CERTIFYING NO-WINNER DECISIONS IN LARGE LANGUAGE MODEL JUDGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used as automatic judges of response quality and preference alignment. Existing guarantee-based evaluation frameworks typically assume that each instance has a unique human-preferred response, an assumption that often fails in open-ended settings where candidates may be similarly poor leaving no meaningful winner. We study reliable LLM judgment under this more general setting. Allowing no-winner outcomes introduces a key challenge: confidence scores can conflate evidence about absolute response quality, relative preference, and candidate ambiguity, complicating the identification of reliable no-winner decisions. Moreover, controlling overall selective risk does not ensure reliability separately for no-winner and response-selection decisions. We address this challenge with a two-stage certified selection framework. First, conformal prediction constructs a localized label set over candidate responses and the no-winner label, retaining the reference label with a distribution-free, finite-sample marginal coverage guarantee under exchangeability. Second, conditioned on the localized set, separate calibrated hypothesis tests certify either a no-winner or response-selection decision. The resulting framework provides high-probability selective correctness guarantees while abstaining or escalating when evidence is insufficient. Extensive experiments demonstrate the benefits of our two stage framework for no-winner recognition and characterize its effectiveness across controlled and human-rated evaluation settings.

1 INTRODUCTION

Large language models (LLMs) are increasingly used as automatic evaluators of output quality (Li et al., 2024; Gu et al., 2026; Dubois et al., 2024), preference alignment (Tennant et al., 2025; Ji et al., 2023; Deng et al., 2026), instruction following (Li et al., 2026a; Kaur et al., 2025; Yin et al., 2023), and safety (Mou et al., 2024; Kumar et al., 2023; Ji et al., 2023), offering a scalable alternative to costly human annotation. Yet their growing use raises a fundamental reliability question: when an LLM judge selects a response as best, can we trust that its judgment agrees with human preference?

Recent work addresses this question through selective evaluation (Geifman & El-Yaniv, 2017; Jung et al., 2025): rather than accepting every judgment, the evaluator accepts only predictions whose estimated confidence exceeds a calibrated threshold, enabling high-probability guarantees on human–LLM agreement (Jung et al., 2025). However, existing formulations typically focus on pairwise comparisons and assume that every instance admits a unique human-preferred response. This assumption is restrictive in open-ended evaluation. Given a query and m candidate responses, the candidates may all be poor, unsafe, irrelevant, factually incorrect. Forcing a single winner in such cases creates artificial preference labels and can overstate evaluation reliability. We therefore allow a no-winner outcome, $y \in \{0, 1, \ldots, m\}$, where $y = i > 0$ denotes that response i is human-preferred, while $y = 0$ indicates that all candidates fail a prespecified quality rubric.

Introducing a no-winner label requires assessing the reliability of both response-selection and no-winner decisions. Unlike $y = i > 0$, which identifies a specific candidate, $y = 0$ represents a global judgment over the candidate set. A reliable evaluator must therefore distinguish between two decisions: whether the evidence supports declaring that no candidate meets the prespecified quality criterion, and whether it supports selecting a particular candidate as preferred (Liu et al., 2024). When

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.

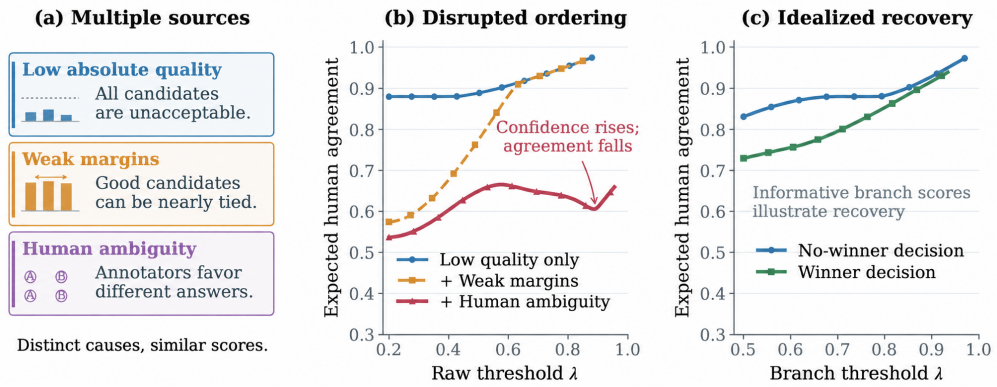


Figure 1: **Heterogeneous confidence in no-winner judgments.** (a) Low absolute quality, weak preference margins, and human preference ambiguity can produce null-confidence signals with different evidential meanings. (b) In a constructed mixture, adding heterogeneous sources makes agreement among examples above a confidence threshold nonmonotone. (c) Informative branch-specific scores can recover nondecreasing agreement for both no-winner and winner decisions. Curves are analytical illustrations, not LLM measurements; panel (c) uses idealized correctness scores and does not establish that localization alone guarantees monotonicity.

neither conclusion is sufficiently supported, the evaluator should abstain and defer the instance to a stronger judge or human evaluator.

Standard single-stage confidence thresholding does not directly address these two reliability objectives. Confidence in the no-winner label can arise from qualitatively different sources, including uniformly low response quality, small preference margins, and substantial ambiguity among candidates (Figure 1(a)), even though the latter do not themselves imply that all candidates are unacceptable. Such heterogeneous signals can weaken the relationship between confidence and correctness, limiting the usefulness of threshold-based selection (Figure 1(b)). Moreover, low overall selective risk can conceal high error among accepted no-winner decisions.

To address this challenge, we propose a two-stage certified selection framework that first localizes plausible reference labels (Angelopoulos & Bates, 2023) and then performs decision-specific certification. In the first stage, conformal prediction constructs a localized set over the m response labels and the no-winner label, retaining the true reference label with marginal probability at least $1 - \alpha$. This yields a potentially smaller set of labels for the subsequent decision. In the second stage, the localized set determines which branch is tested, enabling decision-specific selection (Figure 1(c)). If it contains the no-winner label, a calibrated hypothesis test determines whether the no-winner acceptance rule meets its prescribed risk target; otherwise, a separate test evaluates the response-selection rule. Empty localized sets lead to abstention. When the corresponding test does not pass, the system abstains. The framework therefore returns a no-winner verdict, a selected response, or abstention, with high-probability control of the conditional error rate in each accepted decision branch under the stated assumptions. To summarize, the main contributions of this work are as follows:

- **No-winner-aware judgment.** We formulate LLM-as-a-judge evaluation in a generalized label space that explicitly allows for the possibility that no candidate response is human-preferred. This formulation better captures realistic open-ended evaluation scenarios in which all candidate responses fail a prespecified quality rubric.
- **Compositional guarantees.** We develop a two-stage framework introducing conformal localization with branch-specific fixed-sequence testing. The resulting framework provides two complementary guarantees: marginal retention of the true label, including the no-winner label, in the localized set, and high-probability control of the selective error rate within each accepted decision branch.
- **Extensive empirical evaluation.** We evaluate the framework on human-rating-based HelpSteer2 and TL;DR benchmarks and an Arena-Controlled synthetic benchmark, using diverse judges including Qwen, Llama, GPT-OSS, and DeepSeek. The results demonstrate that the two-stage policy can substantially improve no-winner recognition under matched branch-risk budgets. Evaluations on human-rated benchmarks further highlight the distinction between reliable winner selection and reliable no-winner recognition.

