF-based calibration transforms the raw confidence* $c(x)$ *into* $\widetilde{F}_m(c(x))$*, approximately mapping scores from different candidate settings onto a common* $\mathrm{Uniform}(0,1)$ *scale. This restores comparability across heterogeneous candidate counts, allowing the guarantee in Theorem 3.2 to remain valid under the random-candidate setting.*

4 Experiments

Section 3 provides theoretical analyses that CDF-normalized confidence scores are comparable across candidate counts and that the resulting selective decisions are reliable. We now turn to empirical evaluation: does this comparability hold in practice, and does the CDF transformation improve on raw confidence baselines? Specifically, we investigate three questions: **(Q1)** Does the CDF transformation restore distribution comparability and the monotonic confidence–agreement relationship under variable m (Section 4.2)? **(Q2)** Does the CDF-calibrated confidence yield higher guarantee success rates and coverage than previous raw confidence baselines (Section 4.3)? **(Q3)** Can the framework extend to cascaded multi-model architectures under variable m (Section 4.4)?

4.1 Experimental Setup

Below we show the shared settings; method-specific settings are introduced in each subsection.

Datasets. We conduct experiments on six benchmarks that span distinct preference-evaluation scenarios: **TL;DR** (Stiennon et al., 2020) for summarization quality, **Chatbot Arena** (Zheng et al., 2023) for open-ended dialogue preference, **HH-RLHF** (Bai et al., 2022) for helpfulness and harmlessness judgments, **AlpacaEval** (Dubois et al., 2023) for instruction-following assessment, **HelpSteer3** (Wang et al., 2025) for code, STEM, and multilingual helpfulness, and **Community Alignment** (Zhang et al., 2025) for cross-cultural preference evaluation. Each benchmark provides human-annotated ground-truth preference labels. To create the variable-m evaluation setting central to this work, we generate multi-candidate instances with $m \in \{2, 5, 10, 20\}$ per query ($\approx$3,000 instances per m value) and pool them uniformly so that each candidate count contributes an equal 25% share of the combined evaluation set. We also evaluate non-uniform m-distribution settings in Appendix K. Details of the multi-candidate construction procedure are provided in Appendix E.

Judge Models. Our judge pool comprises eleven open-weight LLMs from seven families, spanning 7B to 120B parameters: **MiMo** (7B) (Xiaomi et al., 2025), **Qwen** (8B, 32B, 72B) (Yang et al., 2025; Qwen et al., 2025), **Llama-3** (8B, 70B) (Grattafiori et al., 2024), **Ministral** (14B) (Liu et al., 2026), **DeepSeek** (16B, 67B) (Liu et al., 2024; Bi et al., 2024), **Gemma-3** (27B) (Team et al., 2025), and **GPT-OSS** (120B) (Agarwal et al., 2025).

Confidence Estimation. We adopt the Simulated Annotators protocol of Jung et al. (2025): each instance is evaluated by N=15 simulated annotators, each prompted with K=15 few-shot demonstrations. Details of the annotator protocol and a per-instance computational cost breakdown are provided in Appendices F and H.

Scoring Functions. We evaluate four baseline confidence scores plus the proposed CDF normalization (full definitions in Appendix F). Two top-1 baselines use the winner's score as confidence: $\max_i \mathbb{S}(x, g_i)$ (EMP) and $\max_i \mathbb{S}_{\text{Vote}}(x, g_i)$ (Vote). Two margin baselines measure the gap between the top two candidates: $c(x)$ (EMP-based, (1)) and $c_{\text{KL}}(x)$ (KL-weighted). These span different scoring families to verify that cross-m incomparability is structural rather than an artifact of a particular scoring choice. Finally, we define $u(x) = \widetilde{F}_m(c(x))$ as the CDF-normalized confidence proposed in (4), where $c(x)$ is the EMP-based confidence. In addition, results for CDF normalization applied to other base scores are provided in Appendix G.

Bandwidth Selection. The kernel-smoothed CDF estimator (3) requires specifying a per-slice bandwidth h_m. We adopt a Silverman-type rule adapted for each m-slice:

$$h_m = \alpha \cdot \widehat{\text{IQR}}_m \cdot n_m^{-1/5}, \tag{8}$$

where $\widehat{\text{IQR}}_m$ is the interquartile range of $\{c(x_j) : j \in \mathcal{I}_m\}$ and $\alpha = 0.79$ is a constant following the classical Silverman rule (Silverman, 2018). The IQR-based rule adapts the kernel width to each m-slice's scale while remaining robust to outliers; sensitivity analysis over α is in Appendix I.

Evaluation Protocol. To account for split-level variability, we adopt a repeated-splitting protocol: for every (dataset, model) pair, the pooled variable-m data are randomly partitioned into equally sized calibration and test halves, the threshold λ^* is calibrated on the first half, and performance is measured on the second. This procedure is repeated 1,000 times, and we aggregate two metrics: **(1)** *Coverage (Cov.)*: the fraction of test instances on which the system issues a decision (i.e., does not abstain), averaged across all 1,000 splits. **(2)** *Guarantee Success Rate (GSR)*: the fraction of splits for which the empirical agreement on accepted test instances meets the target level. A well-calibrated method should attain GSR $\geq 1-\delta$.

4.2 Distribution Comparability and Monotonicity

We first verify the two central empirical claims underlying our method: (i) the CDF transformation maps heterogeneous confidence distributions onto a common scale, and (ii) this restores the monotonic confidence–agreement relationship.

Table 1: **Distribution comparability and monotonicity under mixed** $m \in \{2, 5, 10, 20\}$. Max KS Dist.: maximum Kolmogorov–Smirnov distance to $\mathrm{Uniform}(0, 1)$ across m-slices (lower represents more comparable). $\mathcal{L}_{\mathrm{rank}}$: ranking loss with T=20 thresholds (lower represents more monotone).

Dataset	Judge	Max KS Dist. (↓)					$\mathcal{L}_{\mathrm{rank}}$ (↓)				
		$\max_i \mathbb{S}$	Vote	$c(x)$	c_{KL}	$u(x)$	$\max_i \mathbb{S}$	Vote	$c(x)$	c_{KL}	$u(x)$
TL;DR	MiMo-7B	0.43	0.41	0.29	0.27	**0.06**	0.39	0.38	0.25	0.22	**0.05**
	Gemma-3-27B	0.38	0.34	0.24	0.21	**0.04**	0.32	0.29	0.20	0.16	**0.03**
	GPT-OSS-120B	0.33	0.33	0.19	0.18	**0.03**	0.25	0.24	0.13	0.12	**0.02**
Chatbot Arena	MiMo-7B	0.46	0.44	0.32	0.28	**0.06**	0.42	0.41	0.28	0.23	**0.05**
	Gemma-3-27B	0.41	0.37	0.27	0.24	**0.05**	0.36	0.33	0.22	0.18	**0.04**
	GPT-OSS-120B	0.34	0.34	0.21	0.19	**0.02**	0.27	0.27	0.13	0.14	**0.02**
HH-RLHF	MiMo-7B	0.41	0.37	0.27	0.24	**0.03**	0.36	0.33	0.24	0.20	**0.02**
	Gemma-3-27B	0.36	0.34	0.22	0.19	**0.02**	0.30	0.28	0.17	0.14	**0.01**
	GPT-OSS-120B	0.31	0.29	0.17	0.17	**0.01**	0.23	0.21	0.10	0.11	**0.01**
AlpacaEval	MiMo-7B	0.44	0.42	0.30	0.26	**0.04**	0.41	0.40	0.27	0.22	**0.04**
	Gemma-3-27B	0.39	0.36	0.24	0.22	**0.03**	0.33	0.29	0.19	0.17	**0.03**
	GPT-OSS-120B	0.34	0.32	0.19	0.17	**0.02**	0.26	0.25	0.14	0.11	**0.01**
HelpSteer3	MiMo-7B	0.42	0.39	0.28	0.26	**0.05**	0.38	0.34	0.24	0.22	**0.03**
	Gemma-3-27B	0.38	0.36	0.24	0.21	**0.04**	0.30	0.29	0.18	0.14	**0.02**
	GPT-OSS-120B	0.32	0.30	0.18	0.15	**0.02**	0.24	0.22	0.12	0.10	**0.01**
Community Alignment	MiMo-7B	0.44	0.40	0.30	0.25	**0.05**	0.40	0.36	0.26	0.21	**0.04**
	Gemma-3-27B	0.40	0.37	0.25	0.22	**0.04**	0.33	0.30	0.19	0.15	**0.02**
	GPT-OSS-120B	0.33	0.31	0.19	0.16	**0.03**	0.25	0.23	0.13	0.11	**0.01**

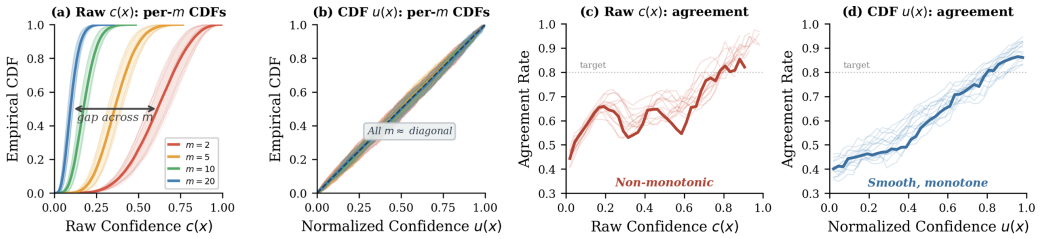

Figure 3: **Comparability and monotonicity. (a)** The CDF of raw $c(x)$ shows clear cross-m gaps. **(b)** After CDF normalization, all m-slices align to a common scale. **(c)** Agreement-vs-confidence is non-monotone for raw $c(x)$,**(d)** but smooth and monotone for $u(x)$. Thin lines denote individual dataset–judge pairs; bold lines and shaded bands show the mean ± std across all combinations (a, b). The bold line in (c, d) highlights Gemma-3-27B on TL;DR.

Comparability: To systematically quantify how well CDF normalization improves cross-m comparability, we measure the maximum Kolmogorov–Smirnov (KS) distance between each per-m confidence distribution and $\mathrm{Uniform}(0, 1)$ across $m \in \{2, 5, 10, 20\}$. Lower maximum KS distances indicate superior comparability. As shown in Table 1, all four raw baselines exhibit large KS distances: top-1 scores ($\max_i \mathbb{S}$, Vote) range from 0.29 to 0.46, while margin scores ($c(x)$, c_{KL}) range from 0.15 to 0.32, indicating heavily m-dependent distributions. In contrast, the CDF-normalized score $u(x)$ reduces the KS distance to ≤ 0.06 across all settings. This confirms a near-uniform alignment, consistent with Theorem 3.1. Furthermore, Appendix J demonstrates that this convergence adheres to the predicted $\mathcal{O}(n_m^{-1/2})$ rate; notably, even modest calibration sets ($n_m \geq 50$) achieve KS distances well below those of the raw scores.

Monotonicity: We next evaluate the restoration of the monotonic confidence-agreement relationship, where higher confidence should imply higher agreement. We discretize the confidence axis using $T = 20$ equally spaced thresholds and record the empirical agreement rate a_t at each. To quantify this, we report the ranking loss as $\mathcal{L}_{\mathrm{rank}} = \binom{T}{2}^{-1} \sum_{i<j} \mathbf{1}[a_i < a_j]$, which measures the fraction of threshold pairs violating the monotonic order. Here, $\mathcal{L}_{\mathrm{rank}} = 0$ indicates perfect monotonicity, while $\mathcal{L}_{\mathrm{rank}} \approx 0.5$ suggests a random permutation. Table 1 demonstrates that top-1 baselines suffer from high ranking losses (0.21–0.42). While the margin scores ($(c(x), c_{\mathrm{KL}})$) improve upon this (0.10–0.28), they still exhibit substantial monotonicity violations. Conversely, our CDF-normalized $u(x)$ achieves a near-zero ranking loss (≤ 0.05), validating that the transformation successfully restores the monotonic confidence–agreement relationship. This performance gap persists even with the strongest judge: for GPT-OSS-120B, raw EMP yields KS distances of 0.31–0.34, whereas

Table 2: **Single-model guarantee under mixed** $m \in \{2, 5, 10, 20\}$ (target agreement $= 0.80$, δ=0.10). Cov. and Agr. are reported as mean $\pm$ std over 1,000 random splits.

