

Multi-Expert Conformal Risk Control for Pairwise LLM Judging in Open-Ended Dialogue

Ming Cheng Yusheng Dai Qihong Ke[†] Zhaolin Chen Lizhen Qu[†]

Monash University

{ming.cheng, yusheng.dai, qihong.ke, zhaolin.chen, lizhen.qu}@monash.edu

Abstract

In this paper, we explore multi-expert Conformal Risk Control (CRC) algorithms for pairwise LLM-as-a-Judge evaluation in open-ended dialogue. Our core insight is that multi-expert aggregation offers a complementary remedy to CRC: whereas CRC controls risk at the decision threshold through abstention, aggregation sanitizes the scoring function at its source. Guided by this, we first design two multi-expert CRC methods: *Score Averaging* and *Decision Voting*, which aggregate at the score and decision levels, respectively. While both strategies outperform single-expert methods on homogeneous expert panels, on heterogeneous LLM judges they remain risk-valid but recover only limited coverage, because a uniform threshold cannot match the experts’ distinct scoring scales. To resolve this issue, we further propose *Marginal-Calibrated Conformal Consensus* (MC³): it captures distinct per-expert scales via initial threshold ratios, while jointly tuning a unified decision function $C_t(x)$ applied identically in both calibration and test, thereby preserving exchangeability. To evaluate our framework, we construct *Panel*, a 1,800-pair human pairwise-preference benchmark for open-ended dialogue. It is built on responses generated by four open-weight LLMs over dialogue contexts from three domains (ESConv, MSC, DREAM), with full logit access. In experiments, we find that both Score Averaging and Decision Voting substantially improve accuracy and acceptance rate on homogeneous panels. Notably, MC³ extends these gains to heterogeneous panels by accommodating distinct per-expert scoring scales across all three datasets.

1 Introduction

LLM-as-a-judge has become the standard for evaluating open-ended dialogue, replacing costly human preference labels with model-based pairwise comparisons (Zheng et al., 2023; Kim et al., 2024).

Yet, in high-stakes domains such as mental-health support, relying on automated judges carries significant risk: even minor misjudgments, driven by overconfidence or systematic biases, can have severe consequences (Wang et al., 2024; Koo et al., 2024). Consequently, the safe and reliable use of these LLM evaluators in such settings demands explicit mechanisms for strict risk control.

To operationalize this risk control, we formulate LLM-based pairwise evaluation as a *selective prediction* problem (Jung et al., 2025; Chen et al., 2023) under the Conformal Risk Control (CRC) framework (Angelopoulos et al., 2024) (detailed in Section 2). Instead of forcing a verdict, the LLM judge commits to a preference only when sufficiently confident; otherwise, it abstains and returns the full candidate set (optionally deferring the pair to human review). By calibrating the abstention threshold λ on the calibration set, CRC provides a formal guarantee that the error risk on the unseen test set is bounded by a user-specified level α (Angelopoulos et al., 2024). Such a guarantee is finite-sample and distribution-free, requiring only two conditions: the exchangeability of the calibration and test pairs and the non-increasing monotonicity of the loss function in λ . Fundamentally, this selective paradigm achieves risk control by trading coverage for accuracy, abstaining whenever the judge is unsure. Consequently, preserving the practical utility of this approach requires a discriminative, human-aligned scoring function to maximize evaluation coverage while strictly adhering to the specified risk bound.

However, prior work on CRC for LLM evaluation commonly relies on a single-expert paradigm, which is empirically miscalibrated for pairwise comparisons (Jung et al., 2025; Badshah et al., 2026). While advanced models partially mitigate shallow artifacts such as position bias (Zheng et al., 2023) and self-enhancement (Panickssery et al., 2024) (Table 1), deeper model-specific biases per-

[†]Corresponding Authors.

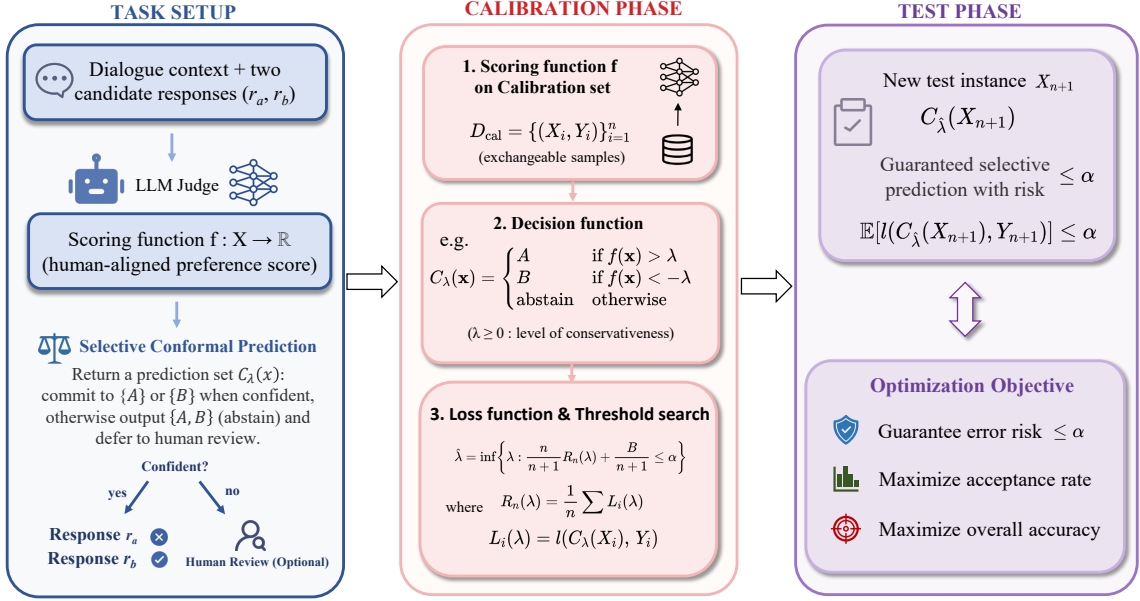


Figure 1: Overview of the selective prediction pipeline under Conformal Risk Control (CRC). **Left:** the task setup defines evaluation instances and the calibration set. **Middle:** the calibration phase constructs a scoring function f and a parameterized decision function C_λ , then searches for the smallest threshold $\hat{\lambda}$ that keeps the empirical risk below α . **Right:** the test phase deploys $C_{\hat{\lambda}}$ on new instances with a distribution-free guarantee, while the optimization objective is to maximize acceptance rate and accuracy under the risk constraint.

sist as a fundamental bottleneck (Koo et al., 2024; Dorner et al., 2025): intrinsic preferences shaped by training data, prompt sensitivities, and domain-specific competence gaps remain embedded in any individual evaluator (Feuer et al., 2025; Stureborg et al., 2024). Lacking explicit ground truth, open-ended dialogues severely amplify these idiosyncrasies, exposing a structural limitation unsolvable by any solitary judge. Judge ensembles offer a natural solution (Verga et al., 2024): whether implemented through multi-agent debate or multi-model panels, they build a robust consensus that yields more discriminative and human-aligned evaluation (Chan et al., 2024; Zhao et al., 2025b; Verga et al., 2024; Zhao et al., 2025a; Qian et al., 2026). Yet how multi-expert mechanisms can be integrated into CRC, and how much coverage they could recover at a fixed risk level, therefore remain largely unexplored.