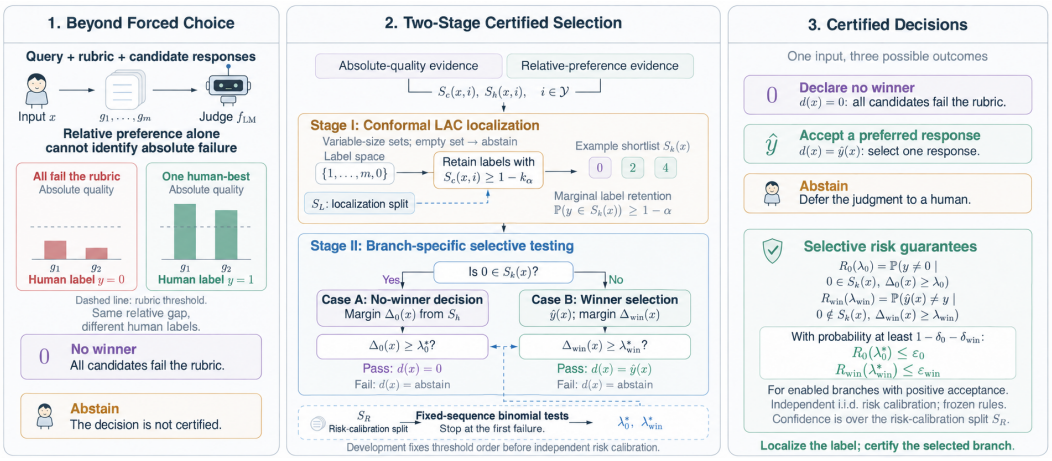


Figure 2: **Null-aware selective judging beyond forced choice. Left:** relative quality gap can occur when both candidates fail the rubric or when acceptable candidates admit a preferred response; a no-winner verdict is distinct from abstention. Middle: LAC constructs a variable-size label set using localization data. Nonempty sets containing N enter the no-winner branch; the remaining nonempty sets enter the winner branch. Empty sets and failed acceptance conditions yield abstention. Development data fix the threshold sequences before independent risk calibration, which stops at the first failed test. Right: the framework outputs N, a selected response, or abstention. Under the stated assumptions, the conditional risks of the enabled branches with positive population acceptance are jointly controlled with allowable confidence.

2 PRELIMINARY

2.1 PROBLEM SETUP

Let $f_{\text{LM}} : \mathcal{X} \to \mathcal{Y}$ denote an LLM judge. Let $x \in \mathcal{X}$ denote an input consisting of a query and a set of m candidate responses $\{g_1, \ldots, g_m\}$. Unlike standard multi-candidate preference evaluation (Jung et al., 2025), we do not assume that one candidate must be human-best. Instead, the human label is defined as $y \in \mathcal{Y} = \{1, \ldots, m, 0\}$, where $y = i > 0$ indicates that g_i is the human-best response, and $y = 0$ indicates that all candidate responses fail a prespecified quality criterion.

Let (x, y) denote a labeled instance drawn from an unknown distribution $\mathcal{D}$. We assume access to a calibration set $\mathcal{S} = \{(x_j, y_j)\}_{j=1}^n$ and a future test point (x_{n+1}, y_{n+1}). In this work, we assume an LLM judge provides scoring functions over the extended label space: $\mathbb{S}_c(x, i), \mathbb{S}_h(x, i)$. Here, $\mathbb{S}_c$ will be used for first-stage conformal localization, and $\mathbb{S}_h$ will used for second-stage selective decision making. Larger scores indicate stronger model evidence that label is the human-best label (or 0 for no-winner label). For ordinary response labels $i \in \{1, \ldots, m\}$, $\mathbb{S}_c(x, i)$ and $\mathbb{S}_h(x, i)$ can be derived from preference probabilities, simulated annotator votes, reward scores, or pairwise aggregation. In this work, we normally use the preference probabilities as the scores. For the no-winner label 0, the score can be obtained by explicitly prompting the judge with a "no best response" option or by constructing an implicit ambiguity score, for example using margin, entropy, absolute quality signals.

The objective is to output one of three types of decisions:

$$d(x) \in \{1, \ldots, m, 0, \text{abstain}\}.$$

A response-selection decision $d(x) = i > 0$ asserts that response g_i is the human-best response, accompanied by a calibrated probabilistic guarantee. A no-winner decision $d(x) = 0$ asserts that all candidates fail the prespecified quality criterion, with a calibrated selective-risk guarantee. Finally, an abstention decision indicates that the system lacks sufficient statistical evidence to certify either conclusion and should therefore defer the instance to human evaluation or escalate it to a stronger judge model.

2.2 ASSUMPTIONS

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

Assumption 1 (Sampling and independence). Examples are sampled i.i.d. from $\mathcal{D}$. Development, localization-calibration, risk-calibration, and test splits are mutually independent. In particular, localization examples and a future test example are exchangeable.

Assumption 2 (Well-defined human label). For each input x, the human-best label is well-defined in the extended label space $y \in \{1, \ldots, m, 0\}$. A nonzero reference label identifies a unique preferred response. Label 0 means that all candidates fail the prespecified quality criterion. Acceptable ties are outside the retained empirical label construction.

3 Two-Stage Certified Selection

We now present the proposed two-stage framework. Stage I employs conformal prediction to localize a small set of plausible labels that contains the true human label with high probability. Stage II then performs branch-specific selective testing according to whether the no-winner label 0 is included in the localized set. If 0 is contained in the localized set, the framework tests whether there is sufficient statistical evidence to certify a no-winner decision. If 0 is not contained in the localized set, the framework instead tests whether the selected response label can be certified as reliable. In both cases, the system abstains whenever the corresponding decision cannot be statistically guaranteed.We use disjoint development, localization-calibration $\mathcal{S}_L$, and risk-calibration $\mathcal{S}_R$ splits. The scorer is frozen before localization, and development data fix the threshold sequences before $\mathcal{S}_R$ is examined.

3.1 Stage I: Conformal Label Localization

For a labeled example $(x, y) \in \mathcal{S}_L$, we define the least ambiguous set-valued classifier (LAC) nonconformity score $1 - \mathbb{S}_c(x, y)$. This score measures the lack of support for the true label under the first-stage scoring function $\mathbb{S}_c$, with smaller values indicating better conformity. For a target localization error level $\alpha \in (0, 1)$, let k_α be the $\lceil (n_L+1)(1-\alpha) \rceil$-th smallest nonconformity score in $\mathcal{S}_L$, where $n_L = |\mathcal{S}_L|$.If the required order exceeds n_L, set $k_\alpha = +\infty$. For a new input x, construct the conformal shortlist

$$S_k(x) = \{y' \in \mathcal{Y} : 1 - \mathbb{S}_c(x, y') \leq k_\alpha\}. \tag{1}$$

Thus, $S_k(x)$ contains all labels with $\mathbb{S}_c(x, i) \geq 1 - k_\alpha$ over the extended label space $\{1, \ldots, m, 0\}$, and its size may vary across inputs.

Proposition 1 (Conformal localization guarantee). Under the assumptions in Section 2.2, the candidate set S_k constructed above satisfies

$$\mathbb{P}(y \in S_k(x)) \geq 1 - \alpha,$$

This proposition guarantees that the true label, which may be the no-winner label 0, is retained in the first-stage shortlist with marginal probability at least $1 - \alpha$.

3.2 Stage II: Branch-Specific Selective Testing

Stage I reduces the original judgment problem to a localized set $S_k(x)$. Stage II then considers two cases. When $0 \in S_k(x)$, the framework evaluates whether the evidence is strong enough to certify a no-winner decision. Conversely, when $0 \notin S_k(x)$, the framework tests whether the selected response label can be certified as the human-best response. In both branches, the system abstains whenever the corresponding conclusion cannot be supported with the prescribed statistical guarantee.