Judge	Method	TL;DR			Chatbot Arena			HH-RLHF			AlpacaEval			HelpSteer3			Comm. Align.		
		Cov.	GSR	Agr.	Cov.	GSR	Agr.	Cov.	GSR	Agr.	Cov.	GSR	Agr.	Cov.	GSR	Agr.	Cov.	GSR	Agr.
MiMo-7B	No Selection	100%	0%	36% $\pm$1.3	100%	0%	32% $\pm$1.4	100%	0%	39% $\pm$1.2	100%	0%	34% $\pm$1.4	100%	0%	33% $\pm$1.4	100%	0%	35% $\pm$1.3
	$\max_i \mathbb{S}$ (EMP)	21% $\pm$3.8	7.8%	73% $\pm$5.1	17% $\pm$4.1	3.3%	70% $\pm$5.7	25% $\pm$3.5	20.1%	76% $\pm$4.8	19% $\pm$3.9	6.3%	72% $\pm$5.4	18% $\pm$4.0	5.1%	71% $\pm$5.3	20% $\pm$3.8	5.7%	72% $\pm$5.2
	$\max_i \mathbb{S}_{\text{Vote}}$ (Vote)	24% $\pm$3.5	9.8%	74% $\pm$4.8	20% $\pm$3.8	6.6%	72% $\pm$5.3	28% $\pm$3.2	32.2%	78% $\pm$4.4	22% $\pm$3.6	12.1%	74% $\pm$5.0	21% $\pm$3.7	7.1%	73% $\pm$4.9	22% $\pm$3.6	8.4%	73% $\pm$5.0
	$c(x)$ (margin)	27% $\pm$3.3	18.4%	76% $\pm$4.5	23% $\pm$3.6	10.8%	74% $\pm$4.9	31% $\pm$3.0	23.3%	77% $\pm$4.2	25% $\pm$3.4	13.6%	75% $\pm$4.7	24% $\pm$3.5	13.8%	75% $\pm$4.6	25% $\pm$3.4	14.1%	75% $\pm$4.6
	$c_{\text{KL}}(x)$ (margin)	28% $\pm$3.2	23.6%	77% $\pm$4.3	24% $\pm$3.5	14.5%	75% $\pm$4.7	32% $\pm$2.9	39.2%	79% $\pm$3.9	26% $\pm$3.3	17.9%	76% $\pm$4.4	25% $\pm$3.3	18.6%	76% $\pm$4.5	26% $\pm$3.3	18.2%	76% $\pm$4.3
	$u(x)$ **(ours)**	36% $\pm$2.7	**93.1%**	83% $\pm$1.8	32% $\pm$2.9	**90.2%**	82% $\pm$1.5	40% $\pm$2.5	**94.7%**	84% $\pm$1.5	35% $\pm$2.8	**93.5%**	83% $\pm$1.7	34% $\pm$2.8	**90.4%**	82% $\pm$1.5	35% $\pm$2.7	**91.8%**	82% $\pm$1.6
Gemma-3-27B	No Selection	100%	0%	46% $\pm$1.1	100%	0%	42% $\pm$1.2	100%	0%	49% $\pm$1.0	100%	0%	45% $\pm$1.1	100%	0%	44% $\pm$1.2	100%	0%	45% $\pm$1.2
	$\max_i \mathbb{S}$ (EMP)	33% $\pm$3.0	17.1%	76% $\pm$4.1	28% $\pm$3.3	8.0%	74% $\pm$4.4	37% $\pm$2.7	38.8%	79% $\pm$3.7	31% $\pm$3.1	12.2%	75% $\pm$4.2	30% $\pm$3.1	11.5%	75% $\pm$4.3	31% $\pm$3.0	11.8%	75% $\pm$4.2
	$\max_i \mathbb{S}_{\text{Vote}}$ (Vote)	36% $\pm$2.8	22.1%	77% $\pm$3.8	31% $\pm$3.0	11.2%	75% $\pm$4.1	40% $\pm$2.5	49.6%	80% $\pm$3.4	34% $\pm$2.9	15.8%	76% $\pm$3.9	33% $\pm$2.9	16.3%	76% $\pm$4.0	34% $\pm$2.8	15.2%	76% $\pm$3.9
	$c(x)$ (margin)	40% $\pm$2.6	38.5%	79% $\pm$3.5	35% $\pm$2.9	15.1%	76% $\pm$3.8	44% $\pm$2.4	49.5%	80% $\pm$3.2	38% $\pm$2.7	20.4%	77% $\pm$3.6	37% $\pm$2.7	20.9%	77% $\pm$3.7	38% $\pm$2.6	19.7%	77% $\pm$3.6
	$c_{\text{KL}}(x)$ (margin)	41% $\pm$2.5	38.6%	79% $\pm$3.3	36% $\pm$2.8	19.9%	77% $\pm$3.6	45% $\pm$2.3	62.9%	81% $\pm$3.0	39% $\pm$2.6	38.9%	79% $\pm$3.4	38% $\pm$2.6	28.4%	78% $\pm$3.5	39% $\pm$2.5	27.6%	78% $\pm$3.4
	$u(x)$ **(ours)**	52% $\pm$2.2	**94.0%**	83% $\pm$1.6	47% $\pm$2.4	**91.3%**	82% $\pm$1.4	56% $\pm$2.0	**94.8%**	85% $\pm$1.4	50% $\pm$2.3	**93.7%**	83% $\pm$1.6	49% $\pm$2.3	**93.5%**	83% $\pm$1.7	50% $\pm$2.2	**92.1%**	83% $\pm$1.5
GPT-OSS-120B	No Selection	100%	0%	62% $\pm$1.0	100%	0%	58% $\pm$1.2	100%	0%	65% $\pm$0.9	100%	0%	61% $\pm$1.1	100%	0%	60% $\pm$1.1	100%	0%	59% $\pm$1.1
	$\max_i \mathbb{S}$ (EMP)	46% $\pm$2.6	28.1%	78% $\pm$3.4	41% $\pm$2.9	14.2%	76% $\pm$3.7	50% $\pm$2.4	50.3%	80% $\pm$3.1	44% $\pm$2.7	29.0%	78% $\pm$3.5	43% $\pm$2.7	19.5%	77% $\pm$3.5	42% $\pm$2.8	18.3%	77% $\pm$3.6
	$\max_i \mathbb{S}_{\text{Vote}}$ (Vote)	49% $\pm$2.4	37.6%	79% $\pm$3.1	45% $\pm$2.7	18.7%	77% $\pm$3.4	53% $\pm$2.2	63.2%	81% $\pm$2.8	47% $\pm$2.5	38.1%	79% $\pm$3.2	47% $\pm$2.5	26.1%	78% $\pm$3.2	46% $\pm$2.6	24.8%	78% $\pm$3.3
	$c(x)$ (margin)	53% $\pm$2.2	64.5%	81% $\pm$2.8	49% $\pm$2.5	26.6%	78% $\pm$3.1	57% $\pm$2.0	79.1%	82% $\pm$2.5	51% $\pm$2.3	49.5%	80% $\pm$2.9	50% $\pm$2.3	49.7%	80% $\pm$3.0	49% $\pm$2.4	46.3%	79% $\pm$3.0
	$c_{\text{KL}}(x)$ (margin)	54% $\pm$2.1	77.7%	82% $\pm$2.6	50% $\pm$2.4	35.8%	79% $\pm$2.9	58% $\pm$1.9	89.4%	83% $\pm$2.4	52% $\pm$2.2	64.6%	81% $\pm$2.7	51% $\pm$2.2	63.7%	81% $\pm$2.8	50% $\pm$2.3	58.5%	80% $\pm$2.8
	$u(x)$ **(ours)**	67% $\pm$1.8	**94.3%**	84% $\pm$1.3	62% $\pm$2.0	**92.8%**	83% $\pm$1.4	71% $\pm$1.7	**94.8%**	85% $\pm$1.2	65% $\pm$1.9	**94.1%**	84% $\pm$1.3	64% $\pm$1.9	**93.2%**	83% $\pm$1.4	63% $\pm$2.0	**92.5%**	83% $\pm$1.4

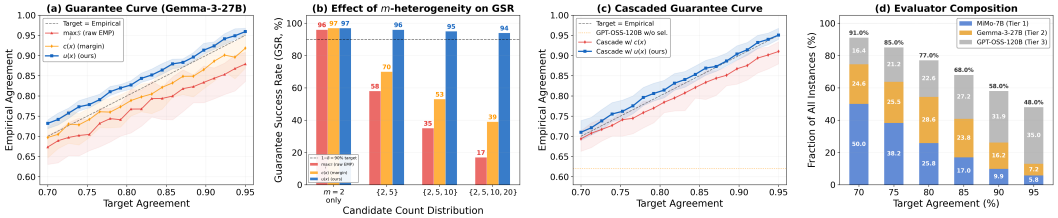

Figure 4: **Guarantee and cascade results** (TL;DR, mixed $m \in \{2, 5, 10, 20\}$; (a)–(b): Gemma-3-27B; (c)–(d): MiMo-7B $\rightarrow$ Gemma-3-27B $\rightarrow$ GPT-OSS-120B). **(a)** For a single model, $u(x)$ (blue) is always above the diagonal, while $c(x)$ and $\max_i \mathbb{S}$ are below; shaded bands show min/max over 1,000 splits. **(b)** All methods start near 97% GSR in single-m; under mixed m, baselines gradually collapse while $u(x)$ declines only to 94%. **(c)** In the cascade, empirical agreement tracks the target throughout. **(d)** In this cascade, tighter targets shift test instances from weaker to stronger tiers.

$u(x)$ suppresses them to ≤ 0.03. Visual summaries in Figure 1(a)–(d) and Figure 3 further confirm consistent improvements across all datasets and judges.

4.3 Single-Model Guarantee

Having confirmed that CDF normalization restores comparability and monotonicity, we now test whether this translates into practical guarantee improvements. We compare all five scoring functions under the same fixed-sequence threshold framework, with target agreement $= 0.80$ and δ=0.10.

CDF normalization yields valid guarantees while retaining broad coverage. Table 2 summarizes results for three representative judges across all six datasets; results for the remaining eight judges are in Appendix M. A first observation is that selective evaluation is indispensable: the unselected baseline attains at most 58–65% agreement (GPT-OSS-120B), far from the 80% target. Among selective methods, all four raw-confidence baselines fall short of the $1-\delta = 90\%$ GSR threshold (range: 3–89%), whereas the CDF-normalized $u(x)$ exceeds 90% GSR in every cell (90–95%) while covering 8–13 percentage points more instances. Because the threshold calibration is identical across scoring functions, the observed gains isolate the contribution of the CDF transformation itself.

Guarantees remain stable across varying targets and growing m-heterogeneity. Figure 4(a) traces empirical agreement as the target sweeps continuously from 70% to 95%. The $u(x)$ curve (blue) stays above the diagonal across the entire range, while both baselines remain consistently below; min/max bands confirm tighter variance for the CDF-calibrated method. Figure 4(b) isolates the role of m-heterogeneity: starting from m=2 alone ($\sim$97% GSR for all methods) and incrementally mixing in $m \in \{5, 10, 20\}$, raw-EMP collapses from 96% to 17% and the margin from 97% to 39%, whereas $u(x)$ declines only slightly from 97% to 94%. Appendix K extends this analysis to non-uniform m-distributions and the single-m extreme, confirming that $u(x)$ is robust across all tested settings and does not degrade when normalization is unnecessary.

Table 3: **Cascaded evaluation on TL;DR** (mixed $m \in \{2, 5, 10, 20\}$, target agreement $= 0.80$, δ=0.10). T_i: the i-th tier (weakest first). "Comp.": each tier's share of covered instances. Results on the remaining datasets are provided in Appendix N.

Configuration	$\max_i \mathbb{S}$		$c(x)$		$c_{\mathrm{KL}}(x)$		$u(x)$ **(ours)**				
	Cov.	GSR	Cov.	GSR	Cov.	GSR	Cov.	GSR	Comp. (%)		
									T_1	T_2	T_3
Llama-3-70B only	38% ± 2.9	22.5%	45% ± 2.5	47.3%	46% ± 2.4	48.6%	58% ± 2.1	94.3%	–	–	100.0
Qwen2.5-72B only	39% ± 2.8	24.0%	46% ± 2.4	49.1%	47% ± 2.3	51.4%	59% ± 2.0	94.5%	–	–	100.0
GPT-OSS-120B only	46% ± 2.6	28.1%	53% ± 2.2	64.5%	54% ± 2.1	77.7%	67% ± 1.8	94.3%	–	–	100.0
Qwen3-8B → 32B → Qwen2.5-72B	49% ± 3.0	25.3%	56% ± 2.7	42.8%	58% ± 2.6	48.3%	**75%** ± 1.8	**91.6%**	38.2	35.4	26.4
Llama-8B → 70B → GPT-120B	57% ± 2.8	30.6%	63% ± 2.5	52.4%	65% ± 2.4	60.1%	**81%** ± 1.7	**92.0%**	31.9	39.1	29.0
Ministral-14B → Llama-70B → GPT-120B	53% ± 2.9	27.8%	60% ± 2.6	47.5%	62% ± 2.5	54.2%	**79%** ± 1.8	**91.3%**	39.2	35.6	25.2
MiMo-7B → Gemma-27B → GPT-120B	51% ± 2.9	26.1%	58% ± 2.6	45.3%	60% ± 2.5	51.8%	**77%** ± 1.8	**91.0%**	33.5	37.1	29.4
Llama-8B → 70B	50% ± 3.1	21.4%	57% ± 2.8	38.6%	59% ± 2.7	43.9%	**74%** ± 2.1	**90.5%**	41.3	58.7	–
DeepSeek-16B → Llama-70B	47% ± 3.4	18.3%	54% ± 3.1	33.8%	56% ± 3.0	39.5%	**72%** ± 2.3	**90.1%**	44.6	55.4	–

4.4 Cascaded Architecture

Having validated per-model guarantees, we next test multi-model cascading: instances are routed through judges ordered by capacity, each accepting or escalating based on its CDF-calibrated threshold. Since $u(x)$ restores valid guarantees at every tier, reliability composes naturally across the cascade.

Table 3 reports six cascade configurations that explore same-family scaling (e.g., Qwen3-8B → Qwen3-32B → Qwen2.5-72B), cross-family combinations, and two-tier variants, together with three single-model reference points. All configurations follow the same evaluation protocol as Section 4.3; at each tier an instance is accepted if the CDF-normalized confidence exceeds the tier-specific calibrated threshold and escalated otherwise.

Cascading with $u(x)$ preserves reliability; other scoring functions do not. No cascade using uncalibrated baselines reaches the $1-\delta = 90\%$ GSR threshold: $\max_i \mathbb{S}$ ranges from 18–31%, $c(x)$ from 34–52%, and $c_{\mathrm{KL}}(x)$ from 40–60%—stacking more tiers cannot compensate for the cross-m distributional mismatch within each tier. Switching to $u(x)$ resolves this: all CDF-calibrated cascades exceed 90% GSR (90–92%), while broadening coverage to 72–81% compared with 58–67% for the best individual model. Notably, the most capable model handles fewer than 30% of accepted instances in every three-tier cascade, translating reliability gains into substantial inference savings.

Guarantee curve and workload distribution. Figure 4(c) plots the empirical agreement of a representative three-tier cascade (MiMo-7B → Gemma-3-27B → GPT-OSS-120B) as the target sweeps from 70% to 95%. With CDF calibration (blue), the cascade tracks the target across the entire range; without it (red), agreement falls below the diagonal at most operating points. Figure 4(d) visualizes how the workload redistributes across tiers: under a lenient 70% target the weakest model resolves the majority of instances; as the target tightens, more instances escalate to stronger tiers, reducing overall coverage yet maintaining the guarantee.

5 Conclusion

In this work, we investigated how to provide reliable human–LLM agreement guarantees for LLM-based evaluators in realistic variable-candidate judgment settings, where a judge must select a single preferred response from a varying number of alternatives. Although prior confidence-thresholding methods can provide guarantees in the pairwise setting, they do not directly extend to this more general scenario. Confidence scores derived from normalized judge-model preference probabilities are not necessarily comparable across candidate counts. This scale mismatch can break the monotonicity assumption required by confidence-thresholding guarantees, thereby undermining their validity.

To address this challenge, we proposed a kernel-based Gaussian CDF transformation that calibrates confidence scores across heterogeneous candidate sizes. We theoretically showed that this transformation approximately maps raw confidence scores onto a common $\mathrm{Uniform}(0, 1)$ scale, restoring cross-candidate comparability and enabling reliable confidence-based selection. Experiments across multiple datasets and judge LLMs demonstrate that our method effectively recovers monotonicity, improves guarantee success rates, and achieves stronger coverage than existing baselines.