In this paper, we present the first formal multi-expert CRC framework for pairwise LLM-as-a-Judge evaluation in open-ended dialogue settings. Our framework is built on the core insight that multi-expert aggregation offers a complementary remedy to CRC: whereas CRC controls risk at the decision threshold through abstention, multi-expert aggregation sanitizes the scoring function at its source. Guided by this, we design two CRC-

adapted multi-expert strategies: Score Averaging, which aggregates at the scoring function f , and Decision Voting, which aggregates at the decision function C_λ . We study these strategies on two kinds of expert panels: *homogeneous* (the same model under varied prompts or random seeds) and *heterogeneous* (different models evaluating independently). While both clearly improve over single-expert approaches in the homogeneous setting, in the heterogeneous setting they remain risk-valid yet recover less coverage, because a single shared threshold cannot match the experts’ distinct scoring scales and silences the narrow-range ones (§3.2).

To further address the heterogeneous setting, we propose *Marginal-Calibrated Conformal Consensus* (MC³). MC³ uses per-expert independent calibration as initialization to estimate the ratio between experts’ thresholds, then performs joint threshold searching over a single global scalar t that scales all per-expert thresholds proportionally as $\lambda_j(t) = t \cdot \lambda_j^{(0)}$. By defining a unified joint vote decision function $C_t(x)$ used identically in both calibration and test phases, MC³ preserves exchangeability and the finite-sample, distribution-free risk guarantee of CRC (Appendix C). Empirically, by retaining the multi-expert ensemble while explicitly accounting for the heterogeneity across experts, MC³ attains the highest acceptance rate among

CRC-compliant methods on all three datasets, at accuracy comparable to Decision Voting and while preserving the formal risk guarantee throughout.

Existing dialogue benchmarks either annotate within-model preferences (e.g., ESC-Pro (Zhao et al., 2025c), EmPO (Sotolar et al., 2024)) or rely on closed-source LLMs without logit access (e.g., HEART (Iyer et al., 2026), o2mDial (Lee et al., 2025)), leaving white-box CRC evaluation unsupported. To fill this gap, we construct PANEL: 1,800 exhaustive pairwise comparisons between four open-weight LLMs across three open-ended dialogue domains (ESConv (Liu et al., 2021), MSC (Xu et al., 2022), DREAM (Sun et al., 2019)), with human preference labels under a multi-dimensional rubric and full logit access.

We make three key contributions:

- Through a systematic empirical study, we identify the optimal conformity score for pairwise LLM-as-a-Judge CRC in the open-ended dialogue settings, filling a gap in prior work.
- Based on this, we propose the first formal multi-expert CRC framework, with *Score Averaging* and *Decision Voting* for homogeneous expert settings, and MC³ for heterogeneous expert settings.
- We construct PANEL, on which our framework substantially improves accuracy and acceptance rate over single-expert baselines while maintaining the target risk level.

2 Problem Formulation

Overview. As illustrated in Figure 1, we formulate LLM-based pairwise evaluation as a *selective prediction* problem under the Conformal Risk Control (CRC) framework (Angelopoulos et al., 2024). Let $\mathcal{X}$ denote the space of evaluation instances, where each $x \in \mathcal{X}$ comprises a dialogue context paired with two candidate responses (r_a, r_b), and let $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$ denote an exchangeable calibration set of instances with human preference labels. Given an LLM judge that compares such pairs, our system performs selective prediction by returning a prediction set $C_\lambda(x) \subseteq \mathcal{Y}$. A singleton output, $C_\lambda(x) = \{A\}$ or $\{B\}$, indicates that the judge is confident enough to commit to a preference (r_a or r_b , respectively). Conversely, when uncertain, the judge outputs the full set $C_\lambda(x) = \{A, B\}$, effectively *abstaining* and deferring the instance to human evaluation. The CRC pipeline consists of (i) a *calibration phase* that uses $\mathcal{D}_{\text{cal}}$ to search

a threshold $\hat{\lambda}$, and (ii) a *test phase* that deploys $\hat{\lambda}$ on unseen instances with the formal risk guarantee. Our objective is to maximize both the *accuracy*, which measures the judge’s alignment with human preferences over all pairs regardless of acceptance, and the *acceptance rate*, which represents the overall coverage of accepted predictions across all samples.

Scoring Function and Decision Function. A scoring function $f: \mathcal{X} \rightarrow \mathbb{R}$ assigns each instance a real-valued preference score reflecting the confidence of the LLM judge in preferring one response. A threshold parameter $\lambda \geq 0$ then defines the prediction set:

$$C_\lambda(x) = \begin{cases} \{A\} & \text{if } f(x) > \lambda \\ \{B\} & \text{if } f(x) < -\lambda \\ \{A, B\} & \text{otherwise (abstain)} \end{cases} \quad (1)$$

Here λ controls conservativeness: larger λ raises the acceptance bar, increasing abstentions at the cost of coverage.

Loss Function and Threshold Searching. For each calibration instance (X_i, Y_i) , the loss is the *miscoverage* of the prediction set:

$$L_i(\lambda) = \mathbf{1}[Y_i \notin C_\lambda(X_i)] \quad (2)$$

which is non-increasing in λ (Lemmas 2 and 4): larger λ accepts fewer instances and thus incurs fewer errors. CRC then selects $\hat{\lambda}$ as:

$$\hat{\lambda} = \inf \left\{ \lambda : \frac{n}{n+1} R_n(\lambda) + \frac{B}{n+1} \leq \alpha \right\} \quad (3)$$

where $R_n(\lambda) = \frac{1}{n} \sum_{i=1}^n L_i(\lambda)$ is the empirical risk and B is a uniform upper bound on the loss ($L_i(\lambda) \leq B$ for every λ). The term $B/(n+1)$ bounds the worst-case contribution of the single unseen test point to the risk, which is what lets the guarantee hold at any finite n . Since the loss in Eq. 2 is the 0/1 miscoverage indicator, it deterministically takes values in $\{0, 1\}$: $L_i(\lambda) = 1$ only when the judge commits to a single response that disagrees with the human label, whereas an abstention returns $\{A, B\}$ which trivially contains Y_i and hence $L_i = 0$. Consequently, $B = 1$ is the tightest valid upper bound throughout. The inf yields the smallest qualifying λ , maximizing acceptance under the risk constraint.