3.2.1 Case A: no-winner label within the localized set

When 0 is included in the conformal shortlist, the system cannot rule out the possibility that no human-best response exists. We therefore test whether the evidence for 0 is sufficiently strong to certify a no-winner decision. We define the no-winner margin

$$\Delta_0(x) = \mathbb{S}_h(x, 0) - \max_{i \in S_k(x) \setminus \{0\}} \mathbb{S}_h(x, i).$$

A large value of $\Delta_0(x)$ indicates that the no-winner label is preferred over all ordinary response labels retained in the shortlist. For a threshold λ_0, we define the selective false-no-winner risk as

$$R_0(\lambda_0) = \mathbb{P}(y \neq 0 \mid 0 \in S_k(x),\, \Delta_0(x) \geq \lambda_0).$$

This risk measures the probability of falsely declaring that no human-best response exists. On the independent risk-calibration split $\mathcal{S}_R$, we define the empirical false-no-winner risk as

$$\widehat{R}_0(\lambda_0) = \frac{E_0(\lambda_0) := \sum_{(x,y)\in\mathcal{S}_R} \mathbf{1}\{y \neq 0,\, 0 \in S_k(x),\, \Delta_0(x) \geq \lambda_0\}}{N_0(\lambda_0) := \sum_{(x,y)\in\mathcal{S}_R} \mathbf{1}\{0 \in S_k(x),\, \Delta_0(x) \geq \lambda_0\}}.$$

Then, we compute the exact binomial upper confidence bound

$$\widehat{R}_0^+(\lambda_0) = \sup\{r \in [0,1] : \mathbb{P}(\mathrm{Bin}(N_0(\lambda_0), r) \leq E_0(\lambda_0)) \geq \delta_0\}. \tag{2}$$

For a target risk level ε_0, let $\lambda_{0,1} > \cdots > \lambda_{0,K_0}$ be the finite candidate sequence fixed using development data before risk calibration. Test candidates in this order and stop at the first failure. Define

$$j_0^\star = \max\left(\{0\} \cup \left\{j \in \{1,\ldots,K_0\} : \widehat{R}_0^+(\lambda_{0,\ell}) \leq \varepsilon_0 \text{ for every } \ell \leq j\right\}\right).$$

If $j_0^\star > 0$, set $\lambda_0^\star = \lambda_{0,j_0^\star}$. Otherwise, disable the no-winner branch. The candidate construction is specified in Appendix D.

Proposition 2 (Certified no-winner decision). With probability at least $1-\delta_0$ over the calibration samples, given a new test example, the selected threshold $\lambda_0^\star$ satisfies

$$\mathbb{P}(y \neq 0 \mid 0 \in S_k(x),\, \Delta_0(x) \geq \lambda_0^\star) \leq \varepsilon_0.$$

3.2.2 Case B: no-winner label outside the localized set

When $0 \notin S_k(x)$, the conformal shortlist excludes the no-winner label, and the localized evidence therefore favors the existence of a human-best response among the retained candidates. In this case, the remaining task is to certify whether the judge can reliably select one response from $S_k(x)$.

We define the predicted response within the shortlist as

$$\hat{y}(x) = \arg\max_{i \in S_k(x)} \mathbb{S}_h(x,i),$$

and the corresponding response-selection margin as

$$\Delta_{\mathrm{win}}(x) = \mathbb{S}_h(x,\hat{y}(x)) - \max_{i \in S_k(x),\, i \neq \hat{y}(x)} \mathbb{S}_h(x,i).$$

A larger value of $\Delta_{\mathrm{win}}(x)$ indicates stronger evidence that the selected response dominates the other labels in the localized set. For a threshold λ_{win}, we define the selective winner error as

$$R_{\mathrm{win}}(\lambda_{\mathrm{win}}) = \mathbb{P}(\hat{y}(x) \neq y \mid 0 \notin S_k(x),\, \Delta_{\mathrm{win}}(x) \geq \lambda_{\mathrm{win}}).$$

This risk measures the probability of certifying an incorrect human-best response. Analogously to Case A, on the independent risk-calibration split $\mathcal{S}_R$, we estimate this risk by

$$\widehat{R}_{\mathrm{win}}(\lambda_{\mathrm{win}}) = \frac{E_{\mathrm{win}}(\lambda_{\mathrm{win}}) := \sum_{(x,y)\in\mathcal{S}_R} \mathbf{1}\{\hat{y}(x) \neq y,\, 0 \notin S_k(x),\, \Delta_{\mathrm{win}}(x) \geq \lambda_{\mathrm{win}}\}}{N_{\mathrm{win}}(\lambda_{\mathrm{win}}) := \sum_{(x,y)\in\mathcal{S}_R} \mathbf{1}\{0 \notin S_k(x),\, \Delta_{\mathrm{win}}(x) \geq \lambda_{\mathrm{win}}\}}.$$

We then compute the exact binomial upper confidence bound $\widehat{R}_{\mathrm{win}}^+(\lambda_{\mathrm{win}})$ with confidence $1-\delta_{\mathrm{win}}$. Given a target risk level $\varepsilon_{\mathrm{win}}$, we select the threshold by the same fixed-sequence testing rule and get $\lambda_{\mathrm{win}}^\star$. The certified response-selection rule is therefore

$$d(x) = \hat{y}(x) \quad \text{if} \quad 0 \notin S_k(x) \text{ and } \Delta_{\mathrm{win}}(x) \geq \lambda_{\mathrm{win}}^\star.$$

Otherwise, the system abstains or escalates.

Proposition 3 (Certified winner selection). With probability at least $1-\delta_{\mathrm{win}}$ over the calibration samples, given a new test example, the selected threshold $\lambda_{\mathrm{win}}^\star$ satisfies

$$\mathbb{P}(\hat{y}(x) \neq y \mid 0 \notin S_k(x),\, \Delta_{\mathrm{win}}(x) \geq \lambda_{\mathrm{win}}^\star) \leq \varepsilon_{\mathrm{win}}.$$

Table 1: Simulations at $\varepsilon = 0.15$. All entries are percentages (%).

	Total coverage		One-stage		Two-stage	
Scenario	One-stage	Two-stage	C_0	GSR	C_0	GSR
Clear quality	98.1	**98.3** ↑	23.8	100.0	**24.0** ↑	100.0
Ties and quality	63.5	**70.2** ↑	1.1	97.9	**8.3** ↑	98.1
Mixed evidence	58.7	**67.5** ↑	6.0	100.0	**15.2** ↑	100.0
Weak evidence	0.0	**0.6** ↑	0.0	-	0.0	99.6

3.3 Overall Decision Rule

The final certified decision rule is

$$d(x) = \begin{cases} 0, & \text{if } 0 \in S_k(x) \text{ and } \Delta_0(x) \geq \lambda_0^\star, \\ \hat{y}(x), & \text{if } 0 \notin S_k(x) \text{ and } \Delta_{\text{win}}(x) \geq \lambda_{\text{win}}^\star, \\ \text{abstain}, & \text{otherwise.} \end{cases}$$

This rule has an important structural interpretation. A singleton or confidently dominant 0 branch certifies that no human-best response exists. A confident ordinary-label branch certifies a response-level decision. Mixed or low-margin cases are not forced into artificial decisions and are instead abstained or escalated. In the experiments, we normally let $\varepsilon = \varepsilon_0 = \varepsilon_{\text{win}}$, $\delta = \delta_0 = \delta_{\text{win}}$.

4 Experiments

In the experiments, we investigate three questions: (1) whether localized scoring improves selection over a one-stage framework; (2) whether localization increases correctly accepted no-winner recall under matched branch-risk budgets; and (3) how these effects vary across judges and benchmark constructions. We address these questions through controlled simulations, empirical risk comparisons, and matched branchwise evaluations on human-annotated benchmarks.

4.1 Controlled simulations

We compare one-stage, which certifies pooled risk using a single hypothesis testing framework, and our Two-stage method. The primary policy uses rank-conformal localization and abstains if its routed branch fails certification. We report total coverage and no-winner coverage C_0. In simulation, joint population guarantee success rate (GSR) counts repetitions in which both nonempty branches meet the risk target. For two-stage method, we set $\varepsilon = 0.15$, $\delta = 0.05$, and $\alpha = 0.10$, while $\varepsilon = 0.15$, $\delta = 0.145$ for One-stage method. The simulation specification appears in Appendix H.