AI USE STATEMENT

We used generative AI tools to assist in debugging parts of the experiment code and to improve the language and typesetting of the manuscript. They did not generate any data or experimental results. AI-assisted code was reviewed and tested by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies the reliability of LLM-based evaluation using publicly available benchmarks and existing human-preference annotations. No additional human annotation was conducted for this study. The benchmarks and LLM judges may reflect social or cultural biases. Our guarantees concern agreement with benchmark preferences under the stated assumptions and should be interpreted within that scope. Applications beyond the evaluated settings should account for distribution shift and retain appropriate human oversight.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Pranjal Aggarwal, Aman Madaan, Ankit Anand, Srividya Pranavi Potharaju, Swaroop Mishra, Pei Zhou, Aditya Gupta, Dheeraj Rajagopal, Karthik Kappaganthu, Yiming Yang, et al. Automix: Automatically mixing language models. *arXiv preprint arXiv:2310.12963*, 2023.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *CoRR*, abs/2107.07511, 2021. URL `https://arxiv.org/abs/2107.07511`.

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

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

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Peter L Bartlett and Marten H Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(8), 2008.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

Peter Bauer. Multiple testing in clinical trials. *Statistics in medicine*, 10(6):871–890, 1991.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

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

Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15607–15631, 2023.

C Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 2003.

Gianluca Detommaso, Martin Bertran, Riccardo Fogliato, and Aaron Roth. Multicalibration for confidence scoring in llms, 2024. URL https://arxiv.org/abs/2404.04689.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Ruhle, Laks VS Lakshmanan, and Ahmed Hassan Awadallah. Hybrid llm: Cost-efficient and quality-aware query routing. *arXiv preprint arXiv:2404.14618*, 2024.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069, 2023.

Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

Chen Feng, Minghe Shen, Ananth Balashankar, Carsten Gerner-Beuerle, and Miguel RD Rodrigues. Noisy but valid: Robust statistical evaluation of llms with imperfect judges. *arXiv preprint arXiv:2601.20913*, 2026.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6577–6595, 2024.

Patrizio Giovannotti. Calibration of natural language understanding models with venn–abers predictors. In *Conformal and Probabilistic Prediction with Applications*, pp. 55–71. PMLR, 2022.

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

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *The Innovation*, 2024.

Yu Gui, Ying Jin, and Zhimei Ren. Conformal alignment: Knowing when to trust foundation models with guarantees. *Advances in Neural Information Processing Systems*, 37:73884–73919, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: Llm judges with provable guarantees for human agreement. *ICLR*, 2025.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

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

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*, 2023.

Minjae Lee, Kyungmin Kim, Taesoo Kim, and Sangdon Park. Selective generation for controllable language models. *Advances in Neural Information Processing Systems*, 37:50494–50527, 2024.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2757–2791, 2025.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.

Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026.

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

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. *arXiv preprint arXiv:2402.10978*, 2024.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632, 2005.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Waleed Kadous, and Ion Stoica. Routellm: Learning to route llms with preference data. *arXiv preprint arXiv:2406.18665*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802, 2024.

Sangwoo Park, Matteo Zecchin, and Osvaldo Simeone. Adaptive prediction-powered autoeval with reliability and efficiency guarantees. *arXiv preprint arXiv:2505.18659*, 2025.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. Conformal language modeling. *arXiv preprint arXiv:2306.10193*, 2023.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Huanxin Sheng, Xinyi Liu, Hangfeng He, Jieyu Zhao, and Jian Kang. Analyzing uncertainty of llm-as-a-judge: Interval evaluations with conformal prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11297–11339, 2025.

Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

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

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. *arXiv preprint arXiv:2410.12784*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin,

Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, 2023.

Zailong Tian, Zhuoheng Han, Yanzhe Chen, Haozhe Xu, Xi Yang, Richeng Xuan, Houfeng Wang, and Lizi Liao. Overconfidence in llm-as-a-judge: Diagnosis and confidence-driven solution. *arXiv preprint arXiv:2508.06225*, 2025.

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

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

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. Helpsteer3-preference: Open human-annotated preference data across diverse tasks and languages. *arXiv preprint arXiv:2505.11475*, 2025.

LLM Xiaomi, Bingquan Xia, Bowen Shen, Dawei Zhu, Di Zhang, Gang Wang, Hailin Zhang, Huaqiu Liu, Jiebao Xiao, Jinhao Dong, et al. Mimo: Unlocking the reasoning potential of language model–from pretraining to posttraining. *arXiv preprint arXiv:2505.07608*, 2025.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, et al. Mitigating llm hallucinations via conformal abstention. *arXiv preprint arXiv:2405.01563*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Murong Yue, Jie Zhao, Min Zhang, Liang Du, and Ziyu Yao. Large language model cascades with mixture of thoughts representations for cost-efficient reasoning. *arXiv preprint arXiv:2310.03094*, 2023.

Lily Hong Zhang, Smitha Milli, Karen Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, et al. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. *arXiv preprint arXiv:2507.09650*, 2025.

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

A LIMITATIONS

Our theoretical analyses rely on the three assumptions stated in Section 2.2. Assumptions 2.2 (unique human-best label) and 2.3 (continuous, bounded-density confidence distribution) are mild regularity conditions that are generally satisfied in standard preference-evaluation settings. The most consequential in practice is Assumption 2.1 (exchangeability), which underlies most distribution-free inference methods: if the deployment distribution shifts substantially relative to the calibration corpus (e.g., due to a change in query domain or user population), the per-m CDF estimates may become less accurate and the approximation bound in Theorem 3.1 may loosen. While fully addressing distribution shift remains an open problem shared by the broader distribution-free inference literature, our consistent results across six benchmarks spanning diverse domains (summarization, dialogue, safety, instruction-following, code/STEM, and cross-cultural preference) suggest that the framework is broadly applicable when calibration data is drawn from the target domain.

The kernel-smoothed CDF estimator relies on having a reasonable number of calibration instances per candidate-count slice n_m. When a particular value of m is underrepresented in the calibration set, the CDF estimate for that slice may become noisier and the approximation bound in Theorem 3.1 may loosen correspondingly. Our analysis in Appendix J shows that even $n_m \geq 50$ already yields substantial improvements over raw scores, suggesting that the practical requirement is modest; for particularly rare m values, binning neighboring counts into coarser slices is a natural extension that we leave to future investigation.

Finally, the Simulated Annotators protocol used for confidence estimation requires N forward passes per instance, each with a K-shot prompt. While this cost is inherited from prior work (Jung et al., 2025) and is independent of our CDF contribution (the CDF transformation itself adds negligible post-processing overhead; see Appendix H), it may pose additional overhead in latency-sensitive deployments. The framework is agnostic to the confidence estimator, so lighter alternatives (e.g., single-pass predictive probability or verbalized confidence) can be substituted whenever they provide sufficient discriminative power for threshold calibration.

B RELATED WORK

LLM-as-a-Judge. Using LLMs as automated evaluators has become a scalable alternative to human annotation for assessing output quality and preference alignment (Zheng et al., 2023; Dubois et al., 2023; Chiang & Lee, 2023). Zheng et al. (2023) introduce MT-Bench and Chatbot Arena, establishing the LLM-judge paradigm and showing strong correlation with human rankings. Follow-up work improves the evaluation paradigm along multiple axes: chain-of-thought scoring (Liu et al., 2023) and dedicated judge models such as Prometheus (Kim et al., 2024) enhance prompting strategies and alignment with human judgments through task-specific fine-tuning; multi-judge approaches such as PoLL (Verga et al., 2024) reduce single-model cost and bias by aggregating diverse evaluators. Despite these advances, LLM judges exhibit well-documented failure modes including position bias (Wang et al., 2024; Zheng et al., 2023), self-preference bias (Panickssery et al., 2024), and verbosity bias (Li et al., 2025); several surveys and benchmarks provide comprehensive overviews (Gu et al., 2024; Tan et al., 2024). While these methods improve accuracy or reduce biases empirically, none provides a formal guarantee on the resulting error rate. Furthermore, the vast majority of these studies operate in the pairwise regime ($m=2$), whereas practical pipelines such as best-of-N decoding in RLHF (Stiennon et al., 2020; Ouyang et al., 2022; Rafailov et al., 2023; Gao et al., 2023) require judging among $m>2$ candidates, a setting where probability dilution introduces additional reliability challenges that our work addresses.

Uncertainty quantification and calibration in LLMs. Quantifying the reliability of LLM outputs is an active area of research (Geng et al., 2024). Kadavath et al. (2022) show that LLMs can partially self-assess correctness via token probabilities, and Lin et al. (2022) explore verbalized uncertainty expressions. Xiong et al. (2023) systematically evaluate confidence elicitation strategies including prompting-based and logit-based methods. Sampling-based approaches such as self-consistency (Wang et al., 2022) and semantic entropy (Kuhn et al., 2023) derive confidence from agreement across multiple sampled outputs, while Manakul et al. (2023) detect hallucinations via sampling-based self-checks and Tian et al. (2023) propose calibration strategies for RLHF-tuned models. More recently, Detommaso et al. (2024) introduce multi-calibration for LLM confidence

scoring, targeting calibration across multiple subgroups simultaneously, and Tian et al. (2025) systematically diagnose overconfidence in LLM judges and propose an ensemble-based calibration framework. Classical calibration techniques such as Platt scaling (Platt et al., 1999) and temperature scaling (Guo et al., 2017; Niculescu-Mizil & Caruana, 2005; Giovannotti, 2022) recalibrate softmax outputs but treat all inputs identically regardless of structural differences such as candidate count. In contrast, our CDF-based approach performs per-m-slice calibration, addressing the distributional shift induced by varying candidate counts—a form of subgroup-level miscalibration that arises structurally rather than from model error.

Distribution-free guarantees and conformal prediction. Conformal prediction provides finite-sample coverage guarantees under exchangeability (Vovk et al., 2005; Angelopoulos & Bates, 2021; Lei et al., 2018) and has been extended to distribution-free risk control (Bates et al., 2021; Angelopoulos et al., 2022). These tools have been adapted to LLM settings along several directions: Mohri & Hashimoto (2024) obtain factuality guarantees for language generation, Quach et al. (2023) develop conformal language modeling for open-ended text, Kumar et al. (2023) adapt conformal methods to multiple-choice QA, Yadkori et al. (2024) propose conformal abstention to selectively defer uncertain LLM judgments, Sheng et al. (2025) construct prediction intervals for LLM judge scores via conformal methods, and Gui et al. (2024) provide conformal alignment guarantees for foundation models. While these methods provide formal guarantees for LLM generation and prediction tasks, they do not directly address the LLM-as-a-judge evaluation setting where the label space grows with m. Closest to our work, Jung et al. (2025) derive exact disagreement-risk guarantees via fixed-sequence hypothesis testing on a calibration set, but their framework assumes a fixed m=2. Concurrently, Badshah et al. (2026) propose SCOPE, which improves pairwise reliability through Bidirectional Preference Entropy, Park et al. (2025) study prediction-powered inference (Angelopoulos et al., 2023) for automated evaluation, and Feng et al. (2026) provide statistical guarantees for LLM certification under noisy judges. We build on the testing procedure of Jung et al. (2025) while introducing CDF normalization to extend the guarantee to the variable-m setting, a challenge not addressed by any of the above.

Selective prediction and abstention. Selective prediction allows a model to abstain when uncertain, trading coverage for reliability (Chow, 2003; El-Yaniv et al., 2010). The framework has been extended to deep neural networks (Geifman & El-Yaniv, 2017) and, more recently, to LLM generation (Lee et al., 2024). In the LLM-judge context, Jung et al. (2025) formalize this as a confidence-thresholding problem: instances with confidence above a calibrated threshold are accepted, and the remainder are deferred to human annotators or stronger models. The validity of this approach hinges on a monotonicity assumption—higher confidence should correspond to lower disagreement risk—which implicitly requires that confidence scores be comparable across all instances in the pool. Our work preserves this framework but identifies and resolves a key failure mode: when candidate counts vary across instances, raw confidence scores exhibit different scales for different m (Figure 1(b)), violating the monotonicity assumption and causing threshold-based guarantees to fail. The CDF normalization restores cross-m comparability and thereby re-enables valid selective prediction.

Cascaded and cost-efficient LLM inference. Cascading routes inputs through a sequence of models of increasing capability and cost, accepting early when confidence is high. Chen et al. (2023) (FrugalGPT) propose cost-aware routing among API-based LLMs, Aggarwal et al. (2023) automatically mix language models by routing between small and large models based on confidence, Yue et al. (2023) introduce mixture-of-thoughts cascades for cost-efficient reasoning, Ding et al. (2024) train a quality-aware router that dynamically assigns queries to small or large models, and Ong et al. (2024) learn routing policies from preference data to balance cost and quality. These methods typically assume a fixed task format (e.g., classification with a fixed label set), so the confidence scores used for routing decisions are homogeneous. In the variable-m evaluation setting, however, a single cascade tier may encounter inputs with different candidate counts, making raw confidence scores unreliable for routing. We show that CDF-calibrated confidence enables reliable cascading under variable m and that the reliability guarantees compose across tiers (Section 4.4).

Distinction from prior work. All of the above lines of work implicitly assume a fixed number of candidates m across inputs. Our work addresses the variable-m setting: we identify that heterogeneous candidate counts induce incomparable confidence scales—a structural issue orthogonal to

model choice—and resolve it with a lightweight, per-m CDF normalization that plugs into existing selective-prediction and cascade pipelines without modifying the underlying judge.

C PROOFS

Proof 1 (Proof of Theorem 3.1) *Firstly, given a specific value of m, we define the empirical CDF on the m-slice as*

$$\widehat{F}_m(t) := \frac{1}{n_m} \sum_{i \in \mathcal{I}_m} \mathbf{1}\{c(x_i) \le t\},$$

where $\mathcal{I}_m$ is defined in (2).

Fix a bandwidth $h_m > 0$ and define the Gaussian kernel-smoothed CDF estimator (as in (3)):

$$\widetilde{F}_m(t) := \frac{1}{n_m} \sum_{i \in \mathcal{I}_m} \Phi\left(\frac{t - c(x_i)}{h_m}\right) = \int \Phi\left(\frac{t-s}{h_m}\right) d\widehat{F}_m(s),$$

where Φ is the standard normal CDF.

Define the kernel-CDF-normalized score for a test point with (x, m, y) as

$$u(x) := \widetilde{F}_m(c(x)) \in [0, 1].$$

By the Dvoretzky–Kiefer–Wolfowitz inequality, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the calibration sample, we have

$$\sup_{s \in \mathbb{R}} \left|\widehat{F}_m(s) - F_m(s)\right| \le \eta_m, \qquad \eta_m = \sqrt{\frac{\log(2/\delta)}{2n_m}}, \tag{9}$$

where $F_m(s)$ is the true CDF over all samples with candidate size m. Fix a calibration draw for which (9) holds.