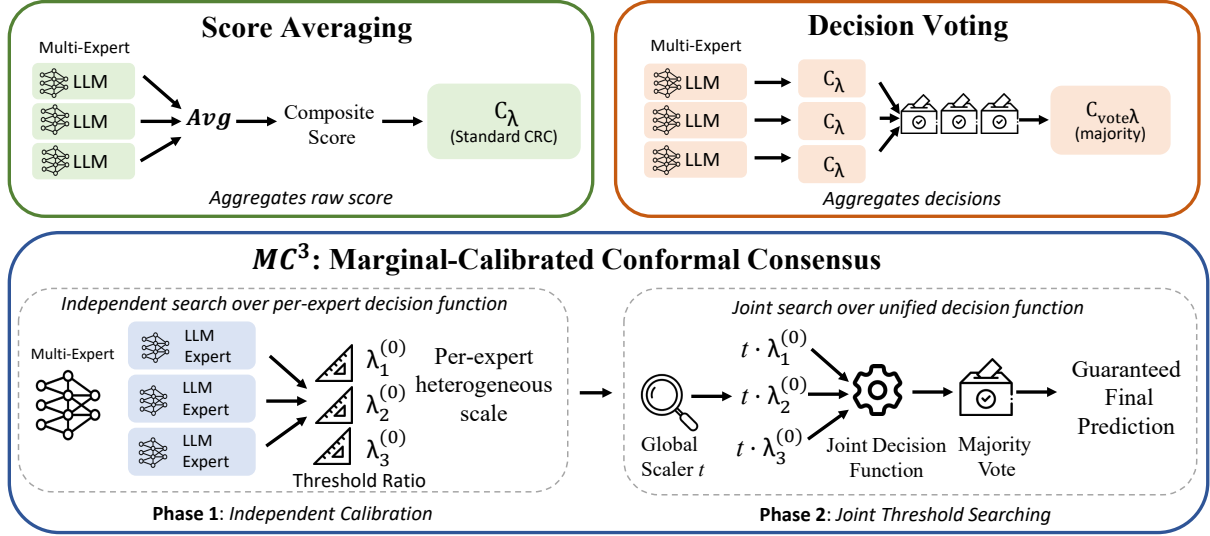


Figure 2: Overview of the three multi-expert CRC strategies. **Score Averaging** (top left) aggregates per-expert scores into a composite f_{avg} . **Decision Voting** (top right) lets each expert make an independent decision $C_{\lambda}^{(j)}$ with a shared threshold λ , then aggregates via majority vote. **MC³** (bottom) first initializes per-expert threshold ratios $\lambda_j^{(0)}$ via independent calibration (Phase 1), then scales them proportionally by a global scalar t and performs joint threshold searching over a unified decision function $C_t(x)$ (Phase 2), simultaneously preserving per-expert heterogeneity and the exchangeability required by CRC.

Deployment Guarantee. For a new test instance X_{n+1} drawn exchangeably with the calibration data, the calibrated predictor $C_{\hat{\lambda}}$ satisfies:

$$\mathbb{E}[\ell(C_{\hat{\lambda}}(X_{n+1}), Y_{n+1})] \leq \alpha \quad (4)$$

This guarantee is distribution-free, model-agnostic, and finite-sample, requiring only exchangeability and loss monotonicity in λ .

3 Multi-Expert Conformal Consensus

Surface biases such as position bias and self-enhancement fade as models scale up (Table 1), but model-specific tendencies, distinct scoring scales, domain preferences, and prompt sensitivities, persist regardless of scale (Koo et al., 2024). Multi-expert aggregation reduces the variance from these residual biases, which single-model improvements cannot. Yet no prior work has studied multi-expert CRC with formal risk guarantees, a gap we address in this section.

3.1 Per-Expert Base Score Selection

We first conduct an extensive empirical study of base scoring functions for CRC in open-ended dialogue (see §6 for full results). Among all candidates, *logit-based pairwise preference probability* emerges as the optimal single-expert conformity score. Let $z_A^{(j)}, z_B^{(j)}$ denote the judge’s logits at the

first generated token position for tokens A, B on instance x . Each expert’s scoring function is a binary softmax over these two logits:

$$f_j(x) = \frac{\exp(z_A^{(j)})}{\exp(z_A^{(j)}) + \exp(z_B^{(j)})} - 0.5 \quad (5)$$

Being continuous, this score allows the threshold to be searched at fine granularity, and it satisfies the monotonicity required by CRC. It also achieves the strongest alignment with human preferences among all candidates (Table 1). We therefore adopt $f_j(x)$ as the default per-expert base score for multi-expert CRC throughout this work.

3.2 Score Averaging and Decision Voting

Within the CRC pipeline of §2, multiple experts can be combined at two natural points: the scoring function f and the decision function C_{λ} . We design one CRC-adapted strategy for each point, with all experts sharing a single threshold λ for simplicity:

Score Averaging modifies the scoring function f while leaving the pipeline unchanged. The K per-expert scores are averaged into a composite:

$$f_{\text{avg}}(x) = \frac{1}{K} \sum_{j=1}^K f_j(x) \quad (6)$$

and standard CRC is applied to f_{avg} : decision function C_λ , loss $L_i(\lambda)$, and threshold searching all remain identical to the single-expert case.

Decision Voting retains each per-expert scoring function $f_j(x)$ but modifies the decision function C_λ . All experts share a unified threshold λ ; each independently produces a per-expert decision $C_\lambda^{(j)}(x)$ following Eq. 1, and a majority vote aggregates them into a joint decision function:

$$C_\lambda^{\text{vote}}(x) = \text{MajVote}(C_\lambda^{(1)}(x), \dots, C_\lambda^{(K)}(x)) \quad (7)$$

The system accepts a prediction when a strict majority of experts agree on a direction, and abstains otherwise.

Adaptation on Homogeneous and Heterogeneous multi-experts. When all K experts are instances of the same model under varied prompts or seeds (*homogeneous* setting), every f_j shares the same scoring scale by construction. Score Averaging and Decision Voting therefore preserve the underlying distribution, and a shared λ treats all experts fairly: both strategies maintain the risk guarantee (Appendix C.3 and C.4) and substantially improve accuracy over a single expert (Table 2). When experts come from different models (*heterogeneous* setting), their scoring distributions become incompatible. When experts operate at different scoring scales (e.g., $\lambda \approx 0.05$ for a 12B model vs. $\lambda \approx 0.20$ for an 8B model), a single shared threshold treats them asymmetrically: the narrow-scale expert never crosses the threshold and is effectively *silenced*, always casting an abstain vote. The system thereby degrades from a multi-expert ensemble into a single-expert one, since only the wide-scale expert can ever contribute a non-abstaining vote. *Score Averaging* compounds this by first collapsing the incompatible f_j distributions into a composite whose scale matches no individual expert; per-expert calibration (e.g., Platt scaling (Platt et al., 1999)) could address this but at prohibitive complexity. Neither strategy resolves the underlying issue, so both attain low acceptance rates and fail to fully exploit multi-expert benefits.

4 Marginal-Calibrated Conformal Consensus

4.1 Marginal Calibration Initialization

A natural approach to this is *Test-time Voting*: assign each expert j an independent threshold via the

Algorithm 1 MC³

Require: Calibration set $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$, per-expert scores $f_1, \dots, f_K$, risk level α , loss ℓ with upper bound B ; per-expert decision $C_\lambda^{(j)}$ from Eq. 1.

Phase 1: Marginal Calibration $\triangleright$ *per-expert scale*

- 1: **for** $j = 1, \dots, K$ **do**
- 2: $R_n^{(j)}(\lambda) \leftarrow \frac{1}{n} \sum_{i=1}^n \ell(C_\lambda^{(j)}(X_i), Y_i)$
- 3: $\lambda_j^{(0)} \leftarrow \inf\{\lambda \geq 0 : R_n^{(j)}(\lambda) \leq \alpha\}$
- 4: **end for**

Phase 2: Joint Threshold Searching $\triangleright$ *global scalar t*

- 5: Define $\lambda_j(t) \leftarrow t \cdot \lambda_j^{(0)}$ for all j
- 6: $C_t(x) \leftarrow \text{MajVote}(C_{\lambda_1(t)}^{(1)}(x), \dots, C_{\lambda_K(t)}^{(K)}(x))$
- 7: $R_n(t) \leftarrow \frac{1}{n} \sum_{i=1}^n \ell(C_t(X_i), Y_i)$
- 8: $\hat{t} \leftarrow \inf\{t \geq 0 : \frac{n}{n+1} R_n(t) + \frac{B}{n+1} \leq \alpha\}$
- 9: **return** $\hat{t}$ and $\{\lambda_j = \hat{t} \cdot \lambda_j^{(0)}\}_{j=1}^K$

Deployment. On test X_{n+1} , output $C_{\hat{t}}(X_{n+1})$.