We provide four scenarios vary label ambiguity, score noise, and misleading confidence in no-winner predictions. In each of 1,000 repetitions, we draw 500 localization, 1,500 certification, and 2,000 test observations independently. The informative scenarios assume that the localizer can exclude misleading candidates; they do not model a particular human benchmark or LLM.

In ties and quality, Two-stage increases total coverage from One-stage's 63.5% to 70.2%, including 8.3% no-winner coverage, while in mixed evidence, Two-stage increases total coverage from One-stage's 58.7% to 67.5%, including 15.2% no-winner coverage. This supports the value of localization beyond access to two scores. Conversely, weak evidence yields no no-winner coverage despite high GSR, statistical validity alone does not ensure useful decisions.

4.2 Empirical results on human-annotated datasets

Experimental protocol. We examine whether calibrated selection achieves useful coverage while controlling accepted-decision error. Our risk comparison evaluates three methods using 500 paired calibration repartitions per judge and a fixed test set. We consider $\varepsilon \in \{0.10, 0.15, 0.20, 0.25, 0.30\}$ and global failure budgets $\delta \in \{0.05, 0.025, 0.01\}$. All methods share the calibration partitions and test examples within each judge. Conformal localization uses $\alpha = 0.10$, separately from the

Table 2: **Retrospective global-risk replay at $\varepsilon = 0.25$ and LAC $\alpha = 0.10$.** We report empirical guarantee success rate (GSR, %), the number of nonempty judge-runs (N), and mean acceptance rate (%). There are 500 paired repetitions per judge, corresponding to 3,000 runs for the six-judge aggregates and 500 runs for Gemini Flash. One-stage H uses maximum label probability. All methods use a single global threshold; the count-adaptive replay is retrospective and does not constitute a branchwise certification result.

		One-stage H			Localized Direct			Localized Quality-W		
Dataset	δ	**GSR**	**N**	**Accept.**	**GSR**	**N**	**Accept.**	**GSR**	**N**	**Accept.**
(a) Six open judges										
	0.05	33.33	33	0.118	**79.13**	**230**	**1.042**	63.16	19	0.051
HelpSteer2	0.025	30.77	26	0.096	**86.05**	**172**	**0.751**	85.71	7	0.021
	0.01	13.33	15	0.051	89.41	**85**	**0.424**	**100.00**	2	0.007
	0.05	**100.00**	2	0.010	**100.00**	**18**	**0.051**	**100.00**	16	0.047
TL;DR	0.025	**100.00**	1	0.008	**100.00**	**17**	**0.043**	**100.00**	9	0.027
	0.01	**100.00**	1	0.004	**100.00**	**8**	**0.020**	**100.00**	4	0.015
	0.05	88.53	1987	49.962	97.99	2585	83.737	**98.59**	**2834**	**87.373**
Arena-Controlled	0.025	93.93	1877	46.156	99.17	2537	83.267	**99.74**	**2714**	**86.162**
	0.01	97.47	1542	39.210	99.88	2504	82.699	**99.96**	**2621**	**85.000**
(b) Gemini Flash										
	0.05	96.43	56	1.308	13.68	**117**	**4.256**	**98.68**	76	1.714
HelpSteer2	0.025	**100.00**	13	0.338	9.20	**87**	**2.856**	**100.00**	27	0.648
	0.01	**100.00**	4	0.104	9.80	**51**	**1.994**	**100.00**	3	0.090
	0.05	0.00	403	48.482	0.00	139	20.663	**1.40**	**500**	**63.477**
TL;DR	0.025	0.00	319	37.937	0.00	65	9.620	**7.20**	**500**	**62.264**
	0.01	0.00	168	20.020	0.00	16	2.439	**24.60**	**500**	**60.008**

certification failure budget δ. Human-rating-based benchmarks, HelpSteer2 and TL;DR, involve subtle quality distinctions and potentially ambiguous failure cases. We therefore complement them with Arena-Controlled, where both-bad outcomes are constructed through explicit response failures. This controlled setting allows us to examine whether two-stage selection translates identifiable no-winner evidence into higher correctly accepted N recall than one-stage selection.

Judges and baselines. We evaluate Qwen2.5-1.5B-Instruct, Qwen2.5-7B-Instruct (Hui et al., 2024), DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025), Llama-3.1-8B-Instruct(Grattafiori et al., 2024), GPT-OSS-20B (Agarwal et al., 2025), and Gemma-4-E4B-it (Team et al., 2026). We additionally evaluate Gemini-2.5-Flash(Comanici et al., 2025) on HelpSteer2 (Wang et al., 2024) and TL;DR (Stiennon et al., 2020). One-stage predicts the highest-scoring label under $\mathbb{S}_h$ and uses its probability as the acceptance score, calibrating a single global threshold without localization.

Scoring variants. We compare two ways to score a proposed winner $\hat{y}(x)$. DIRECT uses the judge's preference confidence directly. QUALITY-W additionally assesses whether the selected response is acceptable on its own, since the better of two poor responses may still be poor, and uses the lower of the preference and acceptability scores. Quality-W keeps the same proposed answer and modifies only the winner acceptance score; no-winner scoring remains unchanged. We use DIRECT as the default in the global-risk comparison and QUALITY-W as an ablation.

Metrics. For the global-risk comparison, each cell reports empirical GSR (%), the number of nonempty runs, and mean coverage (%). A run is nonempty if it accepts at least one test decision. Empirical GSR is the percentage of nonempty runs whose pooled accepted test error is at most ε, including both winner and no-winner decisions. Mean coverage averages the accepted fraction of test examples over all runs, including empty runs. A dash denotes undefined GSR when no run is nonempty. Empirical GSR measures finite-test performance and is distinct from the population-level confidence guarantee.

4.2.1 GLOBAL-RISK COMPARISON WITH ONE-STAGE SELECTION

A selective evaluator must do more than achieve low error on a small accepted subset: it must also provide useful decisions. High empirical GSR alone can mask a policy that rarely accepts

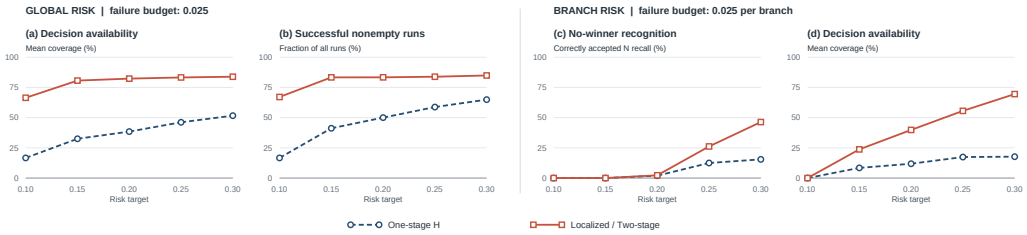


Figure 3: **Risk-target sensitivity on Arena-Controlled.** Results use LAC $\alpha = 0.10$, averaged over six judges and 500 paired repartitions per judge (3,000 runs). **Global risk (a–b):** with failure budget $\delta = 0.025$, we report (a) mean coverage and (b) the fraction of all runs that are nonempty and satisfy the pooled test-risk target. **Branch risk (c–d):** with $\delta_0 = \delta_{\text{win}} = 0.025$, we report (c) correctly accepted no-winner recall and (d) mean coverage under matched branchwise risk control. One-stage H selects using the maximum S_b probability. Blue dashed curves with circles denote One-stage H; red solid curves with squares denote Localized Direct in (a–b) and the Two-stage method in (c–d).

any judgments. We therefore compare one-stage and localized selection under matched global-risk budgets, asking whether localized routing and scoring improve the tradeoff between empirical reliability and decision availability.