Define the population smoothed CDF as

$$F_m^{(h)}(t) := \int \Phi\left(\frac{t-s}{h_m}\right) dF_m(s).$$

Then

$$\widetilde{F}_m(t) - F_m(t) = \underbrace{\left(\widetilde{F}_m(t) - F_m^{(h)}(t)\right)}_{\text{stochastic (finite-}n_m\text{) term}} + \underbrace{\left(F_m^{(h)}(t) - F_m(t)\right)}_{\text{kernel bias term}}.$$

For fixed t, the map $s \mapsto \Phi\left(\frac{t-s}{h_m}\right)$ is monotone in s and has total variation 1 (it decreases from 1 to 0 as s increases). Hence, by integration by parts / bounded-variation domination,

$$\left|\widetilde{F}_m(t) - F_m^{(h)}(t)\right| = \left|\int \Phi\left(\frac{t-s}{h}\right) d(\widehat{F}_m - F_m)(s)\right| \le \sup_{s \in \mathbb{R}} |\widehat{F}_m(s) - F_m(s)|.$$

Taking supremum over t and applying (9) yields

$$\sup_{t \in \mathbb{R}} \left|\widetilde{F}_m(t) - F_m^{(h)}(t)\right| \le \eta_m. \tag{10}$$

Write $F_m^{(h)}(t) = \int \Phi\left(\frac{t-s}{h_m}\right) f_m(s)\, ds$. Then

$$F_m^{(h)}(t) - F_m(t) = \int \left[\Phi\left(\frac{t-s}{h_m}\right) - \mathbf{1}\{s \le t\}\right] f_m(s)\, ds.$$

Taking absolute values and using $\sup_s f_m(s) \le f_{\max,m}$,

$$|F_m^{(h)}(t) - F_m(t)| \le f_{\max,m} \int \left|\Phi\left(\frac{t-s}{h_m}\right) - \mathbf{1}\{s \le t\}\right| ds.$$

Change variables $u = (t-s)/h_m$, so $ds = h_m\,du$. By symmetry,

$$\int |\Phi(u) - \mathbf{1}\{u \geq 0\}|\,du = 2\int_0^\infty (1-\Phi(u))\,du = 2\cdot\frac{1}{\sqrt{2\pi}} = \sqrt{\frac{2}{\pi}}.$$

Therefore, uniformly over t,

$$\sup_{t\in\mathbb{R}} |F_m^{(h)}(t) - F_m(t)| \leq \sqrt{\frac{2}{\pi}}\, f_{\max,m}\, h_m. \tag{11}$$

Combining (10) and (11) gives

$$\sup_{t\in\mathbb{R}} |\widetilde{F}_m(t) - F_m(t)| \leq \eta_m + \sqrt{\frac{2}{\pi}}\, f_{\max,m}\, h_m := \varepsilon_m. \tag{12}$$

By continuity of F_m, we have

$$F_m(c(x)) \sim \mathrm{Uniform}(0,1).$$

From (12), we have

$$F_m(c(x)) - \varepsilon_m \leq \widetilde{F}_m(c(x)) \leq F_m(c(x)) + \varepsilon_m.$$

Fix $t \in [0,1]$. If $u(x) = \widetilde{F}_m(c(x)) \leq t$, then $F_m(c(x)) \leq t + \varepsilon_m$, so

$$\mathbb{P}(u(x) \leq t) \leq \mathbb{P}(F_m(c(x)) \leq t + \varepsilon_m) = t + \varepsilon_m.$$

Similarly, if $F_m(c(x)) \leq t - \varepsilon_m$ then $u(x) \leq t$, hence

$$\mathbb{P}(u(x) \leq t) \geq \mathbb{P}(F_m(c(x)) \leq t - \varepsilon_m) = t - \varepsilon_m.$$

Thus, for all $t \in [0,1]$,

$$|\mathbb{P}(u(x) \leq t) - t| \leq \varepsilon_m = \sqrt{\frac{2}{\pi}}\, f_{\max,m}\, h.$$

Accounting for the probability $1-\delta$ event on which (9) holds yields the stated bound in the theorem statement. □

Proof 2 (Proof of Theorem 3.2) *The argument follows the fixed-sequence testing framework of Jung et al. (2025).*

Step 1: Validity of the UCB at each threshold. *Fix any threshold $\lambda \in (0,1)$. Conditional on the number of accepted instances $N(\lambda)$, the error count $Y(\lambda) = \sum_{j\in A(\lambda)} \mathbf{1}\{\hat{y}_j \neq y_j\}$ satisfies $Y(\lambda) \mid N(\lambda) \sim \mathrm{Bin}(N(\lambda), R(\lambda))$ under the exchangeability of calibration samples, where $R(\lambda) = \mathbb{P}(\hat{y} \neq y \mid u(x) \geq \lambda)$ is the true selective risk. By construction of the UCB in (5),*

$$\mathbb{P}\big(R(\lambda) > \widehat{R}^+(\lambda)\big) \;\leq\; \delta,$$

i.e., $\widehat{R}^+(\lambda)$ is a valid $(1-\delta)$ upper confidence bound for $R(\lambda)$.

Step 2: Fixed-sequence error control. *The threshold selection rule (6) examines thresholds in a* fixed descending *order $\lambda_1 > \lambda_2 > \cdots$. At each step, we accept the null hypothesis $H_k : R(\lambda_k) \leq \varepsilon$ if $\widehat{R}^+(\lambda_k) \leq \varepsilon$, and stop at the first rejection. Because the testing sequence is fixed (not data-dependent), the fixed-sequence principle (Bauer, 1991) ensures that the family-wise error rate is controlled at level δ without any multiplicity correction.*

In particular, for the selected threshold λ^*, all hypotheses H_k with $\lambda_k \geq \lambda^*$ have been accepted. Since the family-wise error rate is at most δ, we have

$$\mathbb{P}\big(R(\lambda^*) > \varepsilon\big) \;\leq\; \delta.$$

Rewriting $R(\lambda^*) = \mathbb{P}(\hat{y} \neq y \mid u(x) \geq \lambda^*)$ yields the stated guarantee (7). □

D CONFIDENCE DISTRIBUTION AND MONOTONICITY DETAILS

Figure 1 in the main text illustrates the core motivation of our CDF transformation using a single representative setting (TL;DR, Llama-3-70B). This appendix provides (i) details on how the four-panel visualization is constructed, (ii) evidence that the scale mismatch affects all scoring functions evaluated in this paper, not only $c(x)$, and (iii) additional visualizations for other (dataset, judge) combinations demonstrating that the phenomenon generalizes broadly. All experiments in this section share the same configuration described in Section 4.1: mixed $m \in \{2, 5, 10, 20\}$ with uniform pooling, N=15 simulated annotators, and K=15 few-shot demonstrations.

Panel construction. **Panel (a)** computes the confidence–agreement curve by placing equally spaced thresholds along the confidence axis, computing the empirical agreement rate $a_t = \frac{1}{|\{i : c(x_i) \geq \tau_t\}|} \sum_{i:c(x_i)\geq\tau_t} \mathbf{1}[y_i = \hat{y}_i]$ at each threshold, and applying a moving-average smoothing window of size 3 to reduce noise. The "Mixed (all m)" curve pools data from all m-slices; the per-m curves restrict to a single m value. **Panel (b)** displays normalized histograms of the raw margin confidence $c(x)$ for each m-slice, with dashed lines indicating per-m medians. **Panel (c)** shows normalized histograms of the CDF-normalized scores $u(x)$ in 40 equally spaced bins, with the dashed line at density 1 marking the theoretical $\mathrm{Uniform}(0, 1)$ level. **Panel (d)** overlays the pooled confidence–agreement curves before and after CDF normalization to visualize the monotonicity restoration.

Scale mismatch across all scoring functions. Figure 1(b) uses the margin confidence $c(x)$ for illustration. Figure 5 shows that the scale-mismatch phenomenon is not specific to $c(x)$: all four scoring functions evaluated in this paper—$\max_i \mathbb{S}$ (EMP), $\max_i \mathbb{S}_{\mathrm{Vote}}$ (Vote), $c(x)$ (margin), and $c_{\mathrm{KL}}(x)$ (KL margin)—exhibit drastically different distributions across m. In every case, the median shifts by a factor of $5{-}6\times$ between m=2 and m=20, confirming that a direct threshold on any raw scoring function is invalid when m varies.

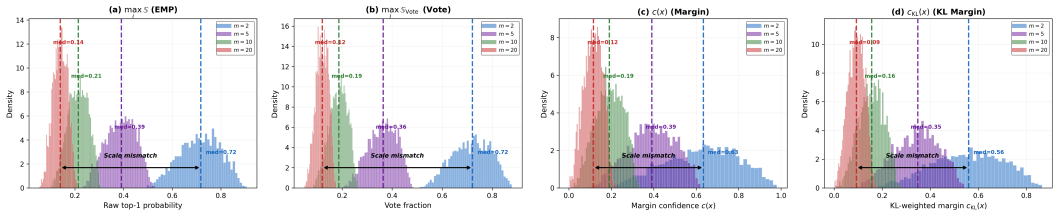

Figure 5: **Raw confidence distributions across m for all scoring functions** (TL;DR, Llama-3-70B). Each panel shows the per-m distribution of one scoring function. All four exhibit the same qualitative pattern: medians shift dramatically with m, making raw scores incomparable across different candidate counts.

Generalization across datasets and judges. Figures 6 and 7 replicate the four-panel visualization for two additional (dataset, judge) combinations: (Chatbot Arena, Qwen2.5-72B) and (HelpSteer3, Qwen3-8B). These combinations are chosen to cover both a large model (72B) on a dialogue-preference benchmark and a small model (8B) on a helpfulness benchmark with diverse query types. In all cases, the same qualitative pattern emerges: (i) raw confidence distributions exhibit drastically different scales across m (panel b), leading to a non-monotonic pooled confidence–agreement curve (panel a); (ii) after CDF normalization, all per-m distributions collapse onto approximately $\mathrm{Uniform}(0, 1)$ (panel c); and (iii) the pooled CDF-normalized curve restores smooth monotonicity (panel d).

Summary. Across all 66 (dataset, judge) combinations in our evaluation (6 datasets $\times$ 11 judges), we observe the same pattern: raw confidence distributions are strongly m-dependent, leading to non-monotonic pooled confidence–agreement curves; the kernel-based CDF transformation resolves both issues. Full quantitative results (Max KS Distance and ranking loss) for all combinations are reported in Table 1 (main text) and Table 8 (appendix).

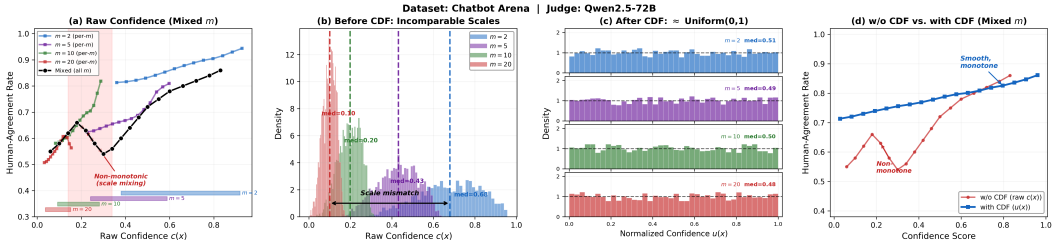

Figure 6: **Confidence distribution and monotonicity analysis** (Dataset: Chatbot Arena; Judge: Qwen2.5-72B). Same four-panel layout as Figure 1. The pattern generalizes: raw distributions are incomparable across m (b), CDF normalization produces near-uniform distributions (c), and monotonicity is restored (d).

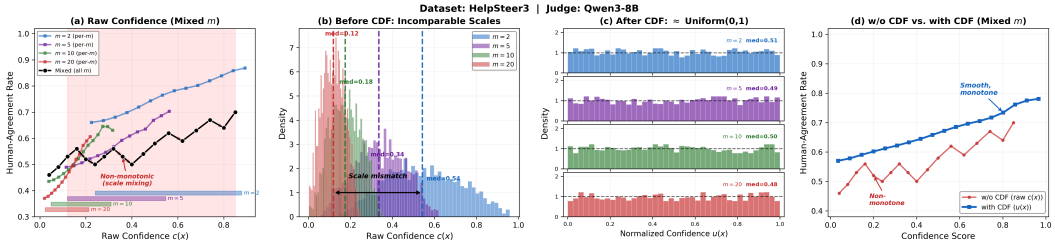

Figure 7: **Confidence distribution and monotonicity analysis** (Dataset: HelpSteer3; Judge: Qwen3-8B). Even with a smaller 8B judge on a different benchmark, the CDF transformation effectively normalizes the per-m distributions and restores the monotonic confidence–agreement relationship.

E Multi-Candidate Data Construction

For each base dataset (TL;DR, Chatbot Arena, HH-RLHF, AlpacaEval, HelpSteer3, Community Alignment), we construct multi-candidate instances as follows. Each original benchmark provides pairwise human-preference annotations: a query paired with two candidate responses and a label indicating the preferred one.

Candidate pool. For each query, we collect all available candidate responses across pairwise records and, when necessary, supplement the pool with responses drawn from other queries in the same dataset to ensure a sufficient number of distinct alternatives. Duplicate responses are removed via exact-string matching.

Instance assembly. For each target candidate count $m \in \{2, 5, 10, 20\}$, we assemble an instance by including the human-preferred response together with $m-1$ randomly sampled alternatives from the pool, then randomly shuffle the candidate order and record the position of the preferred response as the ground-truth label y. For m=2, the original pairwise instances are used directly. This procedure yields approximately 3,000 instances per $(dataset, m)$ pair. All judge models are evaluated on the identical instance set for each dataset, enabling post-hoc composition of arbitrary cascade configurations.

F Simulated Annotators Protocol and Scoring Function Definitions

Simulated Annotators protocol. We follow the Simulated Annotators framework of Jung et al. (2025). For each instance x with m candidates, N=15 simulated annotators independently evaluate the instance. Each annotator receives a distinct prompt consisting of (i) a task instruction describing the evaluation criterion, (ii) K=15 few-shot demonstrations randomly sampled from the calibration pool (excluding the current instance), each containing a query, its candidate responses, and the human-preferred label, and (iii) the target query with its m candidate responses. The judge LLM is queried with greedy decoding (temperature 0) and the top-token log-probabilities over the candidate

labels (e.g., "A", "B", ...) are extracted from the first generated token. These log-probabilities are exponentiated and normalized to form a probability vector $\mathbf{p}_j = (p_{j,1}, \ldots, p_{j,m})$ for annotator j, where $p_{j,i}$ denotes the predicted probability that candidate g_i is human-preferred. Varying the few-shot subset across annotators introduces diversity analogous to inter-annotator variation in human labeling, enabling robust confidence estimation from a single judge model.

Scoring functions. Each scoring function takes as input the probability vectors $\{\mathbf{p}_j\}_{j=1}^N$ produced by the protocol above.

Ensemble Mean Probability (EMP). $\mathbb{S}(x, g_i) = \bar{p}_i = \frac{1}{N}\sum_{j=1}^N p_{j,i}$.