Guarantee (Theorem 2): $\mathbb{E}[\ell(C_{\hat{t}}(X_{n+1}), Y_{n+1})] \leq \alpha$.

per-expert analogue of Eq. 3:

$$\lambda_j^{(0)} = \inf\{\lambda : R_n^{(j)}(\lambda) \leq \alpha\} \quad (8)$$

and aggregate via majority vote at test time. These thresholds capture the scoring-scale heterogeneity identified in §3.2, but $\lambda_j^{(0)}$ cannot serve as the final threshold: during calibration each expert is evaluated in isolation, while at test time a collective vote produces the prediction. This mismatch breaks exchangeability of the loss random variables required by CRC (Appendix C.6).

Our insight is to use $\lambda_j^{(0)}$ as *ratio initialization*, not as the final threshold. It estimates the approximate *ratio* between per-expert thresholds at the target risk level. §4.2 uses this ratio as a fixed shape to define a unified decision function $C_t(x)$ that preserves exchangeability (Theorem 2).

4.2 Joint Threshold Searching

Using the ratios from §4.1 as a fixed shape, we introduce a single global scalar $t \geq 0$ that scales all per-expert thresholds proportionally:

$$\lambda_j(t) = t \cdot \lambda_j^{(0)} \quad (9)$$

We define the *joint vote decision function*, replacing $C_\lambda(x)$ in §2:

$$C_t(x) = \text{MajVote}(C_{\lambda_1(t)}^{(1)}(x), \dots, C_{\lambda_K(t)}^{(K)}(x)) \quad (10)$$

Analogous to λ in Eq. 1, t encodes the level of conservativeness: as t increases, all $\lambda_j(t)$ scale up proportionally, raising the abstention bar for

each expert and producing more abstentions, so $\ell(C_t(X_i), Y_i)$ is non-increasing in t (Lemma 4), satisfying the monotonicity requirement of CRC. Threshold searching is performed directly over t :

$$\hat{t} = \inf \left\{ t \geq 0 : \frac{n}{n+1} R_n(t) + \frac{B}{n+1} \leq \alpha \right\} \quad (11)$$

yielding final per-expert thresholds $\lambda_j = \hat{t} \cdot \lambda_j^{(0)}$. Sharing a single scalar t across experts is primarily a search-efficiency choice, reducing a K -dimensional joint search to a one-dimensional problem along the linear ray $\lambda_j(t) = t \cdot \lambda_j^{(0)}$, rather than the mechanism securing the CRC bound; the bound itself rests on the two conditions (P1) monotonicity and (P2) exchangeability. Concretely, because each calibration loss $L_i = \ell(C_t(X_i), Y_i)$ and the test loss $L_{n+1} = \ell(C_t(X_{n+1}), Y_{n+1})$ are computed by the same function applied to exchangeable data, the loss random variables $\{L_1, \dots, L_{n+1}\}$ remain exchangeable, and the CRC bound applies (Theorem 2). Appendix C shows formal proofs for multi-expert methods.

MC³ simultaneously achieves three properties that no approach from §3 provides: (1) risk guarantee via a shared C_t across phases; (2) per-expert heterogeneity, as λ_j differ across experts; and (3) continuous parameterization over the scalar t . The procedure is summarized in Algorithm 1.

5 Benchmark Construction

Evaluating CRC for open-ended dialogue requires a pairwise preference benchmark in which every candidate model exposes its logits, yet no such benchmark currently exists. We construct such a benchmark by sampling 100 dialogue contexts from each of three datasets spanning diverse conversational scenarios: ESConv (Liu et al., 2021) (emotional support), MSC (Xu et al., 2022) (multi-session social chat), and DREAM (Sun et al., 2019) (dialogue comprehension). For each context, candidate responses are generated by four open-weight LLMs: gemma-3-12b-it (Kamath et al., 2025), Mistral-Nemo-Instruct-2407 (Mistral AI, 2024), Qwen2.5-7B-Instruct (Yang et al., 2024), and Llama-3.1-8B-Instruct (Grattafiori et al., 2024). We enumerate all $\binom{4}{2} = 6$ response pairs per context, yielding 1,800 annotated pairs in total with human pairwise preference labels across five quality dimensions.¹ The detailed annotation

protocol is provided in Appendix A. Regarding potential data contamination: although the dialogue contexts predate the judge models’ pre-training cut-offs, both the candidate responses and the human preference labels are newly produced in this work, so the response–label pairs used for CRC calibration remain uncontaminated; our meta-evaluation of judge–human alignment therefore gains no shortcut from context memorization.

The four LLMs in PANEL admit two complementary orderings: generation quality (measured by human preference win rates) and single-judge evaluation accuracy (measured against human pairwise labels). We report both in Appendix A (Figs. 3 and 4); they agree on the same model being strongest along both axes but otherwise diverge, motivating a multi-expert framework robust to individual-judge weakness. The generator ordering directly governs CRC calibration difficulty: well-separated generators (e.g., on ESConv) yield large pairwise margins that make calibration tractable, whereas clustered generators (e.g., on MSC) compress margins and demand more discriminative conformity scores.

6 Experiments

6.1 Experimental Setup

Setup. We evaluate all scoring functions on each dataset and report results averaged over 5 random 1:1 calibration/test splits, partitioned by *conversation ID* to prevent leakage between turns of the same conversation. The risk level is set to $\alpha = 0.1$ throughout for all methods. All multi-expert methods are based on pairwise logit-based preference unless otherwise noted. All experiments are conducted on NVIDIA A800 80GB GPUs.

Scoring Functions. We evaluate pointwise scores (log probability, Self-Certainty (Kang et al., 2025), DeepConf (Fu et al., 2025), Causal (Feng et al., 2025), Consistency (Wang et al., 2022)) and pairwise scores (Pref. Prob., Verbalized Confidence, Rubric). For pointwise scores, the pairwise conformity signal is the difference between the two response scores, with the sign indicating preference and the magnitude indicating confidence. Per-method implementation details are in Appendix B.

Evaluator Selection. We use the top- K strongest evaluators throughout: $K=1$ for single-judge experiments, and $K=3$ for multi-expert experiments, with the joint decision committed via standard majority voting. Evaluators are ranked by

¹PANEL is publicly available at <https://huggingface.co/datasets/EstellaCheng42/panel>.

Table 1: Main results across scoring functions and three datasets ($\alpha = 0.1$). *Pointwise*: each response scored independently, pairwise signal derived from score difference. *Pairwise*: a separate judge LLM directly compares two responses via logit extraction. *Multi.*: multi-expert ensembles using homogeneous experts (same model, varied prompts/seeds). **AccR**: acceptance rate; **Acc**: pairwise accuracy; **AUC**: area under the ROC curve; **Risk**: fraction of all pairs committed to an incorrect preference. CRC requires Risk $\leq \alpha$. Within each category and dataset, the best result is in **bold**. [†]Pref. Prob. w/ swap is the bidirectional preference scoring (BPE) of Badshah et al. (2026).