All methods use the same calibration partitions and test examples and calibrate a single global threshold. One-stage H uses maximum label probability without localization; Direct and Quality-W use localized proposals and their corresponding acceptance scores. Table 2 reports GSR together with nonempty runs and mean acceptance, so that reliability gains can be distinguished from simply withholding more decisions.

For the open judges on HelpSteer2, Direct improves both GSR and coverage over H at every listed budget. At $\delta = 0.025$, GSR rises from 30.77% to 86.05%, while coverage increases approximately eightfold, from 0.096% to 0.751%. On Arena-Controlled, Direct raises coverage from 46.16% to 83.27% and GSR from 93.93% to 99.17%. Quality-W further increases these to 86.16% coverage and 99.74% GSR. With Gemini on HelpSteer2, Quality-W nearly doubles coverage from 0.338% to 0.648% at $\delta = 0.025$, while matching H's 100% empirical GSR. These results show that localized selection can improve both empirical reliability and decision availability, with the preferred scoring variant depending on the judge and task.

Sensitivity to the risk target. Figure 3(a)(b) extends the global comparison across five risk targets on Arena-Controlled. At $\varepsilon = 0.10$, Direct achieves 66.43% coverage, compared with 16.67% for H; the fractions of all runs that are nonempty and meet the test-risk target are 67.10% and 16.67%, respectively. At $\varepsilon = 0.20$, coverage reaches 82.30% versus 38.45%, with successful nonempty runs accounting for 83.37% versus 50.03%. The advantage therefore persists when empty runs are included, rather than arising solely from conditioning GSR on rare acceptance.

4.2.2 No-Winner Recognition under Matched Branch-Risk Budgets

We next compare the full two-stage policy with one-stage selection, holding the branch-risk budget and threshold-design policy fixed across both methods. One-stage H calibrates its full-space rule per branch; Two-stage adds LAC localization and no-winner-first routing. The comparison measures the combined effect of shortlist pruning, no-winner-first routing, and localized scoring against an equally calibrated, unlocalized baseline. As shown in Table 3, at $\varepsilon = 0.30$, Two-stage increases accepted N recall from 15.47% to 46.36% on Arena-Controlled, approximately a threefold improvement, with 100% observed N precision for both methods. Total coverage rises from 17.70% to 69.50%, and joint empirical GSR increases from 96.48% to 100%. At $\varepsilon = 0.25$, N recall increases from 12.47% to 26.16%. Neither method produces N outputs at .15. Figure 3(c)(d) separates this change in N recognition from the increase in total decision availability. Appendix F reports the complete Gemini N recall and precision results on both human-rated benchmarks.

Table 3: **Matched branch-risk results on Arena-Controlled (%).** Results use six judges with 500 paired repartitions each and $\delta_0 = \delta_{\text{win}} = 0.025$. N-recall and coverage are averaged over all runs, whereas N-precision pools accepted N decisions. Joint GSR requires both nonempty branches to satisfy their test-risk targets.

ε	Method	N-Recall ↑	N-Prec. ↑	Coverage ↑	Joint GSR ↑
0.15	One-stage H	**0.00**	–	8.41	97.99
	Two-stage	**0.00**	–	**23.74**	**100.00**
0.20	One-stage H	1.97	100.00	11.86	98.43
	Two-stage	**2.28**	100.00	**39.82**	**100.00**
0.25	One-stage H	12.47	100.00	17.39	97.02
	Two-stage	**26.16**	100.00	**55.61**	**100.00**
0.30	One-stage H	15.47	100.00	17.70	96.48
	Two-stage	**46.36**	100.00	**69.50**	**100.00**

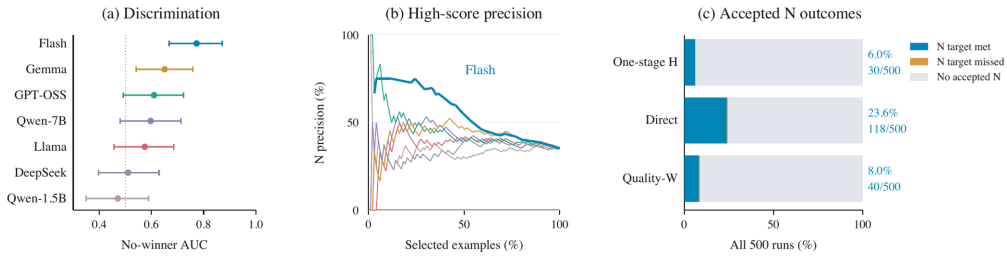


Figure 4: No-winner diagnostics on HelpSteer2. (a–b) Existing shared-example score diagnostics, unchanged by the acceptance-baseline revision (99 examples). Panel (a) retains the archived bootstrap intervals. (c) Recomputed Gemini-2.5-Flash N outcomes on its full 100-example test set, under global calibration at $\varepsilon = 0.30$, $\delta = 0.01$, over 500 repartitions. Numbers above bars count runs with nonempty N output whose empirical N error is at most 0.30.

4.2.3 From No-Winner Scores to Accepted Decisions

Figure 4(a–b) asks whether no-winner recognition tracks a judge's general judging ability, comparing the frontier judge Gemini-2.5-Flash against six open-weight judges on 99 shared HelpSteer2 examples. Flash reaches an AUC of 0.773, versus 0.471–0.650 for the open judges; its top 22 scores are 72.7% precise, and no tie-preserving threshold exceeds 75%. The stronger judge does discriminate no-winner cases better, but even its best-scoring subset stays impure – stronger overall judging ability does not guarantee a high-confidence no-winner subset. Panel (c) then asks whether better ranking actually yields more accepted N decisions. Under the global protocol at $\varepsilon = 0.30$, $\delta = 0.01$, Direct meets the N risk target in 118 of 500 runs, compared with 30 for H and 40 for Quality-W (3, 1, and 3 runs respectively exceed it). Direct thus accepts more target-meeting N decisions at this operating point – though as branch-level diagnostics of global selection, these counts are not independently calibrated N certificates.

5 Conclusion

In this work, we studied certified LLM-as-a-judge evaluation in a label space that explicitly allows a no-winner outcome. Because confidence in a no-winner verdict can reflect qualitatively different signals, a single pooled risk threshold may mask substantial error within either decision branch. To address this, we introduced a two-stage framework that first localizes the plausible label set using a distribution-free conformal guarantee and then performs separate certification for no-winner and response-selection decisions, abstaining when the corresponding criterion is not met. On the synthetic positive-control dataset, this approach achieves substantial certified gains in no-winner recall under matched branch-risk budgets.

AI USE STATEMENT

We used generative AI tools to help debug experiment code and to improve the manuscript's language and typesetting; these tools did not generate any data or experimental results. All AI-assisted code was reviewed and tested by the authors, who take full responsibility for the final content of this work, including text, claims, and artifacts produced with such assistance.

ETHICS STATEMENT

Our human study used only public datasets, involved no personal or sensitive information. All datasets used in this work are publicly available.

REPRODUCIBILITY STATEMENT

All datasets used in this work are publicly available, and our code will be released upon publication.

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.

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

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 2023.

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

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

Tian Bai and Ying Jin. Conformal selective prediction with general risk control. *arXiv preprint arXiv:2603.24704*, 2026.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Xun Deng, Han Zhong, Rui Ai, Fuli Feng, Zheng Wang, and Xiangnan He. Less is more: Improving llm alignment via preference data selection. *Advances in Neural Information Processing Systems*, 38:161259–161285, 2026.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

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

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

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704, 2023.

Gaojie Jin, Yong Tao, Lijia Yu, and Tianjin Huang. Margin-adaptive confidence ranking for reliable llm judgement. *arXiv preprint arXiv:2605.15416*, 2026.