Vote. $\mathbb{S}_{\text{Vote}}(x, g_i) = \frac{1}{N}\sum_{j=1}^N \mathbf{1}[\arg\max_l p_{j,l} = i]$, the fraction of annotators whose top-1 prediction is g_i.

KL-divergence-weighted voting. Unlike EMP, which averages probabilities over all annotators equally, KL-weighted voting only aggregates contributions from annotators who rank g_i first, weighting each by their probability $p_{j,i}$ (derived from the KL divergence $\text{KL}(e_i \| \mathbf{p}_j) = -\log p_{j,i}$ via exponential transform). The score further normalizes by voter consensus and cross-candidate variance:

$$\mathbb{S}_{\text{KL}}(x, g_i) = \frac{\frac{1}{N}\sum_{j=1}^N p_{j,i} \cdot \mathbf{1}\big[\arg\max_l p_{j,l} = i\big]}{\sqrt{\sigma_C^2 + \bar{v}_i}},$$

where $\bar{v}_i$ is the variance of $p_{j,i}$ among annotators who vote for g_i (smaller $\bar{v}_i$ rewards stronger consensus), and σ_C^2 is the variance of $\{\bar{p}_i\}_{i=1}^m$ across candidates (enabling cross-instance comparability).

Confidence functions. Given per-candidate scores $\mathbb{S}(x, g_i)$, we derive a scalar confidence via two types: Top-1 uses the score of the predicted winner, i.e., $\max_i \mathbb{S}(x, g_i)$; Margin uses the normalized gap $c(x) = \Delta(x)/\sigma(x)$ as defined in (1). The named configurations in Tables 2 and 3 combine these as follows: $\max_i \mathbb{S}$ uses EMP with top-1; $\max_i \mathbb{S}_{\text{Vote}}$ uses Vote with top-1; $c(x)$ uses EMP with margin; $c_{\text{KL}}(x)$ uses KL with margin.

G CDF Normalization with Alternative Base Scores

The main text reports results using $u(x) = \widetilde{F}_m(c(x))$, where $c(x)$ is the EMP-based margin confidence. Here we apply the same kernel-smoothed CDF normalization to the remaining three base scores—$\max_i \mathbb{S}$ (EMP top-1), $\max_i \mathbb{S}_{\text{Vote}}$ (Vote top-1), and $c_{\text{KL}}(x)$ (KL margin)—to verify that the benefits of CDF calibration are not specific to a single scoring function. All settings follow the same evaluation protocol as Section 4.3 (target agreement $= 0.80$, δ=0.10, 1,000 random splits).

Table 4: **CDF normalization applied to alternative base scores** (mixed $m \in \{2, 5, 10, 20\}$, target agreement $= 0.80$, δ=0.10). Each row shows a CDF-normalized confidence score; rows with blue shading reproduce the main-text $u(x)$. Raw baselines are in Table 2.

Judge	CDF base score	TL;DR		Chatbot Arena		HH-RLHF		AlpacaEval		HelpSteer3		Comm. Align.	
		Cov.	GSR	Cov.	GSR	Cov.	GSR	Cov.	GSR	Cov.	GSR	Cov.	GSR
MiMo-7B	CDF($\max_i \mathbb{S}$)	30%	90.9%	25%	88.5%	35%	92.8%	29%	91.0%	27%	88.8%	29%	90.3%
	CDF($\max_i \mathbb{S}_{\text{Vote}}$)	33%	91.3%	28%	89.1%	37%	93.4%	32%	91.7%	29%	89.5%	31%	90.8%
	CDF($c(x)$) $= u(x)$	36%	**93.1%**	32%	**90.2%**	40%	**94.7%**	35%	**93.5%**	34%	**90.4%**	35%	**91.8%**
	CDF(c_{KL})	37%	93.6%	34%	90.8%	41%	95.0%	37%	93.8%	35%	90.9%	36%	92.3%
Gemma-3-27B	CDF($\max_i \mathbb{S}$)	44%	92.0%	39%	89.5%	49%	93.2%	42%	91.8%	40%	91.3%	44%	90.0%
	CDF($\max_i \mathbb{S}_{\text{Vote}}$)	48%	92.5%	42%	90.1%	52%	93.7%	46%	92.3%	44%	91.8%	46%	90.6%
	CDF($c(x)$) $= u(x)$	52%	**94.0%**	47%	**91.3%**	56%	**94.8%**	50%	**93.7%**	49%	**93.5%**	50%	**92.1%**
	CDF(c_{KL})	53%	94.4%	49%	91.8%	57%	95.1%	51%	94.0%	50%	93.7%	52%	92.5%
GPT-OSS-120B	CDF($\max_i \mathbb{S}$)	59%	92.5%	53%	91.0%	64%	93.0%	56%	92.0%	54%	91.3%	54%	90.5%
	CDF($\max_i \mathbb{S}_{\text{Vote}}$)	62%	92.9%	56%	91.5%	67%	93.6%	60%	92.5%	57%	91.8%	58%	91.1%
	CDF($c(x)$) $= u(x)$	67%	**94.3%**	62%	**92.8%**	71%	**94.8%**	65%	**94.1%**	64%	**93.2%**	63%	**92.5%**
	CDF(c_{KL})	68%	94.7%	64%	93.2%	73%	95.2%	66%	94.3%	65%	93.6%	65%	92.9%

Results. Table 4 shows that CDF normalization consistently lifts all four base scores above the $1-\delta = 90\%$ GSR threshold across every (dataset, judge) combination, with only a few cells for the weakest base score on the hardest benchmark (Chatbot Arena, MiMo-7B) falling marginally below 90%. This confirms that the benefit of CDF calibration is not an artifact of a particular scoring function: even the simplest top-1 confidence achieves reliable guarantees once the cross-m distributional mismatch is resolved.

Effect of base-score quality. While CDF normalization resolves the cross-m mismatch for every base score, coverage still reflects the discriminative quality of the underlying confidence signal. The margin-based CDF scores (CDF($c(x)$), CDF(c_{KL})) yield 4–9 percentage points higher coverage than the top-1 CDF scores (CDF($\max_i \mathbb{S}$), CDF($\max_i \mathbb{S}_{\mathrm{Vote}}$)), because margin scores better separate confident from uncertain instances. Among the two margin variants, CDF(c_{KL}) and CDF($c(x)$) perform nearly identically (within 2 pp in Cov. and within 1 pp in GSR), suggesting that either is a suitable base for CDF calibration.

Practical implication. The main-text choice of $u(x) = \widetilde{F}_m(c(x))$ is validated: it matches the best-performing CDF variant while using the simpler EMP-based margin as its base. Practitioners who prefer a different scoring function can substitute it without losing the guarantee.

H Computational Cost Analysis

We decompose the per-test-instance cost into three components: (a) confidence estimation via model inference, (b) CDF transformation, and (c) threshold comparison.

(a) Confidence estimation (model inference). The Simulated Annotators protocol requires $N{=}15$ forward passes per instance, each with a $K{=}15$-shot prompt. This model inference step is the dominant cost and is shared with the baseline framework of Jung et al. (2025); it is entirely independent of the CDF normalization proposed in this work.

(b) CDF transformation. For a test instance with candidate count m, the kernel-smoothed CDF (3) evaluates $u(x) = \frac{1}{n_m}\sum_{j\in\mathcal{I}_m} \Phi\big(\frac{c(x)-c(x_j)}{h_m}\big)$, which involves n_m standard Gaussian CDF evaluations and a single summation. With $n_m{\approx}1{,}500$ in our default configuration, this amounts to a vectorized operation that completes in $\sim$0.02 ms on a single CPU core.

(c) Threshold comparison. The calibrated threshold λ^* is precomputed once during calibration via fixed-sequence testing. At test time, the accept/abstain decision reduces to a single comparison $u(x) \geq \lambda^*$, which is negligible.

Table 5: **Per-test-instance wall-clock breakdown.** Measured on NVIDIA GH200 GPUs + 1 CPU core with $m{=}10$ and $n_m{=}1{,}500$ calibration instances per slice.

Component	Wall-clock time			Same as baseline?
	MiMo-7B	Gemma-27B	GPT-OSS-120B	
(a) Simulated Annotators (15 passes)	0.45 s	1.80 s	3.50 s	Yes
(b) CDF transformation ($n_m{=}1{,}500$)	0.02 ms	0.02 ms	0.02 ms	No (additional)
(c) Threshold comparison	$<$0.01 ms	$<$0.01 ms	$<$0.01 ms	No (additional)
CDF overhead (b)+(c)	**0.02 ms**	**0.02 ms**	**0.02 ms**	**0.001–0.004% of (a)**

Table 5 confirms that the CDF transformation adds $\sim$0.02 ms per test instance, which is 3–4 orders of magnitude smaller than the model inference cost. The overhead is constant across judge models since it involves only CPU arithmetic on precomputed calibration scores. In the cascaded setting (Section 4.4), the CDF transformation is applied once per tier, so the total additional overhead scales linearly with the number of tiers (e.g., $\sim$0.06 ms for a three-tier cascade), remaining negligible relative to inference.

I BANDWIDTH SENSITIVITY ANALYSIS

The bandwidth multiplier α in (8) controls the bias–variance trade-off of the kernel CDF estimator: a larger α oversmooths, while a smaller α undersmooths. We examine how this choice affects downstream guarantee quality.

Protocol. We fix all experimental settings as in Sections 4.1 and 4.3 (margin score $c(x)$, mixed $m \in \{2, 5, 10, 20\}$, N=15 annotators, K=15 shots) and sweep $\alpha \in \{0.50,\ 0.65,\ 0.79,\ 0.95,\ 1.10\}$, where α=0.79 is the default Silverman-style choice used in the main text. We use $c(x)$ as the base score because the bandwidth α only affects the kernel smoothing step of the CDF transformation, making the sensitivity pattern independent of the upstream scoring function. For each α, we recompute $u(x)$ on the calibration half of every random split, re-calibrate the selective threshold, and aggregate metrics over 1,000 splits and six datasets, reporting means for **GSR** and **Cov.** for $u(x)$ (target agreement =0.80, δ=0.10). We also report **Max KS Dist.** and $\mathcal{L}_{\text{rank}}$ for $u(x)$ under the same comparability protocol as Table 1, averaged across datasets and restricted to three representative judges (MiMo-7B, Gemma-3-27B, GPT-OSS-120B) for brevity.

Table 6: **Sensitivity to bandwidth multiplier** α (margin $c(x)$, mixed m). GSR and Cov. are mean values over six datasets and 1,000 splits; Max KS and $\mathcal{L}_{\text{rank}}$ for $u(x)$ are averaged over those datasets and three judges.

α	Max KS Dist. (↓)	$\mathcal{L}_{\text{rank}}$ (↓)	GSR (↑)	Cov. (↑)
0.50	0.062	0.073	87.8%	34.8%
0.65	0.048	0.052	90.1%	36.5%
0.79 (default)	**0.038**	**0.041**	**91.6%**	**37.2%**
0.95	0.042	0.043	91.3%	37.1%
1.10	0.058	0.054	90.0%	36.4%

Table 6 shows that performance is stable in a wide band around the default: comparability and ranking monotonicity degrade only when α is noticeably small (undersmoothed CDF) or large (oversmoothed bias), while guarantee metrics track the same trend. This supports using the classical Silverman constant as a practical default without aggressive tuning.

J EFFECT OF CALIBRATION-SET SIZE ON CDF NORMALIZATION

Theorem 3.1 shows that the KS distance between the CDF-normalized confidence $u(x)$ and $\text{Uniform}(0, 1)$ is bounded by $\mathcal{O}(n_m^{-1/2} + h_m)$, where n_m is the number of calibration samples in the m-slice. A natural question is: how many calibration examples per slice are needed in practice for the CDF transformation to produce near-uniform distributions?

Protocol. We fix the margin score $c(x)$ and sweep the calibration size per m-slice over $n_m \in \{25,\ 50,\ 100,\ 200,\ 500,\ 1000\}$. For each value of n_m, we sub-sample n_m calibration instances uniformly at random from each m-slice, fit the kernel-smoothed CDF estimator (3), transform all remaining test instances, and compute the per-m KS distance to $\text{Uniform}(0, 1)$. The procedure is repeated 200 times with independent sub-samples; we report means and standard deviations. Results are averaged over six datasets and three main-text judges (MiMo-7B, Gemma-3-27B, GPT-OSS-120B).

Results. Figure 8(a) plots the per-m KS distance as a function of n_m. All four m-slices exhibit a smooth decay consistent with the $\mathcal{O}(n_m^{-1/2})$ rate predicted by Theorem 3.1 (dashed reference line). Larger m values yield slightly higher KS distances, which is expected because the kernel-bias term in the full bound depends on $f_{\max,m}$, and more concentrated raw-score distributions (higher m) have larger density peaks. Nevertheless, all curves converge rapidly and remain close to each other.

Figure 8(b) shows the worst-case (maximum over m) KS distance. Even at n_m=1,000 the worst-case KS distance is approximately 0.05, and our default calibration split already provides ∼1,500 instances per m-slice, ensuring even tighter alignment. Even with as few as n_m=50 calibration examples per

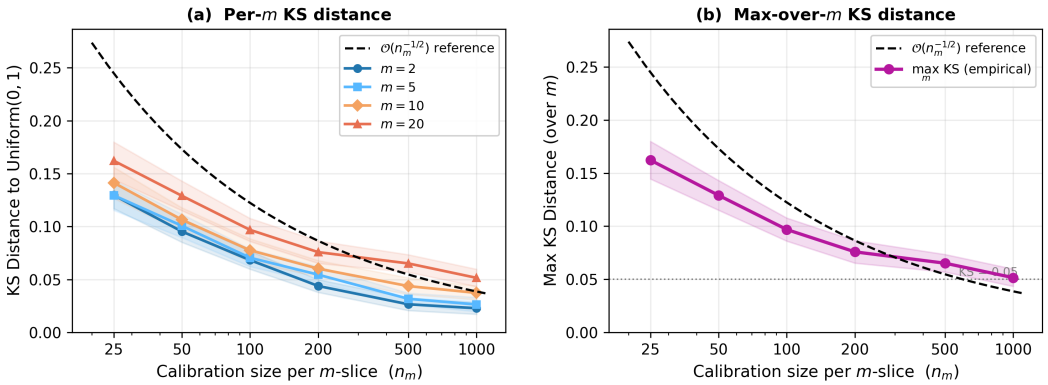

Figure 8: **KS distance to** $\mathrm{Uniform}(0,1)$ **vs. calibration size** n_m**. (a)** Per-m KS distances decay at a rate consistent with $\mathcal{O}(n_m^{-1/2})$ (dashed reference); shaded bands show ± 1 std over 200 re-samples. Larger m yields slightly higher KS due to the density-dependent bias term. **(b)** The maximum-over-m KS distance approaches 0.05 by n_m=1,000 and is already well below raw-score baselines (0.17–0.46) at n_m=50.

slice, the KS distance is below 0.13, substantially lower than the raw-score KS distances reported in Table 1 (0.17–0.46). This demonstrates that the CDF transformation is sample-efficient: a relatively small calibration set per m-slice already yields a large improvement in distributional comparability.