	Score	ESConv				MSC				DREAM			
		AccR $\uparrow$	Acc $\uparrow$	AUC $\uparrow$	Risk $\downarrow$	AccR $\uparrow$	Acc $\uparrow$	AUC $\uparrow$	Risk $\downarrow$	AccR $\uparrow$	Acc $\uparrow$	AUC $\uparrow$	Risk $\downarrow$
Pointwise	Causal	.644	.743	.828	.124	.351	.629	.677	.121	.473	.590	.685	.140
	Consistency	.356	.654	.687	.116	.158	.553	.528	.082	.158	.522	.497	.069
	Self-Certainty	.481	.661	.684	.122	.381	.618	.753	.064	.322	.556	.628	.076
	DeepConf	.480	.668	.624	.185	.309	.667	.690	.094	.255	.532	.565	.096
	Log Prob	.794	.834	.910	.091	.389	.647	.676	.118	.547	.702	.738	.135
	w/ self-judge	.388	.643	.684	.087	.377	.688	.693	.097	.396	.620	.687	.100
Pairwise	Rubric	.314	.619	.713	.042	.196	.486	.535	.079	.172	.573	.678	.036
	Verb. Conf.	.747	.769	.880	.107	.390	.641	.699	.108	.447	.662	.791	.084
	Pref. Prob.	.815	.873	.937	.094	.346	.654	.692	.103	.463	.720	.796	.093
	w/ swap [†]	.847	.854	.913	.097	.348	.663	.743	.074	.479	.728	.809	.094
	w/ self-judge	.852	.876	.936	.092	.405	.672	.736	.095	.548	.692	.784	.101
Multi.	Score Averaging	.951	.882	.940	.094	.489	.712	.800	.088	.550	.751	.830	.085
	Decision Voting	.973	.887	.945	.098	.495	.732	.798	.097	.566	.755	.830	.082

single-judge accuracy against human pairwise labels; detailed rankings are reported in Appendix A. Rows marked “w/ swap” implement Badshah et al. (2026)’s bidirectional preference scoring (BPE): the judge is queried twice with candidate order swapped, and the preference probabilities are averaged. In the multi-expert setting, a *homogeneous* ensemble queries the same model K times with varied random seeds and prompt templates at temperature 0.7; a *heterogeneous* ensemble uses the top- K distinct models.

Metrics. Within CRC, risk is held below α by abstaining on uncertain pairs, so the operative question is how much coverage a method retains under this guarantee. We report four metrics, all computed over non-tie pairs (human-annotated ties are excluded): *Acceptance Rate* (AccR $\uparrow$, the fraction of pairs not abstained), the coverage recovered under the risk constraint; *Accuracy* (Acc $\uparrow$, the fraction of pairs whose score-difference sign matches the human preference label, computed over all pairs regardless of CRC acceptance), which gauges the quality of the aggregated decision independently of abstention; *AUC* ($\uparrow$, area under the ROC curve, capturing the threshold-independent discriminative power of the scoring function); and *Risk* ($\downarrow$, the rate of wrong-and-accepted predictions, with abstentions incurring zero loss; CRC ensures Risk $\leq \alpha$, with a value closer to α indicating tighter calibration), the quantity CRC provably controls.

6.2 Results and Analysis

Base score selection (Table 1). Table 1 reports our empirical study of base scoring functions for CRC in open-ended dialogue. The base score is the most critical component of CRC: it determines whether the calibrated threshold reflects genuine preference uncertainty. Yet no prior work has systematically identified the most effective scoring functions. We address this gap by comparing two families: pointwise scores (e.g., Log Prob, Self-Certainty, DeepConf), where each response is scored independently, and pairwise scores (e.g., Pref. Prob., Verbalized Confidence, Rubric), where the judge directly compares two responses. Empirically, logit-based pairwise extraction (Pref. Prob.) outperforms the strongest pointwise score (Log Prob) in both accuracy and AUC across all three datasets, while keeping risk near α . This advantage reflects a deeper alignment between pairwise judgment and CRC: since open-ended dialogue lacks an absolute quality criterion, relative preference within a shared context provides a more natural conformity signal. We therefore adopt Pref. Prob. as the per-expert base score for all multi-expert experiments.

Bias robustness (Table 1). Known biases in LLM-as-a-Judge evaluation have a diminishing impact with pairwise logit-based scoring. Self-judge bias, which is severe under pointwise scoring (Log Prob drops by 19.1 pp on ESConv), be-

Table 2: Multi-expert aggregation results across three datasets ($\alpha = 0.1$, 5 random 1:1 cal/test splits). **AccR**: acceptance rate; **Acc**: pairwise accuracy; **Risk**: fraction of all pairs committed to an incorrect preference. CRC requires Risk $\leq \alpha$. The best result is in **bold**. [†]Test-time Voting breaks exchangeability between calibration and test, voiding its CRC guarantee, and is excluded from the ranking.

Ensemble	Strategy	ESConv			MSC			DREAM		
		AccR $\uparrow$	Acc $\uparrow$	Risk $\downarrow$	AccR $\uparrow$	Acc $\uparrow$	Risk $\downarrow$	AccR $\uparrow$	Acc $\uparrow$	Risk $\downarrow$
Homogeneous	Score Averaging	.951	.882	.094	.489	.712	.088	.550	.751	.085
	Decision Voting	.973	.887	.098	.495	.732	.097	.566	.755	.082
Heterogeneous	Score Averaging	.845	.853	.092	.396	.685	.086	.508	.715	.089
	Decision Voting	.846	.866	.090	.364	.677	.083	.449	.725	.089
	Test-time Voting [†]	.962	.866	.118	.670	.677	.139	.692	.725	.154
	MC ³	.894	.866	.092	.462	.677	.083	.540	.725	.081

comes negligible in the pairwise setting: enabling self-judging changes accuracy by less than 3 pp across all datasets and in both directions, well within typical run-to-run variation. Position bias likewise yields only marginal gains under swap correction (around 1 pp on average), suggesting its impact is limited at current model scale. These surface biases cannot explain the broader gains of multi-expert aggregation, which addresses model-idiosyncratic tendencies that simple rotation alone cannot reach. Take-away: surface biases (position, self-enhancement) are largely resolved at current model scale under pairwise-logit scoring, yet the persistent gains from multi-expert aggregation in Table 2 point to a deeper, model-specific bias that no single-expert correction can reach, motivating consensus prediction across experts.

Multi-expert in homogeneous settings (Tables 1, 2). When all K experts share the same model (varied prompts and seeds), Decision Voting improves over a single Pref. Prob. judge on both acceptance rate and accuracy across all three datasets, most strongly on acceptance rate (up to +15.8 pp), while Score Averaging gives smaller, less consistent gains. Risk stays at or below α throughout (at most .098), so the guarantee holds with no violation. The edge of Decision Voting reflects vote-level CRC acting directly on the decision function rather than on a composite score. This matches §3.2: homogeneous experts share one scoring scale, so neither score-level nor threshold-level aggregation introduces a distributional mismatch.

The heterogeneous challenge (Table 2). The calibration strategy becomes decisive when experts come from different models. Both *Score Averaging* and *Decision Voting* stay risk-valid here, yet recover only limited coverage, for a shared rea-

son: a single uniform threshold cannot match the experts’ distinct scoring scales, so narrow-range experts are silenced and the ensemble abstains more often. *Score Averaging* compounds this by first collapsing the incompatible per-expert scores into one composite. *Test-time Voting* instead appears to recover high coverage, but it calibrates each expert in isolation and votes only at test time, so the decision function differs between calibration and test; exchangeability breaks, risk climbs above α (up to .154), and we exclude it from the ranking. Take-away: under a single shared threshold, heterogeneous ensembles collapse toward their most wide-scale expert; recovering true multi-expert coverage requires a decision function that preserves per-expert scales while remaining shared across calibration and test, which is exactly what MC³ delivers.