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

Simran Kaur, Simon Park, Anirudh Goyal, and Sanjeev Arora. Instruct-skillmix: A powerful pipeline for llm instruction tuning. In *International Conference on Learning Representations*, volume 2025, pp. 28842–28983, 2025.

Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Aaron Jiaxun Li, Soheil Feizi, and Himabindu Lakkaraju. Certifying llm safety against adversarial prompting. *arXiv preprint arXiv:2309.02705*, 2023.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024.

Xiaomin Li, Zhou Yu, Zhiwei Zhang, Xupeng Chen, Ziji Zhang, Yingying Zhuang, Narayanan Sadagopan, and Anurag Beniwal. When thinking fails: The pitfalls of reasoning for instruction-following in llms. *Advances in Neural Information Processing Systems*, 38:77925–77962, 2026a.

Xinyu Li, Yi Zhou, Guanqun Cao, Zeyu Fu, Tianjin Huang, and Gaojie Jin. Localize-then-decide guarantees for llm judgments. *arXiv preprint arXiv:2608.25824*, 2026b.

Jinsong Liu, Dongdong Ge, and Ruihao Zhu. Reward learning from preference with ties. *arXiv preprint arXiv:2410.05328*, 2024.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah Smith, Hanna Hajishirzi, and Nathan Lambert. Rewardbench 2: Advancing reward model evaluation. In *International Conference on Learning Representations*, volume 2026, pp. 144839–144866, 2026.

Yutao Mou, Shikun Zhang, and Wei Ye. Sg-bench: Evaluating llm safety generalization across diverse tasks and prompt types. *Advances in Neural Information Processing Systems*, 37:123032–123054, 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.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33:3581–3591, 2020.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of machine learning research*, 9(3), 2008.

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 Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. Judgebench: A benchmark for evaluating llm-based judges. In *International Conference on Learning Representations*, volume 2025, pp. 63277–63303, 2025.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. Moral alignment for llm agents. In *International Conference on Learning Representations*, volume 2025, pp. 97477–97510, 2025.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J Zhang, Makesh N Sreedhar, and Oleksii Kuchaiev. Helpsteer 2: Open-source dataset for training top-performing reward models. *Advances in Neural Information Processing Systems*, 37:1474–1501, 2024.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. Helpsteer2-preference: Complementing ratings with preferences. In *International Conference on Learning Representations*, volume 2025, pp. 56548–56575, 2025.

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.

Christopher Yeh, Nicolas Christianson, Adam Wierman, and Yisong Yue. Conformal risk training: End-to-end optimization of conformal risk control. *Advances in Neural Information Processing Systems*, 38:70056–70092, 2026.

Wenpeng Yin, Qinyuan Ye, Pengfei Liu, Xiang Ren, and Hinrich Schütze. Llm-driven instruction following: Progresses and concerns. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pp. 19–25, 2023.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large language models at evaluating instruction following. In *International Conference on Learning Representations*, volume 2024, pp. 40193–40219, 2024.

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 Related work

LLM judges and selective evaluation. LLM judges offer scalable approximations to human evaluation (Li et al., 2024; Gu et al., 2026), but are susceptible to position, verbosity, and other biases (Zheng et al., 2023; Tan et al., 2025; Zeng et al., 2024; Jin et al., 2026). Length-controlled AlpacaEval addresses verbosity as one such confounder (Dubois et al., 2024). Selective prediction trades coverage for lower error on accepted judgments (Geifman & El-Yaniv, 2017), and recent methods apply calibrated selection to pairwise LLM judging (Jung et al., 2025; Badshah et al., 2026). In particular, Jung et al. (2025) provide high-probability human-agreement guarantees using simulated annotators and escalation across judges. Their experiments include ties, while the Arena fine-tuning setup of Zheng et al. (2023) combines tie and both-bad votes into one class. Other preference evaluations discard ties (Ouyang et al., 2022) or assess them explicitly (Liu et al., 2024; Malik et al., 2026; Li et al., 2026b). However, a tie between acceptable responses does not imply that every candidate is unacceptable. The paired ratings and preference annotations of Wang et al. (2025) further motivate distinguishing absolute response quality from relative preference. Our work explicitly models a no-winner outcome and separately certifies the error of accepted no-winner and response-selection decisions.

Conformal prediction and risk control. Conformal prediction provides marginal label-set coverage under exchangeability (Shafer & Vovk, 2008; Angelopoulos & Bates, 2023). Adaptive classification sets seek smaller prediction sets while preserving this marginal guarantee (Romano et al., 2020). Conformal risk control extends calibration to expected losses (Angelopoulos et al., 2024; Yeh et al., 2026), while Learn then Test selects risk-controlling configurations through hypothesis testing (Angelopoulos et al., 2025). Conformal abstention has also been used to withhold uncertain language-model outputs (Yadkori et al., 2024). More recent selective conformal methods control risk among accepted predictions (Bai & Jin, 2026). Our setting additionally distinguishes accepted no-winner verdicts from accepted response selections and calibrates their risks separately. Building on these tools, we first localize plausible labels, including no-winner, and then test the two decision branches separately. A no-winner verdict asserts that the candidates fail the evaluation criterion; abstention withholds a judgment that cannot be certified.

B Proofs and interpretation

Marginal localization. Condition on the data used to freeze the scorer. The localization true-label nonconformity values and the new true-label value are exchangeable. Randomized tie-breaking gives a uniform rank among $n_L + 1$ positions; retaining all boundary ties can only increase inclusion. The ceiling-corrected order statistic in Equation (1), with an infinite cutoff when needed, therefore retains the new reference label with probability at least $1 - \alpha$. The probability averages over the localization sample and the new observation.

Fixed branch tests. Condition on the frozen scorer, localization fit, and development-designed candidates. For a fixed branch and threshold, acceptance is a fixed function of the input. Under i.i.d. risk calibration, the accepted error count, conditional on the number accepted, is binomial with the branch's population conditional error probability. Inverting its lower-tail test yields Equation (2). An empty selection has upper bound one and fails any target below one.

Consider the first candidate in the fixed order whose population conditional risk exceeds its target. Selecting any invalid candidate requires passing this first true null; otherwise testing stops before any later candidate can be selected. Its rejection probability is at most δ_b. The tests need not be independent and population risk need not be monotone in the threshold. If there is no invalid candidate there is no failure event. A union bound gives simultaneous control at $1 - \delta_0 - \delta_{\text{win}}$. Branches with zero population acceptance contribute no conditional-risk violation.

Pooled and branch risk. If w is the proportion of accepted population decisions that are N, then $R_{\text{pool}} = wR_0 + (1 - w)R_{\text{win}}$. For instance, $w = .05$, $R_0 = .80$, and $R_{\text{win}} = .02$ yield pooled risk .059. This arithmetic example explains why pooled calibration does not substitute for N-risk control. It is not a measured benchmark result. Localization misses remain errors in branch testing, so marginal inclusion does not need to be added as another failure event in this risk statement.

Sample support. For zero observed errors, $U_\delta(0, n) = 1 - \delta^{1/n}$; hence a necessary zero-error support level is $n \geq \lceil \log \delta / \log(1 - \varepsilon) \rceil$. This is a property of a fixed test, not a license to skip unsupported candidates using risk-calibration counts. The matched experiment tests its development-frozen sequence from the first candidate, including empty or unsupported failures.

C Datasets and score sources

HelpSteer2-2low defines BAD as at least two of helpfulness, correctness, and coherence being at most 1, and GOOD as all three being at least 3. The older any-low-axis strict construction is not pooled into these results. TL;DR applies thresholds to human overall ratings for each summary. A summary is BAD if its median rating is at most 2 and none of its ratings is at least 6; it is GOOD if its median rating is at least 6 and none is at most 2. Other summaries are excluded. Intermediate summaries are excluded. N requires two BAD summaries of the same source post, while a winner pair requires GOOD and BAD. The eligible post pools need not coincide. One row is retained per post in the replay.