Table 7: **GSR and Coverage under non-uniform** m**-distributions.** Target agreement =0.80, δ=0.10. In mixed-m settings, $u(x)$ substantially outperforms the margin baseline $c(x)$; in single-m settings, where cross-m normalization is unnecessary, the two methods perform comparably (within ± 0.3 pp).

m-distribution	$c(x)$ (margin)		$u(x)$ (ours)	
	GSR	Cov.	GSR	Cov.
Uniform (default)	31.9%	38.7%	**93.0%**	**50.4%**
Skewed-toward-large	27.4%	35.3%	**92.3%**	**48.1%**
Skewed-toward-small	37.9%	42.3%	**93.8%**	**53.5%**
Single m=2	96.8%	52.1%	97.0%	52.0%
Single m=5	95.8%	44.6%	96.1%	44.9%
Single m=10	95.1%	38.4%	95.4%	38.7%
Single m=20	95.0%	33.5%	94.8%	33.2%

K Robustness to Non-Uniform m-Distributions

All main-text experiments use a uniform distribution over $m \in \{2, 5, 10, 20\}$. A natural concern is whether the CDF normalization framework degrades when the distribution of candidate counts is non-uniform, or in the extreme case where all instances share a single m (no mixing at all).

Protocol. We evaluate the CDF-normalized score $u(x)$ and the margin baseline $c(x)$ under four families of m-distributions, using the same evaluation protocol as Section 4.3 (target agreement =0.80, δ=0.10, 1,000 random splits):

1. **Uniform (default):** each $m \in \{2, 5, 10, 20\}$ contributes 25% of instances.
2. **Skewed-toward-large:** m=2, 5, 10 each contribute 15%, m=20 contributes 55%.
3. **Skewed-toward-small:** m=2 contributes 55%, m=5, 10, 20 each contribute 15%.
4. **Single-m:** all instances share the same $m \in \{2, 5, 10, 20\}$ (four separate runs).

Results are averaged over six datasets and three main-text judges (MiMo-7B, Gemma-3-27B, GPT-OSS-120B).

Table 8: **Additional per-model distribution comparability and monotonicity under mixed** $m \in \{2, 5, 10, 20\}$. Extends Table 1 to the eight judges not shown in the main text.

Dataset	Judge	Max KS Dist. (↓)					$\mathcal{L}_{\text{rank}}$ (↓)				
		$\max_i \mathbb{S}$	Vote	$c(x)$	c_{KL}	$u(x)$	$\max_i \mathbb{S}$	Vote	$c(x)$	c_{KL}	$u(x)$
TL;DR	Qwen3-8B	0.41	0.38	0.28	0.26	**0.05**	0.38	0.35	0.23	0.21	**0.04**
	Llama-3-8B	0.42	0.39	0.28	0.25	**0.06**	0.38	0.35	0.24	0.20	**0.05**
	Ministral-14B	0.40	0.37	0.27	0.23	**0.04**	0.36	0.33	0.22	0.19	**0.03**
	DeepSeek-16B	0.40	0.37	0.26	0.24	**0.05**	0.35	0.32	0.21	0.18	**0.04**
	Qwen3-32B	0.37	0.34	0.22	0.19	**0.04**	0.31	0.28	0.17	0.15	**0.03**
	DeepSeek-67B	0.36	0.33	0.23	0.21	**0.03**	0.29	0.25	0.15	0.13	**0.02**
	Llama-3-70B	0.34	0.31	0.20	0.18	**0.02**	0.27	0.26	0.14	0.11	**0.02**
	Qwen2.5-72B	0.35	0.34	0.21	0.17	**0.03**	0.29	0.26	0.16	0.13	**0.02**
Chatbot Arena	Qwen3-8B	0.44	0.42	0.30	0.28	**0.06**	0.40	0.38	0.26	0.22	**0.05**
	Llama-3-8B	0.45	0.42	0.31	0.28	**0.06**	0.41	0.38	0.27	0.23	**0.05**
	Ministral-14B	0.43	0.40	0.29	0.27	**0.05**	0.38	0.36	0.24	0.22	**0.04**
	DeepSeek-16B	0.41	0.39	0.27	0.24	**0.05**	0.36	0.33	0.22	0.20	**0.04**
	Qwen3-32B	0.40	0.37	0.26	0.22	**0.04**	0.34	0.31	0.20	0.17	**0.03**
	DeepSeek-67B	0.37	0.35	0.23	0.21	**0.04**	0.32	0.30	0.19	0.16	**0.03**
	Llama-3-70B	0.38	0.35	0.24	0.20	**0.03**	0.31	0.29	0.18	0.14	**0.02**
	Qwen2.5-72B	0.37	0.36	0.22	0.19	**0.03**	0.30	0.29	0.17	0.15	**0.02**
HH-RLHF	Qwen3-8B	0.39	0.36	0.25	0.23	**0.03**	0.35	0.32	0.21	0.18	**0.02**
	Llama-3-8B	0.40	0.37	0.26	0.23	**0.03**	0.36	0.33	0.22	0.18	**0.02**
	Ministral-14B	0.39	0.36	0.24	0.22	**0.02**	0.34	0.31	0.20	0.18	**0.02**
	DeepSeek-16B	0.38	0.36	0.24	0.20	**0.02**	0.33	0.30	0.19	0.16	**0.01**
	Qwen3-32B	0.34	0.31	0.20	0.17	**0.02**	0.27	0.24	0.14	0.12	**0.01**
	DeepSeek-67B	0.35	0.32	0.21	0.17	**0.01**	0.28	0.25	0.15	0.12	**0.01**
	Llama-3-70B	0.32	0.29	0.18	0.16	**0.01**	0.25	0.22	0.12	0.10	**0.01**
	Qwen2.5-72B	0.33	0.31	0.20	0.17	**0.02**	0.26	0.23	0.13	0.11	**0.01**
AlpacaEval	Qwen3-8B	0.42	0.39	0.28	0.24	**0.04**	0.38	0.35	0.24	0.21	**0.03**
	Llama-3-8B	0.43	0.40	0.29	0.27	**0.05**	0.40	0.37	0.26	0.23	**0.04**
	Ministral-14B	0.41	0.38	0.26	0.23	**0.03**	0.36	0.33	0.22	0.20	**0.03**
	DeepSeek-16B	0.39	0.37	0.25	0.21	**0.04**	0.34	0.31	0.20	0.16	**0.03**
	Qwen3-32B	0.37	0.34	0.22	0.20	**0.03**	0.30	0.27	0.17	0.14	**0.02**
	DeepSeek-67B	0.38	0.35	0.23	0.19	**0.02**	0.29	0.26	0.16	0.12	**0.02**
	Llama-3-70B	0.35	0.32	0.20	0.18	**0.02**	0.26	0.23	0.13	0.11	**0.01**
	Qwen2.5-72B	0.36	0.34	0.21	0.18	**0.03**	0.28	0.25	0.15	0.13	**0.02**
HelpSteer3	Qwen3-8B	0.40	0.37	0.27	0.25	**0.05**	0.37	0.34	0.22	0.20	**0.04**
	Llama-3-8B	0.41	0.38	0.27	0.24	**0.04**	0.37	0.34	0.23	0.19	**0.03**
	Ministral-14B	0.39	0.36	0.26	0.22	**0.04**	0.35	0.32	0.21	0.18	**0.03**
	DeepSeek-16B	0.38	0.36	0.25	0.23	**0.03**	0.34	0.31	0.20	0.18	**0.02**
	Qwen3-32B	0.36	0.33	0.22	0.18	**0.03**	0.29	0.26	0.15	0.12	**0.02**
	DeepSeek-67B	0.37	0.34	0.22	0.19	**0.02**	0.30	0.27	0.16	0.14	**0.02**
	Llama-3-70B	0.33	0.32	0.19	0.17	**0.02**	0.26	0.23	0.13	0.11	**0.01**
	Qwen2.5-72B	0.36	0.33	0.22	0.18	**0.03**	0.29	0.25	0.14	0.10	**0.02**
Community Alignment	Qwen3-8B	0.42	0.38	0.28	0.25	**0.05**	0.38	0.35	0.24	0.20	**0.03**
	Llama-3-8B	0.43	0.39	0.28	0.25	**0.05**	0.39	0.35	0.24	0.21	**0.04**
	Ministral-14B	0.41	0.37	0.27	0.23	**0.04**	0.37	0.33	0.22	0.19	**0.03**
	DeepSeek-16B	0.39	0.37	0.26	0.22	**0.04**	0.35	0.32	0.21	0.17	**0.03**
	Qwen3-32B	0.38	0.35	0.23	0.19	**0.03**	0.31	0.28	0.17	0.14	**0.02**
	DeepSeek-67B	0.38	0.35	0.23	0.20	**0.03**	0.30	0.27	0.17	0.13	**0.02**
	Llama-3-70B	0.35	0.33	0.21	0.18	**0.02**	0.27	0.24	0.14	0.12	**0.02**
	Qwen2.5-72B	0.37	0.34	0.22	0.19	**0.03**	0.29	0.26	0.15	0.12	**0.02**

Non-uniform mixed distributions. The first three rows of Table 7 show that the CDF transformation provides consistent and substantial improvements regardless of how m values are distributed. Under the skewed-toward-large distribution (55% of instances have m=20), the margin baseline degrades further (GSR drops from 32% to 27%) because more instances suffer from the compressed confidence scales of large m; yet $u(x)$ remains robust at $\sim$92% GSR. Under the skewed-toward-small distribution (55% m=2), the baseline improves (GSR rises to 38%) since m=2 instances are inherently easier, but $u(x)$ still dominates at $\sim$94% GSR.

Single-m (extreme case). The bottom four rows address the natural question of whether CDF normalization hurts in the absence of cross-m mismatch. When all instances share the same m, there is no distributional shift to resolve, and CDF normalization reduces to a monotone rescaling of $c(x)$. Since the CDF is a monotone transformation, it preserves the ranking of scores and therefore approximately preserves the threshold selection outcome. Accordingly, $u(x)$ achieves virtually identical GSR and coverage as $c(x)$ in every single-m setting (differences within $\pm$0.3 percentage points, well within noise), confirming that the CDF transformation does not degrade performance when normalization is unnecessary. Note also that single-m baselines already achieve high GSR (95–97%) because the cross-m mismatch problem does not arise; it is precisely the mixed-m setting that causes baselines to fail and where CDF normalization provides its largest gains.

Table 9: **Single-model guarantee: small and mid-size judges** (mixed $m \in \{2, 5, 10, 20\}$, target agreement $= 0.80$, δ=0.10). Extends Table 2.