MC³ in the heterogeneous setting (Table 2). Ratio initialization (§4.1) captures each expert’s native scale via $\lambda_j^{(0)}$, and joint threshold searching (§4.2) finds a global scalar $\hat{t}$ that proportionally rescales all per-expert thresholds through the unified decision function $C_t(x)$. As a result, MC³ recovers the highest acceptance rate among the CRC-valid methods on all three datasets, while keeping risk within α . Its accuracy closely tracks that of the other voting methods, since all use the same majority vote and differ only in threshold selection; the methods separate on AccR and Risk, not on Acc, which scores every pair regardless of acceptance. The full benefit of multi-expert evaluation emerges only when calibration and aggregation use the same decision function. Aggregation should therefore act at the decision level with per-expert thresholds, rather than at the score level through a shared composite.

7 Related Work

Conformal Prediction in LLMs. Conformal prediction has been increasingly applied to quantify uncertainty in LLM generation tasks. Early extensions apply conformal methods to multiple-choice QA (Kumar et al., 2023), open-ended language modeling (Quach et al., 2023), and API-based settings without logit access (Su et al., 2024). Two recent works extend CRC to pairwise LLM-as-a-Judge evaluation: Jung et al. (2025) calibrates a single LLM judge by sampling multiple simulated annotators from one model to provide a provable guarantee on human agreement, and Badshah et al. (2026) optimizes the conformity score function itself to improve selective pairwise judging. However, neither addresses multi-expert pairwise LLM-as-a-Judge CRC.

Preference Datasets for Open-Ended Dialogues. Preference datasets for LLM alignment have grown rapidly (e.g., HH-RLHF (Bai et al., 2022), Ultra-Feedback (Cui et al., 2024)), yet primarily target general instruction-following along a single quality dimension. Open-ended conversational scenarios remain underserved. Many dialogue-focused datasets rely on *within-model* preference pairs and provide no human pairwise comparisons across different LLMs; ESC-Pro (Zhao et al., 2025c) and EmPO (Sotolar et al., 2024) are representative cases. Among the few that attempt *between-model* evaluation, HEART (Iyer et al., 2026) evaluates proprietary black-box models without logit access, and o2mDial (Lee et al., 2025) covers only simple chitchat via closed-source APIs. Neither supports white-box calibration.

8 Discussion

Two design choices in our framework merit clarification. Our framework requires logit access, restricting applicability to open-weight LLMs. This aligns with how LLM judges are actually deployed in high-stakes, privacy-sensitive settings, e.g., clinical or enterprise pipelines under HIPAA/GDPR, where local open-weight judges such as Prometheus (Kim et al., 2024) are the norm; API-only judges instead require alternative conformity scores (Su et al., 2024). Our 0/1 loss further treats abstention as free, an intentional fit for high-stakes settings where errors are intolerable and human deferral is a safe fallback, keeping the target risk directly interpretable as an exact error rate. De-

ferral cost is tracked transparently via Acceptance Rate (e.g., MC³’s AccR of 0.462 on MSC at $\alpha=0.1$ acknowledges a $\sim 54\%$ deferral rate); extending to cost-sensitive settings only requires swapping the abstention cost in the loss.

9 Conclusion

We formulated pairwise LLM-as-a-Judge evaluation as selective prediction under CRC and conducted the first systematic study of multi-expert CRC. We designed two CRC-adapted strategies, *Score Averaging* and *Decision Voting*, that work well for homogeneous experts but recover only limited coverage for heterogeneous ones. To resolve this, we proposed MC³, which uses per-expert threshold ratios under a unified decision function $C_t(x)$ shared by calibration and test, preserving exchangeability. On PANEL, our 1,800-pair human pairwise-preference benchmark with full logit access, MC³ recovers the highest acceptance rate among CRC-valid methods across all three domains, at comparable accuracy and while maintaining the formal risk guarantee.

Acknowledgements

This research was supported by the Australian Government through the Australian Research Council’s Discovery Project funding scheme (Grant No.: DP260100218). Dr Qihong Ke is the recipient of an Australian Research Council Discovery Early Career Researcher Award (project number DE250100030) funded by the Australian Government.

Limitations

Our framework has two open limitations. First, we address only the selective-prediction frontend (commit vs. abstain); the design of the downstream human-review workflow for abstained cases remains open. Second, the CRC guarantee relies on exchangeability, which can break under distribution shift. A practical mitigation is to monitor per-expert threshold ratios $\lambda_j^{(0)}$ on a rolling window and rerun Algorithm 1 once drift exceeds a preset threshold; this recalibration is cheap because CRC is a training-free threshold search.

Ethics Statement

All preference labels in PANEL are produced in-house by three annotators from our research

team, without crowdsourcing, following the multi-dimensional rubric in Appendix A; no separate institutional ethics review was required, as annotation was carried out internally over publicly released, non-identifying dialogue data. All dialogue contexts are used consistent with their original licenses: ESConv (Liu et al., 2021) (CC BY-NC 4.0), MSC (Xu et al., 2022) (CC BY-NC 4.0, Meta ParlAI), and DREAM (Sun et al., 2019) (non-commercial research use); none contain PII, and ESConv’s emotionally sensitive conversations are pseudonymous in the original release. The four candidate response models are used under their public licenses: gemma-3-12b-it (Kamath et al., 2025) (Gemma Terms of Use), Mistral-Nemo-Instruct-2407 (Mistral AI, 2024) (Apache 2.0), Qwen2.5-7B-Instruct (Yang et al., 2024) (Apache 2.0), and Llama-3.1-8B-Instruct (Grattafiori et al., 2024) (Llama 3.1 Community License). All response generation and judge evaluation run locally on institutional GPUs, without transmitting dialogue content to third-party APIs.

References

- Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. 2024. Conformal risk control. In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Sher Badshah, Ali Emami, and Hassan Sajjad. 2026. SCOPE: Selective conformal optimized pairwise LLM judging. *arXiv preprint arXiv:2602.13110*.
- Yuntao Bai, Andy Jones, Kamal Ndousse, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *International Conference on Learning Representations (ICLR)*.
- Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Serkan O. Arik, Tomas Pfister, and Somesh Jha. 2023. Adaptation with self-evaluation to improve selective prediction in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.
- Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Ruobing Xie, Yankai Lin, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2024. Ultrafeedback: Boosting language models with scaled AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*.
- Florian E. Dörner, Vivian Y. Nastl, and Moritz Hardt. 2025. Limits to scalable evaluation at the frontier: LLM as judge won’t beat twice the data. In *International Conference on Learning Representations (ICLR)*. Oral presentation.
- Tao Feng, Lizhen Qu, Xiaoxi Kang, and Gholamreza Haffari. 2025. Causalscore: an automatic reference-free metric for assessing response relevance in open-domain dialogue systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2351–2369.
- Benjamin Feuer, Micah Goldblum, et al. 2025. Style outweighs substance: Failure modes of LLM judges in alignment benchmarking. In *International Conference on Learning Representations (ICLR)*.
- Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. 2025. Deep think with confidence. *arXiv preprint arXiv:2508.15260*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Laya Iyer, Kriti Aggarwal, Sanmi Koyejo, Gail Heyman, Desmond C. Ong, and Subhabrata Mukherjee. 2026. HEART: A unified benchmark for assessing humans and LLMs in emotional support dialogue. *arXiv preprint arXiv:2601.19922*.
- Jaehun Jung, Faeze Brahman, and Yejin Choi. 2025. Trust or escalate: LLM judges with provable guarantees for human agreement. In *International Conference on Learning Representations*.
- Aishwarya Kamath, Johan Ferret, Shreya Pathak, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Zhewei Kang, Xuandong Zhao, and Dawn Song. 2025. Scalable best-of-n selection for large language models via self-certainty. *arXiv preprint arXiv:2502.18581*.
- Seungone Kim, Jamin Shin, Yejin Cho, et al. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations (ICLR)*.
- Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*.
- Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. 2023. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*.
- Jing Yang Lee, Kong Aik Lee, and Woon Seng Gan. 2025. Modeling the one-to-many property in open-domain dialogue with LLMs. In *Proceedings of the*