Table 4: Retained split sizes: localization (L), risk calibration (G), development (D), and test. TL;DR Flash retains one blocked item as abstention in its test denominator.

Dataset / judges	L	G	D	Test
HelpSteer2 / open	72	144	72	294
TL;DR-v14 / open	102	204	51	153
Arena-Controlled / open	40	40	40	80
HelpSteer2 / Flash	50	150	0	100
TL;DR-v14 / Flash	102	202	51	153

Acceptable ties and annotation provenance. N denotes both responses failing the quality criterion, not a tie between acceptable responses. GOOD–GOOD pairs, including acceptable ties, and pairs outside the specified BAD–BAD or GOOD–BAD constructions are excluded. The empirical labels are constructed from the datasets' existing human ratings using these rules; they are not labels collected in a new annotation study. Our AI annotation dry run is a separate assessability diagnostic, not evidence of human inter-rater agreement.

Arena-Controlled uses synthetic response corruptions to produce clear quality contrasts. Its one-pair-per-group replay differs from the earlier 600-pair ablation, whose 120 risk-calibration pairs shared 40 source groups. That dependent-pair ablation is not used as branch-certification evidence here. HH and Alpaca silver-label experiments are outside this recomputed comparison and do not supply independent human both-bad labels.

TL;DR Flash includes 152 scored test rows and one blocked item retained as abstention in the denominator of 153. Two unavailable risk-calibration rows are excluded before repartitioning. The datasets contain distinct sample and source-group IDs within each replay cell. Between-judge aggregates summarize results, not a deployed ensemble.

Heads and readouts. Five heads provide standalone acceptability for A and B, localization preference, direct BOTH_BAD/NOT_BOTH_BAD evidence, and direct winner preference. Human labels and ratings are not inserted into the input. Standalone heads average two opposite output-code mappings; pairwise heads additionally swap candidate order, giving 16 configurations per pair. All outputs are mapped to the original semantic coordinates before averaging.

Qwen, DeepSeek, and Llama normalize probabilities over audited legal code tokens. Gemma additionally handles audited byte-fallback events. GPT-OSS normalizes mean token log probabilities for label-dependent completion suffixes. HelpSteer2 Flash uses structured verbalized probabilities, whereas TL;DR Flash uses mapped categorical votes. These readout differences are part of the evaluated configurations; model identity alone does not explain every contrast.

Let u_A, u_B denote acceptability scores, (v_A, v_B) localization preferences, z the BOTH_BAD score, and (w_A, w_B) direct winner scores. The normalized compositions are

$$\begin{aligned}
\mathbb{S}_c(x, 0) &= (1 - u_A)(1 - u_B), \\
\mathbb{S}_c(x, A) &= u_A(1 - u_B) + u_A u_B v_A, \\
\mathbb{S}_c(x, B) &= (1 - u_A)u_B + u_A u_B v_B, \\
\mathbb{S}_h(x, 0) &= z, \quad \mathbb{S}_h(x, A) = (1 - z)w_A, \quad \mathbb{S}_h(x, B) = (1 - z)w_B.
\end{aligned}$$

These are score constructions, not an assumption that the elicited heads are statistically independent. Winner prompts that posit a GOOD–BAD pair do not establish that the input actually satisfies that premise.

D Development-designed branch calibration

Data separation. Let $\mathcal{S}_D$, $\mathcal{S}_L$, and $\mathcal{S}_R$ denote the development, localization-calibration, and risk-calibration splits, respectively. The scorer is frozen before fitting the localization rule on $\mathcal{S}_L$. Conditional on this rule, $\mathcal{S}_D$ determines the ordered threshold candidates for each branch. These sequences are fixed before $\mathcal{S}_R$ is examined. Under the sampling assumptions, $\mathcal{S}_R$ is independent of all components used to construct the acceptance rules. The test set is not used to fit scores, localization

cutoffs, or threshold sequences. Both matched methods use the following identical design policy within each routed branch. Retain nonnegative development acceptance scores and construct unique descending candidates from 21 empirical quantiles with levels $.8, .76, \ldots, 0$, together with zero. Among candidates with development empirical error at most ε, choose the starting candidate with largest accepted development count, breaking ties toward the higher threshold. If none meets that empirical target, start at the candidate with lowest development error, then largest accepted count, then highest threshold. Retain this candidate and all lower candidates in the list. If there are no nonnegative development scores, the branch has no candidates.

No minimum development sample size or support multiplier is imposed. Development error determines a proposed sequence, not a certificate. Risk calibration independently tests every candidate from this frozen start at level $.025$, stops at the first failure, and returns the last passing threshold if one exists. No candidate is skipped because of risk-calibration count or error. Both branches use the same $\varepsilon \in \{.10, .15, .20, .25, .30\}$. The One-stage H–branch baseline predicts the full-space $\mathbb{S}_h$ argmax and uses its maximum probability as the acceptance score, with deterministic lowest-label-index tie-breaking, and uses these same separate branch tests. Two-stage instead uses LAC localization, N-first routing, and the localized margin, including boundary equality at zero, as specified in the method. The matched branch-risk comparison measures their combined effect under matched budgets; it does not separately identify the contributions of shortlist pruning, routing, and the acceptance statistic.

The policy replaces an earlier heuristic requiring development-estimated support to exceed 1.5 times the zero-error minimum. That heuristic produced no N candidates on Arena and is not used in the reported matched comparison. The current policy was developed during the retrospective analysis; fresh independent calibration and evaluation would be needed for a prospective deployment study.

E AUXILIARY GLOBAL REPLAY

The global experiment preserves the archived protocol. LAC at $\alpha = .10$ retains boundary ties with the saved numerical tolerance and expands an empty set to the highest localization-score label. A singleton N set proposes N. A mixed set proposes N when $z \geq .5$ and otherwise proposes the highest winner-head score among retained response labels. Sets without N propose a retained winner. Direct uses z for N and $w_{\hat{y}}$ for winners; Quality-W uses $\min(w_{\hat{y}}, u_{\hat{y}})$ for winners and the same N score. Binary ties are excluded under the archived rule. H–global instead uses the full decision-vector argmax and its maximum probability as the acceptance score, with deterministic lowest-label-index tie-breaking.

Each method fits one threshold at the listed global δ. Its grid is the union of 5,000 linear points on $[0, 1]$ and 5,000 geometric points from 10^{-8} to one, with a duplicate endpoint removed. The archived fitter skips initial thresholds below the bound-implied minimum accepted count and then stops at its first failed CP test. This count-dependent starting point is not the independent sequence used in the matched branchwise experiment; we report the global table as an empirical selection comparison rather than applying the branchwise theorem to it.

F GEMINI NO-WINNER RECALL ON HUMAN-RATED BENCHMARKS

We distinguish raw N recognition from correctly accepted N recall after selective calibration. Raw predictions use the maximum S_h label. Accepted N recall is the number of correct accepted N decisions divided by all true N examples in the fixed test set, averaged over all 500 repartitions, including runs without N output. N precision pools correct and total accepted N counts across those runs. Repeated runs reuse test examples and are not independent evaluation samples.

Recognition before calibration. Table 6 reports recognition before calibration. Gemini-2.5-Flash correctly predicts 27 of 35 N examples on HelpSteer2 (77.14% recall) and 15 of 51 on TL;DR (29.41%). The corresponding N precision is 47.37% and 53.57%. Thus, nonzero raw recall does not imply reliable accepted N decisions. HelpSteer2 has 100 scored test examples. TL;DR has 153 intended test examples, of which 152 have scores; the blocked example is not N and is treated as abstention rather than removed from the evaluation.