Judge	Method	TL;DR			Chatbot Arena			HH-RLHF			AlpacaEval			HelpSteer3			Comm. Align.		
		Cov.	GSR	Agr.	Cov.	GSR	Agr.	Cov.	GSR	Agr.	Cov.	GSR	Agr.	Cov.	GSR	Agr.	Cov.	GSR	Agr.
Qwen3-8B	No Selection	100%	0%	37% ± 1.2	100%	0%	33% ± 1.4	100%	0%	40% ± 1.1	100%	0%	35% ± 1.3	100%	0%	34% ± 1.3	100%	0%	36% ± 1.3
	$\max_i \hat{S}$ (EMP)	22% ± 3.7	7.9%	73% ± 5.0	18% ± 4.0	3.5%	70% ± 5.6	26% ± 3.4	19.0%	76% ± 4.7	20% ± 3.8	6.3%	72% ± 5.3	19% ± 3.9	3.9%	71% ± 5.2	21% ± 3.7	5.5%	72% ± 5.1
	$\max_i \hat{S}_{\text{Vote}}$ (Vote)	25% ± 3.4	9.8%	74% ± 4.7	21% ± 3.7	6.3%	72% ± 5.2	29% ± 3.1	30.0%	78% ± 4.3	23% ± 3.5	10.4%	74% ± 4.9	22% ± 3.6	7.3%	73% ± 4.8	23% ± 3.5	7.4%	73% ± 4.9
	$c(x)$ (margin)	28% ± 3.2	18.4%	76% ± 4.4	24% ± 3.5	9.9%	74% ± 4.8	32% ± 2.9	31.4%	78% ± 4.1	26% ± 3.3	13.5%	75% ± 4.6	25% ± 3.4	13.0%	75% ± 4.5	26% ± 3.3	13.6%	75% ± 4.5
	$c_{\text{KL}}(x)$ (margin)	29% ± 3.1	22.8%	77% ± 4.2	25% ± 3.4	13.9%	75% ± 4.6	33% ± 2.8	38.5%	79% ± 3.8	27% ± 3.2	16.7%	76% ± 4.3	26% ± 3.2	17.1%	76% ± 4.4	27% ± 3.2	16.1%	76% ± 4.2
	$u(x)$ **(ours)**	38% ± 2.7	**93.2%**	83% ± 1.8	34% ± 2.9	**90.3%**	82% ± 1.5	42% ± 2.5	**94.7%**	84% ± 1.5	36% ± 2.8	**93.5%**	83% ± 1.7	36% ± 2.8	**90.7%**	82% ± 1.5	36% ± 2.7	**91.8%**	82% ± 1.6
Llama-3-8B	No Selection	100%	0%	38% ± 1.2	100%	0%	34% ± 1.3	100%	0%	41% ± 1.1	100%	0%	36% ± 1.3	100%	0%	35% ± 1.3	100%	0%	37% ± 1.2
	$\max_i \hat{S}$ (EMP)	23% ± 3.6	11.1%	74% ± 4.9	19% ± 3.9	4.8%	71% ± 5.4	27% ± 3.3	24.5%	77% ± 4.6	21% ± 3.7	8.5%	73% ± 5.2	20% ± 3.8	5.8%	72% ± 5.1	22% ± 3.6	7.8%	73% ± 5.0
	$\max_i \hat{S}_{\text{Vote}}$ (Vote)	26% ± 3.4	13.9%	75% ± 4.6	22% ± 3.6	8.0%	73% ± 5.1	30% ± 3.1	30.2%	78% ± 4.2	24% ± 3.5	11.1%	74% ± 4.8	23% ± 3.5	9.6%	74% ± 4.7	24% ± 3.4	10.6%	74% ± 4.8
	$c(x)$ (margin)	30% ± 3.2	23.8%	77% ± 4.3	25% ± 3.5	10.7%	74% ± 4.7	34% ± 2.9	31.5%	78% ± 4.0	28% ± 3.3	14.6%	75% ± 4.5	27% ± 3.3	13.1%	75% ± 4.4	28% ± 3.2	17.0%	76% ± 4.4
	$c_{\text{KL}}(x)$ (margin)	31% ± 3.1	24.4%	77% ± 4.1	26% ± 3.4	14.6%	75% ± 4.5	35% ± 2.8	39.8%	79% ± 3.7	29% ± 3.2	23.7%	77% ± 4.2	28% ± 3.2	17.5%	76% ± 4.3	29% ± 3.1	17.8%	76% ± 4.1
	$u(x)$ **(ours)**	39% ± 2.6	**93.3%**	83% ± 1.8	35% ± 2.8	**90.4%**	82% ± 1.5	43% ± 2.4	**94.9%**	84% ± 1.5	38% ± 2.7	**93.7%**	83% ± 1.7	37% ± 2.7	**91.0%**	82% ± 1.5	38% ± 2.6	**91.9%**	82% ± 1.6
Ministral-14B	No Selection	100%	0%	40% ± 1.1	100%	0%	36% ± 1.3	100%	0%	43% ± 1.1	100%	0%	38% ± 1.2	100%	0%	37% ± 1.3	100%	0%	39% ± 1.2
	$\max_i \hat{S}$ (EMP)	26% ± 3.5	13.8%	75% ± 4.7	21% ± 3.8	6.1%	72% ± 5.2	30% ± 3.2	25.2%	77% ± 4.4	24% ± 3.6	10.8%	74% ± 4.9	23% ± 3.6	7.4%	73% ± 4.9	24% ± 3.5	10.4%	74% ± 4.8
	$\max_i \hat{S}_{\text{Vote}}$ (Vote)	29% ± 3.2	17.4%	76% ± 4.4	24% ± 3.5	10.1%	74% ± 4.8	33% ± 2.9	37.4%	79% ± 4.0	27% ± 3.3	14.2%	75% ± 4.6	26% ± 3.4	9.5%	74% ± 4.5	27% ± 3.3	14.0%	75% ± 4.6
	$c(x)$ (margin)	32% ± 3.0	25.0%	77% ± 4.1	28% ± 3.3	12.6%	75% ± 4.5	36% ± 2.8	38.8%	79% ± 3.8	30% ± 3.1	17.9%	76% ± 4.3	29% ± 3.2	16.2%	76% ± 4.2	30% ± 3.1	17.5%	76% ± 4.2
	$c_{\text{KL}}(x)$ (margin)	33% ± 2.9	29.9%	78% ± 3.9	29% ± 3.2	16.7%	76% ± 4.3	37% ± 2.7	49.5%	80% ± 3.5	31% ± 3.0	23.5%	77% ± 4.0	30% ± 3.0	22.2%	77% ± 4.1	31% ± 3.0	22.4%	77% ± 3.9
	$u(x)$ **(ours)**	42% ± 2.5	**93.5%**	83% ± 1.7	38% ± 2.7	**90.6%**	82% ± 1.5	46% ± 2.3	**95.1%**	84% ± 1.5	41% ± 2.6	**93.9%**	83% ± 1.7	40% ± 2.6	**91.6%**	82% ± 1.6	41% ± 2.5	**92.1%**	82% ± 1.6
DeepSeek-16B	No Selection	100%	0%	43% ± 1.1	100%	0%	39% ± 1.3	100%	0%	46% ± 1.0	100%	0%	42% ± 1.2	100%	0%	41% ± 1.2	100%	0%	42% ± 1.2
	$\max_i \hat{S}$ (EMP)	29% ± 3.2	18.5%	76% ± 4.4	25% ± 3.5	7.2%	73% ± 4.8	33% ± 2.9	29.4%	78% ± 4.0	27% ± 3.3	14.0%	75% ± 4.6	26% ± 3.4	9.0%	74% ± 4.6	28% ± 3.2	12.7%	75% ± 4.5
	$\max_i \hat{S}_{\text{Vote}}$ (Vote)	32% ± 3.0	22.2%	77% ± 4.1	28% ± 3.2	12.9%	75% ± 4.5	36% ± 2.7	40.0%	79% ± 3.7	30% ± 3.1	17.6%	76% ± 4.2	29% ± 3.1	12.0%	75% ± 4.3	30% ± 3.0	16.8%	76% ± 4.2
	$c(x)$ (margin)	36% ± 2.8	30.7%	78% ± 3.8	31% ± 3.1	16.5%	76% ± 4.1	40% ± 2.6	41.6%	79% ± 3.5	34% ± 2.9	21.3%	77% ± 3.9	33% ± 2.9	16.9%	76% ± 4.0	34% ± 2.8	21.9%	77% ± 3.9
	$c_{\text{KL}}(x)$ (margin)	37% ± 2.7	31.4%	78% ± 3.6	32% ± 3.0	21.5%	77% ± 3.9	41% ± 2.5	51.0%	80% ± 3.3	35% ± 2.8	27.4%	78% ± 3.7	34% ± 2.8	22.8%	77% ± 3.8	35% ± 2.7	30.0%	78% ± 3.7
	$u(x)$ **(ours)**	47% ± 2.4	**93.7%**	83% ± 1.7	42% ± 2.5	**91.0%**	82% ± 1.4	51% ± 2.1	**95.3%**	85% ± 1.4	46% ± 2.4	**94.1%**	83% ± 1.6	44% ± 2.4	**92.6%**	83% ± 1.6	46% ± 2.4	**92.3%**	83% ± 1.5

Table 10: **Single-model guarantee: medium and large judges** (mixed $m \in \{2, 5, 10, 20\}$, target agreement $= 0.80$, δ=0.10). Extends Table 2.

Judge	Method	TL;DR			Chatbot Arena			HH-RLHF			AlpacaEval			HelpSteer3			Comm. Align.		
		Cov.	GSR	Agr.	Cov.	GSR	Agr.	Cov.	GSR	Agr.	Cov.	GSR	Agr.	Cov.	GSR	Agr.	Cov.	GSR	Agr.
Qwen3-32B	No Selection	100%	0%	48% ± 1.1	100%	0%	44% ± 1.2	100%	0%	51% ± 1.0	100%	0%	47% ± 1.1	100%	0%	46% ± 1.2	100%	0%	47% ± 1.1
	$\max_i \hat{S}$ (EMP)	35% ± 3.0	15.6%	76% ± 4.0	30% ± 3.2	8.3%	74% ± 4.3	39% ± 2.7	38.1%	79% ± 3.6	33% ± 3.1	11.1%	75% ± 4.1	32% ± 3.1	11.5%	75% ± 4.2	32% ± 3.0	10.4%	75% ± 4.1
	$\max_i \hat{S}_{\text{Vote}}$ (Vote)	38% ± 2.8	20.6%	77% ± 3.7	33% ± 3.0	10.5%	75% ± 4.0	42% ± 2.5	49.8%	80% ± 3.3	36% ± 2.9	14.8%	76% ± 3.8	35% ± 2.9	14.5%	76% ± 3.9	36% ± 2.8	14.8%	76% ± 3.8
	$c(x)$ (margin)	42% ± 2.6	39.5%	79% ± 3.4	37% ± 2.9	14.1%	76% ± 3.7	46% ± 2.4	51.1%	80% ± 3.1	40% ± 2.7	18.9%	77% ± 3.5	39% ± 2.7	19.3%	77% ± 3.6	39% ± 2.6	18.9%	77% ± 3.5
	$c_{\text{KL}}(x)$ (margin)	43% ± 2.5	41.2%	79% ± 3.2	38% ± 2.8	18.7%	77% ± 3.5	47% ± 2.2	61.3%	81% ± 2.9	41% ± 2.6	38.0%	79% ± 3.3	40% ± 2.6	26.3%	78% ± 3.4	40% ± 2.5	25.5%	78% ± 3.3
	$u(x)$ **(ours)**	54% ± 2.2	**94.0%**	83% ± 1.6	49% ± 2.4	**91.5%**	82% ± 1.4	58% ± 2.0	**94.8%**	85% ± 1.4	52% ± 2.2	**93.8%**	83% ± 1.6	51% ± 2.2	**93.5%**	83% ± 1.7	52% ± 2.2	**92.1%**	83% ± 1.5
DeepSeek-67B	No Selection	100%	0%	51% ± 1.0	100%	0%	47% ± 1.2	100%	0%	54% ± 1.0	100%	0%	50% ± 1.1	100%	0%	49% ± 1.1	100%	0%	50% ± 1.1
	$\max_i \hat{S}$ (EMP)	37% ± 2.9	22.0%	77% ± 3.9	32% ± 3.2	11.6%	75% ± 4.2	41% ± 2.6	39.1%	79% ± 3.5	35% ± 3.0	15.0%	76% ± 4.0	34% ± 3.0	16.0%	76% ± 4.0	34% ± 2.9	15.6%	76% ± 4.0
	$\max_i \hat{S}_{\text{Vote}}$ (Vote)	40% ± 2.7	27.1%	78% ± 3.6	35% ± 2.9	15.0%	76% ± 3.9	44% ± 2.4	51.0%	80% ± 3.2	38% ± 2.8	20.5%	77% ± 3.7	37% ± 2.8	20.3%	77% ± 3.8	38% ± 2.7	20.8%	77% ± 3.7
	$c(x)$ (margin)	44% ± 2.5	47.0%	80% ± 3.3	39% ± 2.8	19.9%	77% ± 3.6	48% ± 2.3	59.7%	81% ± 3.0	42% ± 2.6	26.6%	78% ± 3.4	41% ± 2.6	29.0%	78% ± 3.5	41% ± 2.5	27.9%	78% ± 3.4
	$c_{\text{KL}}(x)$ (margin)	45% ± 2.4	47.5%	80% ± 3.1	40% ± 2.7	28.1%	78% ± 3.4	49% ± 2.2	75.2%	82% ± 2.8	43% ± 2.5	47.5%	80% ± 3.2	42% ± 2.5	39.0%	79% ± 3.3	42% ± 2.4	38.6%	79% ± 3.2
	$u(x)$ **(ours)**	57% ± 2.1	**94.1%**	83% ± 1.5	52% ± 2.3	**91.8%**	82% ± 1.4	61% ± 1.9	**95.0%**	85% ± 1.3	55% ± 2.2	**94.0%**	83% ± 1.5	54% ± 2.2	**93.7%**	83% ± 1.6	54% ± 2.1	**92.2%**	83% ± 1.5
Llama-3-70B	No Selection	100%	0%	52% ± 1.0	100%	0%	48% ± 1.2	100%	0%	55% ± 1.0	100%	0%	51% ± 1.1	100%	0%	50% ± 1.1	100%	0%	51% ± 1.1
	$\max_i \hat{S}$ (EMP)	38% ± 2.9	22.5%	77% ± 3.8	33% ± 3.1	12.3%	75% ± 4.1	42% ± 2.6	40.2%	79% ± 3.5	36% ± 3.0	16.2%	76% ± 3.9	35% ± 3.0	16.3%	76% ± 4.0	35% ± 2.9	16.4%	76% ± 4.0
	$\max_i \hat{S}_{\text{Vote}}$ (Vote)	41% ± 2.6	27.5%	78% ± 3.5	36% ± 2.9	15.6%	76% ± 3.8	45% ± 2.4	52.3%	80% ± 3.2	39% ± 2.8	21.4%	77% ± 3.6	38% ± 2.8	21.4%	77% ± 3.7	38% ± 2.7	21.4%	77% ± 3.7
	$c(x)$ (margin)	45% ± 2.5	47.3%	80% ± 3.2	40% ± 2.8	21.4%	77% ± 3.5	49% ± 2.2	60.8%	81% ± 2.9	43% ± 2.5	26.9%	78% ± 3.3	42% ± 2.5	30.1%	78% ± 3.4	42% ± 2.5	28.6%	78% ± 3.4
	$c_{\text{KL}}(x)$ (margin)	46% ± 2.4	48.6%	80% ± 3.0	41% ± 2.6	29.4%	78% ± 3.3	50% ± 2.1	76.0%	82% ± 2.8	44% ± 2.5	48.0%	80% ± 3.1	43% ± 2.5	40.2%	79% ± 3.2	43% ± 2.4	39.5%	79% ± 3.2
	$u(x)$ **(ours)**	58% ± 2.1	**94.3%**	83% ± 1.5	53% ± 2.2	**91.9%**	82% ± 1.4	62% ± 1.9	**95.2%**	85% ± 1.3	56% ± 2.1	**94.2%**	83% ± 1.5	55% ± 2.1	**93.9%**	83% ± 1.6	55% ± 2.1	**92.4%**	83% ± 1.5
Qwen2.5-72B	No Selection	100%	0%	53% ± 1.0	100%	0%	49% ± 1.2	100%	0%	56% ± 1.0	100%	0%	52% ± 1.1	100%	0%	51% ± 1.1	100%	0%	52% ± 1.1
	$\max_i \hat{S}$ (EMP)	39% ± 2.8	24.0%	77% ± 3.8	34% ± 3.1	16.8%	76% ± 4.1	43% ± 2.6	41.2%	79% ± 3.4	37% ± 2.9	21.9%	77% ± 3.9	36% ± 2.9	21.8%	77% ± 3.9	36% ± 2.9	22.3%	77% ± 3.9
	$\max_i \hat{S}_{\text{Vote}}$ (Vote)	42% ± 2.6	28.7%	78% ± 3.5	37% ± 2.9	20.1%	77% ± 3.8	46% ± 2.4	53.3%	80% ± 3.1	40% ± 2.7	28.0%	78% ± 3.6	39% ± 2.7	29.0%	78% ± 3.6	39% ± 2.7	29.5%	78% ± 3.6
	$c(x)$ (margin)	46% ± 2.4	49.1%	80% ± 3.2	41% ± 2.7	27.7%	78% ± 3.5	50% ± 2.2	61.1%	81% ± 2.9	44% ± 2.5	29.2%	78% ± 3.3	43% ± 2.5	38.6%	79% ± 3.4	43% ± 2.5	31.6%	78% ± 3.3
	$c_{\text{KL}}(x)$ (margin)	47% ± 2.3	51.4%	80% ± 3.0	42% ± 2.6	36.7%	79% ± 3.3	51% ± 2.1	76.6%	82% ± 2.7	45% ± 2.4	48.7%	80% ± 3.1	44% ± 2.4	50.5%	80% ± 3.2	44% ± 2.4	49.6%	80% ± 3.1
	$u(x)$ **(ours)**	59% ± 2.0	**94.5%**	83% ± 1.5	54% ± 2.2	**92.0%**	82% ± 1.4	63% ± 1.9	**95.4%**	85% ± 1.3	57% ± 2.1	**94.4%**	83% ± 1.5	56% ± 2.1	**94.1%**	83% ± 1.6	56% ± 2.1	**92.6%**	83% ± 1.5

L Additional Comparability and Monotonicity Results

Table 1 in the main text reports Max KS Distance and ranking loss for three representative judges (MiMo-7B, Gemma-3-27B, GPT-OSS-120B). Table 8 extends this to the remaining eight judges. The same pattern holds consistently: $u(x)$ achieves near-zero KS distance and ranking loss across all (dataset, model) combinations, while all baselines exhibit substantial cross-m mismatch.