- Fourth Workshop on Generation, Evaluation and Metrics (GEM)*, pages 276–290.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pages 3469–3483.
- Mistral AI. 2024. Mistral NeMo. <https://mistral.ai/news/mistral-nemo/>.
- Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.
- John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.
- Mengjie Qian, Guangzhi Sun, Mark J.F. Gales, and Kate M. Knill. 2026. Who can we trust? LLM-as-a-jury for comparative assessment. *arXiv preprint arXiv:2602.16610*.
- Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. 2023. Conformal language modeling. *arXiv preprint arXiv:2306.10193*.
- Ondrej Sotolar, Vojtech Formanek, Alok Debnath, Allison Lahnala, Charles Welch, and Lucie Flek. 2024. Empo: Emotion grounding for empathetic response generation through preference optimization. *arXiv preprint arXiv:2406.19071*.
- Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*.
- Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. 2024. Api is enough: Conformal prediction for large language models without logit-access. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 979–995.
- Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.
- Pat Verga, Sebastian Hofstätter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 5180–5197.
- An Yang, Baosong Yang, Beichen Zhang, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Justin Zhao, Flor Miriam Plaza-del Arco, Benjamin Genchel, and Amanda Cercas Curry. 2025a. Language model council: Democratically benchmarking foundation models on highly subjective tasks. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*.
- Ruochoen Zhao, Wenxuan Zhang, Yew Ken Chia, Weiwen Xu, Deli Zhao, and Lidong Bing. 2025b. Autoarena: Automating LLM evaluations with agent peer battles and committee discussions. *arXiv preprint arXiv:2405.20267*.
- Weixiang Zhao, Xingyu Sui, Xinyang Han, et al. 2025c. Chain of strategy optimization makes large language models better emotional supporter. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15361–15381.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

A Dataset and Annotation Details

Datasets. ESConv (Liu et al., 2021) contains dialogues between help-seekers and supporters discussing personal challenges, with emotionally sensitive exchanges that require empathy and strategic support. MSC (Xu et al., 2022) consists of long-term, persona-grounded conversations spanning multiple sessions. DREAM (Sun et al., 2019) is a dialogue-based reading-comprehension dataset comprising short, multi-turn exchanges drawn from English exam materials. Together, these datasets span three complementary registers: short comprehension-oriented exchanges (DREAM), extended persona-grounded social conversation (MSC), and emotionally complex support dialogues (ESConv).

Human Annotation. All pairwise preference labels are obtained via manual annotation by **three independent annotators**. For each response pair generated from the same dialogue context, each annotator independently determines which response is superior along five dimensions: *relevance* to the immediate dialogue history, *specificity* of the information provided, *consistency* with conversational context and common sense, *empathy* in understanding and responding to emotional states, and *overall* quality. The *overall* dimension serves as the final preference label in all subsequent experiments. Annotators may also indicate a *Tie* when the two responses are of comparable quality. To mitigate position bias, the presentation order is randomized for each pair.

Adjudication Protocol. Each of the 1,800 pairs is labeled independently by all three annotators. Pairs on which all three annotators agree on the *overall* dimension are assigned the consensus label directly. For pairs with at least one disagreement, the three annotators convene in a joint discussion session, review the pair together to surface their underlying reasoning, and reach a consensus label through deliberation. Adjudication is applied to the *overall* dimension, which serves as the preference label in all experiments.

Model Quality Rankings. Figures 3 and 4 report the two complementary orderings of the four LLMs in PANEL: human-annotated win rates (generation quality) and pairwise evaluation accuracy (judge quality), respectively.

B Experimental Details

Scoring Function Implementations. The point-wise scores we consider are log probability, Self-Certainty (Kang et al., 2025), and DeepConf (Fu et al., 2025); each requires only a single forward pass per response. The Causal score (Feng et al., 2025) uses fine-tuned RoBERTa classifiers. Consistency (Wang et al., 2022) scoring uses 5 independent samples per response with temperature 0.7, clustered via DeBERTa-based NLI. For Pref. Prob., we extract logits at the *first generated token position* for the tokens A and B as encoded by each model’s own tokenizer; we verified that A and B are single-token under all four model tokenizers given our prompt formatting. The conformity score in Eq. 5 is then computed as a *binary softmax restricted to these two logits*, deliberately excluding the rest of the vocabulary. A full-vocabulary softmax would otherwise absorb probability mass into tokens such as The, leading whitespace, or end-of-sequence that carry no preference content under our prompt template, diluting the comparative signal that motivates pairwise judging. The “w/ swap” variant of Pref. Prob. is the *bidirectional preference scoring* (BPE) introduced by Badshah et al. (2026): we query the judge twice with the candidate order swapped, extract the binary-softmax probabilities under both orderings via the same protocol above, and average them into a position-symmetric per-expert score. We therefore include Badshah et al. (2026)’s BPE directly in our base-score comparison under the “Pref. Prob. w/ swap” row of Table 1. Verbalized Confidence (Verb. Conf.) prompts the judge to output a numerical confidence score along with its preference, while Rubric prompts the judge to select from a 7-point preference scale.

C Proofs of CRC Validity for Multi-Expert Methods

This appendix establishes that the three multi-expert methods of §3 and §4, namely *Score Averaging*, *Decision Voting*, and *MC*³, each satisfy the conditions of the Conformal Risk Control (CRC) theorem of Angelopoulos et al. (2024) and therefore inherit its finite-sample, distribution-free risk guarantee

$$\mathbb{E}[\ell(C_{\hat{\lambda}}(X_{n+1}), Y_{n+1})] \leq \alpha \quad (12)$$

for an exchangeable sequence of calibration and test pairs $(X_i, Y_i)_{i=1}^{n+1}$. Throughout, $\ell(\hat{y}, y) :=$

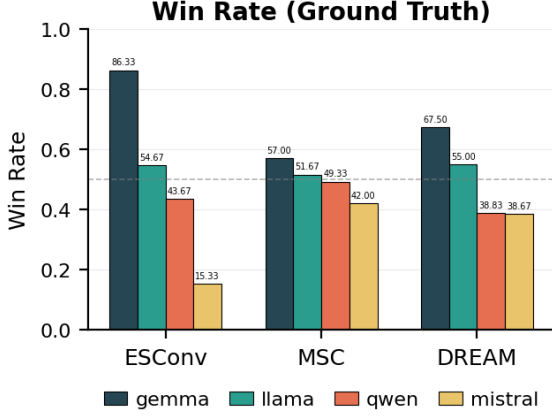


Figure 3: Human-annotated model win rates across the three dialogue datasets. The generator ranking is Gemma-3-12B > Llama-3.1-8B > Qwen2.5-7B > Mistral-Nemo on every dataset.