N recall after global-risk calibration. At $\varepsilon = .30$ and $\delta = .025$ on HelpSteer2, Direct increases accepted N recall from 3.68% for H to 7.03%, with N precision of 72.97% versus 74.19%. At $\delta = .01$, the recall improvement is 1.66% to 6.10%, with nearly identical N precision (74.55% versus 74.41%). These gains describe N outputs from the global-risk protocol, not separate N-risk certificates. On TL;DR, Direct does not improve N recall over H at the listed budgets, and accepted N precision remains below 50% for all three methods.

Separate branch-risk evaluation. On TL;DR, both the maximum-probability H baseline and the N-first two-stage method have zero accepted N recall at $\varepsilon \in \{.10, .15, .20, .25, .30\}$ under $\delta_0 = \delta_{\text{win}} = .025$ and the frozen development-based threshold-design policy. HelpSteer2 Flash is not evaluated under that policy because its frozen input lacks a separate development split; it must not be reported as a failed branchwise certificate.

G METRICS, PARTITIONS, AND AGGREGATION

Stable IDs order the rows. The original localization and risk-calibration pool is permuted using seed 20260916, a dataset-profile hash, and the repetition index, then split at the original sizes. Development and test sets remain fixed. Methods share all 500 permutations within a judge. Repetitions reuse the same observations and are not independent samples for a binomial significance test.

For judge j and run b, let M_{jb} be accepted test decisions, E_{jb} their errors, $M_{jb,0}$ accepted N decisions, $E_{jb,0}$ erroneous accepted N decisions, and $n_{j,0}$ the number of true N test labels. Reported N recall is

$$\overline{\text{Recall}}_0 = \frac{100}{JB} \sum_{j,b} \frac{M_{jb,0} - E_{jb,0}}{n_{j,0}},$$

while N precision pools correctly accepted N counts over accepted N counts across judge-runs. It is undefined with no N output. Coverage averages $100M_{jb}/n_j$ over all judge-runs, including empty runs. Global GSR pools successful and nonempty run counts; joint GSR requires the risk target in each branch with positive test acceptance. Different methods can have different nonempty-run conditioning sets. 100% observed precision or success does not imply zero population risk. These definitions apply equally to favorable and empty outcomes.

H ORIGINAL SIMULATION SPECIFICATION

The original simulation uses 4,096 equiprobable input states, sampled independently with replacement, and a fresh conditional label for each observation. It has rank-conformal localization rather than LAC and the unequal nominal budgets reported in the main text. It is retained as a separate illustration; it does not determine the matched empirical operating points.

I REPRODUCIBILITY AND LIMITATIONS

All empirical tables and the new figure are generated from stored score files without new judge calls. Input hashes, per-run arrays, per-model summaries, and calibration settings accompany the source bundle. The global H and Direct acceptance/error arrays match the earlier independent replay implementation in all 20 dataset–judge cells. The matched calibration checks include stopping at an empty first candidate, stopping at a failed first candidate without jumping ahead, and passing a supported zero-error sequence.

The comparisons use constructed rating-derived subsets, and Arena is a synthetic control. Between-model results also reflect readouts and prompt configurations. Analyses and operating policies were revised on the available data; we use the paired replays to characterize empirical behavior and do not treat their repetitions as fresh human judgments. The main remaining practical challenge is improving high-precision N evidence on human-rated pairs while retaining sufficient calibration support. A future evaluation should freeze the rubric, judge, development policy, and budgets before collecting independent risk-calibration and test groups.

Table 5: Complete global-risk comparison at $\varepsilon = .30$. Cells: GSR (%) / nonempty runs / mean coverage (%). Open aggregates have 3,000 runs; Flash has 500. One-stage H uses maximum label probability. All methods calibrate one global acceptance threshold.

Dataset / judges	δ	One-stage H	Direct	Quality-W
HelpSteer2 / open	0.05	13.91 / 115 / 0.304	88.02 / 459 / 2.170	37.78 / 45 / 0.120
HelpSteer2 / open	0.025	33.33 / 33 / 0.129	89.13 / 230 / 1.280	63.16 / 19 / 0.053
HelpSteer2 / open	0.01	30.77 / 26 / 0.100	97.09 / 172 / 0.817	85.71 / 7 / 0.021
TL;DR / open	0.05	100.00 / 6 / 0.020	89.66 / 29 / 0.119	100.00 / 24 / 0.090
TL;DR / open	0.025	100.00 / 2 / 0.014	100.00 / 18 / 0.079	100.00 / 16 / 0.055
TL;DR / open	0.01	100.00 / 1 / 0.008	100.00 / 17 / 0.043	100.00 / 9 / 0.027
Arena-Controlled / open	0.05	95.33 / 2036 / 54.252	96.88 / 2689 / 84.580	97.35 / 2942 / 90.246
Arena-Controlled / open	0.025	97.99 / 1987 / 51.643	98.53 / 2585 / 83.869	99.19 / 2834 / 87.795
Arena-Controlled / open	0.01	98.45 / 1877 / 47.438	99.41 / 2537 / 83.447	99.74 / 2714 / 86.490
HelpSteer2 / Flash	0.05	99.01 / 101 / 3.020	93.22 / 118 / 6.792	87.39 / 111 / 5.526
HelpSteer2 / Flash	0.025	100.00 / 72 / 1.936	95.90 / 122 / 5.742	94.94 / 79 / 2.676
HelpSteer2 / Flash	0.01	100.00 / 31 / 0.862	99.17 / 121 / 4.688	95.35 / 43 / 1.404
TL;DR / Flash	0.05	24.60 / 500 / 78.706	21.06 / 489 / 80.035	22.80 / 500 / 79.396
TL;DR / Flash	0.025	42.60 / 500 / 73.797	34.37 / 451 / 72.404	40.80 / 500 / 75.514
TL;DR / Flash	0.01	65.10 / 490 / 66.482	48.88 / 356 / 55.937	63.60 / 500 / 70.596

Table 6: Gemini-2.5-Flash before selective calibration. Predictions use $\arg\max S_h$. Recall includes all true N test items; blocked items are abstentions. Rates are percentages.

Dataset	Test n	True N	Correct N	N recall	N precision
TL;DR-v14	153	51	15	29.41	53.57
HelpSteer2	100	35	27	77.14	47.37

Table 7: Gemini N outputs after global-risk calibration at $\varepsilon = .30$. N recall averages over all 500 runs, including empty runs; N precision pools accepted N counts. n_+ counts runs with any accepted N. These are not separate N-risk certificates.

Dataset	δ	Method	N recall	N precision	n_+
HelpSteer2	0.05	One-stage H	5.44	73.01	101
HelpSteer2	0.05	Direct	7.73	71.29	118
HelpSteer2	0.05	Quality-W	6.98	68.42	111
HelpSteer2	0.025	One-stage H	3.68	74.19	72
HelpSteer2	0.025	Direct	7.03	72.97	122
HelpSteer2	0.025	Quality-W	3.85	71.17	79
HelpSteer2	0.01	One-stage H	1.66	74.55	31
HelpSteer2	0.01	Direct	6.10	74.41	121
HelpSteer2	0.01	Quality-W	2.12	72.18	43
TL;DR-v14	0.05	One-stage H	12.04	45.86	500
TL;DR-v14	0.05	Direct	11.40	45.38	489
TL;DR-v14	0.05	Quality-W	13.73	46.67	500
TL;DR-v14	0.025	One-stage H	9.93	44.10	500
TL;DR-v14	0.025	Direct	9.34	44.36	451
TL;DR-v14	0.025	Quality-W	13.73	46.67	500
TL;DR-v14	0.01	One-stage H	7.32	41.21	490
TL;DR-v14	0.01	Direct	6.36	42.98	356
TL;DR-v14	0.01	Quality-W	13.73	46.67	500