M Additional Single-Model Guarantee Results

Table 2 in the main text reports Coverage, GSR, and Agreement for three representative judges across all six datasets. Tables 9 and 10 extend these results to the remaining eight judges under the same evaluation protocol (target agreement $= 0.80$, δ=0.10, 1,000 random splits).

N Additional Cascaded Evaluation Results

Table 3 in the main text reports cascaded results on TL;DR. Tables 11 and 12 extend the cascaded evaluation to the remaining five datasets using the same cascade configurations and evaluation

Table 11: **Cascaded evaluation on Chatbot Arena and HH-RLHF** (mixed $m \in \{2, 5, 10, 20\}$, target agreement $= 0.80$, δ=0.10). Extends Table 3; tier composition is omitted as it follows the same pattern.

	Configuration	$\max_i \mathbb{S}$		$c(x)$		$c_{\mathrm{KL}}(x)$		$u(x)$ **(ours)**	
		Cov.	GSR	Cov.	GSR	Cov.	GSR	Cov.	GSR
Chatbot Arena	Llama-3-70B only	33% $_{\pm 3.1}$	12.3%	40% $_{\pm 2.8}$	21.4%	41% $_{\pm 2.6}$	29.4%	53% $_{\pm 2.2}$	91.9%
	Qwen2.5-72B only	34% $_{\pm 3.1}$	16.8%	41% $_{\pm 2.7}$	27.7%	42% $_{\pm 2.6}$	36.7%	54% $_{\pm 2.2}$	92.0%
	GPT-OSS-120B only	41% $_{\pm 2.9}$	14.2%	49% $_{\pm 2.5}$	26.6%	50% $_{\pm 2.4}$	35.8%	62% $_{\pm 2.0}$	92.8%
	Qwen3-8B → 32B → Qwen2.5-72B	44% $_{\pm 3.2}$	10.6%	51% $_{\pm 2.9}$	19.1%	53% $_{\pm 2.8}$	25.6%	**71%** $_{\pm 2.0}$	**91.1%**
	Llama-8B → 70B → GPT-120B	51% $_{\pm 3.0}$	13.1%	58% $_{\pm 2.7}$	23.6%	60% $_{\pm 2.6}$	31.9%	**77%** $_{\pm 1.9}$	**91.6%**
	Ministral-14B → Llama-70B → GPT-120B	48% $_{\pm 3.1}$	12.7%	55% $_{\pm 2.8}$	23.3%	57% $_{\pm 2.7}$	31.5%	**75%** $_{\pm 2.0}$	**90.8%**
	MiMo-7B → Gemma-27B → GPT-120B	46% $_{\pm 3.1}$	12.5%	53% $_{\pm 2.8}$	23.1%	55% $_{\pm 2.7}$	31.2%	**73%** $_{\pm 2.0}$	**90.5%**
	Llama-8B → Qwen3-32B → Qwen2.5-72B	49% $_{\pm 3.1}$	10.9%	56% $_{\pm 2.8}$	19.2%	58% $_{\pm 2.7}$	25.8%	**75%** $_{\pm 2.1}$	**90.6%**
	Llama-8B → 70B	45% $_{\pm 3.3}$	8.6%	52% $_{\pm 3.0}$	15.6%	54% $_{\pm 2.9}$	21.1%	**70%** $_{\pm 2.3}$	**90.3%**
	DeepSeek-16B → Llama-70B	42% $_{\pm 3.6}$	8.2%	49% $_{\pm 3.3}$	15.2%	51% $_{\pm 3.2}$	20.6%	**68%** $_{\pm 2.5}$	**90.1%**
HH-RLHF	Llama-3-70B only	42% $_{\pm 2.6}$	40.2%	49% $_{\pm 2.2}$	60.8%	50% $_{\pm 2.1}$	76.0%	62% $_{\pm 1.9}$	95.2%
	Qwen2.5-72B only	43% $_{\pm 2.6}$	41.2%	50% $_{\pm 2.2}$	61.1%	51% $_{\pm 2.1}$	76.6%	63% $_{\pm 1.9}$	95.4%
	GPT-OSS-120B only	50% $_{\pm 2.4}$	50.3%	57% $_{\pm 2.0}$	79.1%	58% $_{\pm 1.9}$	89.4%	71% $_{\pm 1.7}$	94.8%
	Qwen3-8B → 32B → Qwen2.5-72B	52% $_{\pm 2.9}$	37.3%	59% $_{\pm 2.6}$	57.1%	61% $_{\pm 2.5}$	64.8%	**78%** $_{\pm 1.7}$	**92.5%**
	Llama-8B → 70B → GPT-120B	59% $_{\pm 2.7}$	47.0%	66% $_{\pm 2.4}$	71.7%	68% $_{\pm 2.3}$	81.3%	**84%** $_{\pm 1.6}$	**92.8%**
	Ministral-14B → Llama-70B → GPT-120B	56% $_{\pm 2.8}$	45.8%	63% $_{\pm 2.5}$	70.9%	65% $_{\pm 2.4}$	80.4%	**82%** $_{\pm 1.7}$	**92.1%**
	MiMo-7B → Gemma-27B → GPT-120B	54% $_{\pm 2.9}$	45.0%	61% $_{\pm 2.6}$	70.2%	63% $_{\pm 2.5}$	79.6%	**80%** $_{\pm 1.8}$	**92.0%**
	Llama-8B → Qwen3-32B → Qwen2.5-72B	57% $_{\pm 2.8}$	38.3%	64% $_{\pm 2.5}$	57.5%	66% $_{\pm 2.4}$	65.2%	**82%** $_{\pm 1.8}$	**91.6%**
	Llama-8B → 70B	53% $_{\pm 3.0}$	30.5%	60% $_{\pm 2.7}$	46.8%	62% $_{\pm 2.6}$	52.3%	**77%** $_{\pm 2.0}$	**91.3%**
	DeepSeek-16B → Llama-70B	50% $_{\pm 3.3}$	29.7%	57% $_{\pm 3.0}$	45.7%	59% $_{\pm 2.9}$	51.2%	**75%** $_{\pm 2.2}$	**90.9%**

protocol (target agreement $= 0.80$, δ=0.10, 1,000 random splits). The same qualitative pattern holds across all datasets: uncalibrated baselines fail to reach the $1-\delta = 90\%$ GSR threshold in every cascade, whereas CDF-calibrated cascades consistently exceed 90% GSR while providing substantially broader coverage than any single model.

Table 12: **Cascaded evaluation on AlpacaEval, HelpSteer3, and Community Alignment** (mixed $m \in \{2, 5, 10, 20\}$, target agreement $= 0.80$, δ=0.10). Extends Table 3; tier composition is omitted as it follows the same pattern.

	Configuration	$\max_i \mathbb{S}$		$c(x)$		$c_{\mathrm{KL}}(x)$		$u(x)$ **(ours)**	
		Cov.	GSR	Cov.	GSR	Cov.	GSR	Cov.	GSR
AlpacaEval	Llama-3-70B only	36% $_{\pm 3.0}$	16.2%	43% $_{\pm 2.5}$	26.9%	44% $_{\pm 2.5}$	48.0%	56% $_{\pm 2.1}$	94.2%
	Qwen2.5-72B only	37% $_{\pm 2.9}$	21.9%	44% $_{\pm 2.5}$	29.2%	45% $_{\pm 2.4}$	48.7%	57% $_{\pm 2.1}$	94.4%
	GPT-OSS-120B only	44% $_{\pm 2.7}$	29.0%	51% $_{\pm 2.3}$	49.5%	52% $_{\pm 2.2}$	64.6%	65% $_{\pm 1.9}$	94.1%
	Qwen3-8B → 32B → Qwen2.5-72B	47% $_{\pm 3.1}$	21.2%	54% $_{\pm 2.8}$	35.1%	56% $_{\pm 2.7}$	46.0%	**73%** $_{\pm 1.9}$	**91.3%**
	Llama-8B → 70B → GPT-120B	54% $_{\pm 2.9}$	26.5%	61% $_{\pm 2.6}$	43.5%	63% $_{\pm 2.5}$	57.0%	**79%** $_{\pm 1.8}$	**91.8%**
	Ministral-14B → Llama-70B → GPT-120B	51% $_{\pm 3.0}$	25.8%	58% $_{\pm 2.7}$	42.9%	60% $_{\pm 2.6}$	56.3%	**77%** $_{\pm 1.9}$	**91.3%**
	MiMo-7B → Gemma-27B → GPT-120B	49% $_{\pm 3.0}$	25.3%	56% $_{\pm 2.7}$	42.4%	58% $_{\pm 2.6}$	55.6%	**75%** $_{\pm 1.9}$	**91.0%**
	Llama-8B → Qwen3-32B → Qwen2.5-72B	52% $_{\pm 3.0}$	21.8%	59% $_{\pm 2.7}$	35.1%	61% $_{\pm 2.6}$	46.0%	**77%** $_{\pm 2.0}$	**90.6%**
	Llama-8B → 70B	48% $_{\pm 3.2}$	17.4%	55% $_{\pm 2.9}$	28.6%	57% $_{\pm 2.8}$	37.6%	**72%** $_{\pm 2.2}$	**90.8%**
	DeepSeek-16B → Llama-70B	45% $_{\pm 3.5}$	16.7%	52% $_{\pm 3.2}$	27.9%	54% $_{\pm 3.1}$	36.6%	**70%** $_{\pm 2.4}$	**90.1%**
HelpSteer3	Llama-3-70B only	35% $_{\pm 3.0}$	16.3%	42% $_{\pm 2.5}$	30.1%	43% $_{\pm 2.5}$	40.2%	55% $_{\pm 2.1}$	93.9%
	Qwen2.5-72B only	36% $_{\pm 2.9}$	21.8%	43% $_{\pm 2.5}$	38.6%	44% $_{\pm 2.4}$	50.5%	56% $_{\pm 2.1}$	94.1%
	GPT-OSS-120B only	43% $_{\pm 2.7}$	19.5%	50% $_{\pm 2.3}$	49.7%	51% $_{\pm 2.2}$	63.7%	64% $_{\pm 1.9}$	93.2%
	Qwen3-8B → 32B → Qwen2.5-72B	46% $_{\pm 3.1}$	14.4%	53% $_{\pm 2.8}$	35.2%	55% $_{\pm 2.7}$	45.2%	**72%** $_{\pm 1.9}$	**91.1%**
	Llama-8B → 70B → GPT-120B	52% $_{\pm 2.9}$	17.8%	59% $_{\pm 2.6}$	43.8%	61% $_{\pm 2.5}$	56.1%	**78%** $_{\pm 1.8}$	**91.4%**
	Ministral-14B → Llama-70B → GPT-120B	50% $_{\pm 3.0}$	17.3%	57% $_{\pm 2.7}$	42.9%	59% $_{\pm 2.6}$	55.0%	**76%** $_{\pm 1.9}$	**90.9%**
	MiMo-7B → Gemma-27B → GPT-120B	48% $_{\pm 3.0}$	17.0%	55% $_{\pm 2.7}$	42.4%	57% $_{\pm 2.6}$	54.3%	**74%** $_{\pm 1.9}$	**90.5%**
	Llama-8B → Qwen3-32B → Qwen2.5-72B	51% $_{\pm 3.0}$	14.9%	58% $_{\pm 2.7}$	35.2%	60% $_{\pm 2.6}$	45.2%	**76%** $_{\pm 2.0}$	**90.7%**
	Llama-8B → 70B	47% $_{\pm 3.2}$	11.7%	54% $_{\pm 2.9}$	28.7%	56% $_{\pm 2.8}$	36.8%	**71%** $_{\pm 2.2}$	**90.4%**
	DeepSeek-16B → Llama-70B	44% $_{\pm 3.5}$	11.2%	51% $_{\pm 3.2}$	27.9%	53% $_{\pm 3.1}$	35.9%	**69%** $_{\pm 2.4}$	**90.2%**
Comm. Align.	Llama-3-70B only	35% $_{\pm 2.9}$	16.4%	42% $_{\pm 2.5}$	28.6%	43% $_{\pm 2.4}$	39.5%	55% $_{\pm 2.1}$	92.4%
	Qwen2.5-72B only	36% $_{\pm 2.9}$	22.3%	43% $_{\pm 2.5}$	31.6%	44% $_{\pm 2.4}$	49.6%	56% $_{\pm 2.1}$	92.6%
	GPT-OSS-120B only	42% $_{\pm 2.8}$	18.3%	49% $_{\pm 2.4}$	46.3%	50% $_{\pm 2.3}$	58.5%	63% $_{\pm 2.0}$	92.5%
	Qwen3-8B → 32B → Qwen2.5-72B	46% $_{\pm 3.1}$	15.2%	53% $_{\pm 2.8}$	34.6%	55% $_{\pm 2.7}$	44.8%	**72%** $_{\pm 1.9}$	**91.0%**
	Llama-8B → 70B → GPT-120B	53% $_{\pm 2.9}$	18.5%	60% $_{\pm 2.6}$	42.7%	62% $_{\pm 2.5}$	55.4%	**79%** $_{\pm 1.8}$	**91.5%**
	Ministral-14B → Llama-70B → GPT-120B	49% $_{\pm 3.1}$	17.8%	56% $_{\pm 2.8}$	42.1%	58% $_{\pm 2.7}$	54.7%	**75%** $_{\pm 2.0}$	**90.8%**
	MiMo-7B → Gemma-27B → GPT-120B	48% $_{\pm 3.0}$	17.3%	55% $_{\pm 2.7}$	41.6%	57% $_{\pm 2.6}$	53.8%	**74%** $_{\pm 1.9}$	**90.6%**
	Llama-8B → Qwen3-32B → Qwen2.5-72B	50% $_{\pm 3.1}$	15.5%	57% $_{\pm 2.8}$	34.8%	59% $_{\pm 2.7}$	45.1%	**75%** $_{\pm 2.0}$	**90.5%**
	Llama-8B → 70B	47% $_{\pm 3.2}$	12.1%	54% $_{\pm 2.9}$	28.1%	56% $_{\pm 2.8}$	36.4%	**71%** $_{\pm 2.2}$	**90.5%**
	DeepSeek-16B → Llama-70B	43% $_{\pm 3.6}$	11.5%	50% $_{\pm 3.3}$	27.4%	52% $_{\pm 3.2}$	35.3%	**68%** $_{\pm 2.5}$	**90.1%**