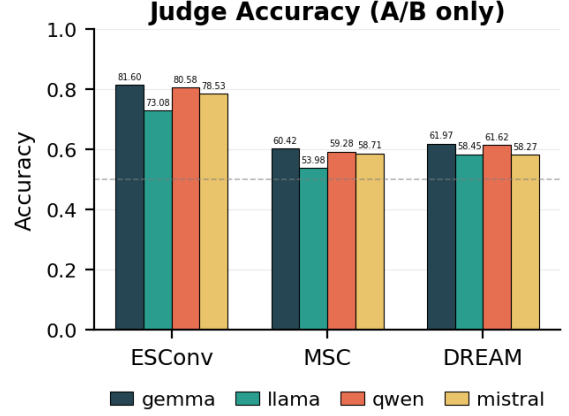


Figure 4: Single-judge accuracy of each LLM against human pairwise labels across the three dialogue datasets. The dataset-averaged ranking is Gemma > Qwen > Mistral > Llama.

$1[\hat{y} \in \mathcal{Y} \wedge \hat{y} \neq y]$ is the indicator loss of Eq. 2, so $\ell \in \{0, 1\}$ and the loss upper bound is $B = 1$.

C.1 Proof Goal and Strategy

Goal. For each of the three methods, with its associated calibrated threshold (denoted $\hat{\lambda}$ for Score Averaging and Decision Voting, and $\hat{t}$ for MC³), we wish to verify Eq. 12.

Strategy. We invoke Theorem 1 of Angelopoulos et al. (2024), which we restate here for completeness:

Theorem 1 (CRC; Angelopoulos et al. 2024, Theorem 1). *Let $\{L_i(\cdot)\}_{i=1}^{n+1}$ be an exchangeable collection of random functions $L_i : \Lambda \rightarrow (-\infty, B]$ satisfying, almost surely, $L_i(\lambda_{\max}) \leq \alpha$ and $L_i(\lambda)$ non-increasing and right-continuous in λ . Define*

$$\hat{\lambda} = \inf \left\{ \lambda : \frac{n}{n+1} \hat{R}_n(\lambda) + \frac{B}{n+1} \leq \alpha \right\},$$

where $\hat{R}_n(\lambda) = \frac{1}{n} \sum_{i=1}^n L_i(\lambda)$. Then $\mathbb{E}[L_{n+1}(\hat{\lambda})] \leq \alpha$.

It suffices to verify two properties for each method:

(P1) Monotonicity. For every sample (X_i, Y_i) , the function $\lambda \mapsto L_i(\lambda)$ (or $t \mapsto L_i(t)$ for MC³) is non-increasing.

(P2) Exchangeability. The collection $\{L_i(\cdot)\}_{i=1}^{n+1}$ is exchangeable as a sequence of random functions. Boundedness ($L_i \in [0, 1]$) and right-continuity (each L_i is a step function in its threshold variable) are immediate. Existence of $\lambda_{\max}$ with $L_i(\lambda_{\max}) = 0$ follows because for any

threshold strictly exceeding $\max_j |f_j(X_i)|$ (resp. $\max_j |f_j(X_i)|/\lambda_j^{(0)}$ for MC³, assuming $\lambda_j^{(0)} > 0$), every expert abstains and the loss vanishes. We can then take $\lambda_{\max} = \infty$ (or any finite upper bound satisfying the same).

We establish (P1) via the voting lemmas of §C.2, and (P2) via the following one-line observation, used uniformly across all three methods:

Lemma 1 (Pointwise exchangeability). *Let $\{Z_i\}_{i=1}^{n+1}$ be an exchangeable sequence of random variables and g a deterministic measurable function. Then $\{g(Z_i)\}_{i=1}^{n+1}$ is exchangeable.*

Proof. For any permutation π of $\{1, \dots, n+1\}$, $(g(Z_{\pi(1)}), \dots, g(Z_{\pi(n+1)}))) = g_*(Z_{\pi(1)}, \dots, Z_{\pi(n+1)}) \stackrel{d}{=} g_*(Z_1, \dots, Z_{n+1}) = (g(Z_1), \dots, g(Z_{n+1}))$, where g_* denotes pointwise application of g . $\square$

C.2 Voting Lemmas

Throughout this subsection, let K be the number of experts and $M = \lfloor K/2 \rfloor + 1$ the strict-majority threshold, satisfying $K - M < M$.

Lemma 2 (Per-expert and vote-count monotonicity). *For each expert j , the per-expert decision $C_{\lambda}^{(j)}(x)$ defined in Eq. 1 satisfies, for any $\lambda_1 \leq \lambda_2$:*

$$C_{\lambda_1}^{(j)}(x) = A \implies C_{\lambda_2}^{(j)}(x) \in \{A, \text{abstain}\},$$

$$C_{\lambda_1}^{(j)}(x) = B \implies C_{\lambda_2}^{(j)}(x) \in \{B, \text{abstain}\},$$

$$C_{\lambda_1}^{(j)}(x) = \text{abstain} \implies C_{\lambda_2}^{(j)}(x) = \text{abstain}.$$

Consequently, the vote counts $v_A(\lambda) := |\{j : C_{\lambda}^{(j)}(x) = A\}|$, $v_B(\lambda) := |\{j : C_{\lambda}^{(j)}(x) = B\}|$,

and $v_{\perp}(\lambda) := |\{j : C_{\lambda}^{(j)}(x) = \text{abstain}\}|$ (with $v_A + v_B + v_{\perp} = K$) satisfy: $\lambda \mapsto v_A(\lambda)$ and $\lambda \mapsto v_B(\lambda)$ are non-increasing, while $\lambda \mapsto v_{\perp}(\lambda)$ is non-decreasing.

Proof. By Eq. 1, $C_{\lambda}^{(j)}(x) = A \iff f_j(x) > \lambda$ and $C_{\lambda}^{(j)}(x) = B \iff f_j(x) < -\lambda$; as λ increases, both events can only turn from true to false and abstention is absorbing, giving the per-expert containment. Each expert transition is therefore one of $A \rightarrow A$, $A \rightarrow \text{abstain}$, $B \rightarrow B$, $B \rightarrow \text{abstain}$, or $\text{abstain} \rightarrow \text{abstain}$, so v_A and v_B can only decrease and $v_{\perp}$ can only increase. $\square$

Lemma 3 (No vote-flip). *For any $\lambda_1 \leq \lambda_2$, the majority-vote decision $C_{\lambda}^{\text{vote}}(x) = \text{MajVote}(C_{\lambda}^{(1)}(x), \dots, C_{\lambda}^{(K)}(x))$ satisfies the same containment as Lemma 2 (with $C^{(j)}$ replaced by C^{vote}).*

Proof. We treat the three cases, using throughout the identity $v_A + v_B + v_{\perp} = K$ from Lemma 2.

Case 1: $C_{\lambda_1}^{\text{vote}}(x) = A$. Then $v_A(\lambda_1) \geq M$, so $v_B(\lambda_1) \leq K - v_A(\lambda_1) \leq K - M$. By Lemma 2, $v_B(\lambda_2) \leq v_B(\lambda_1) \leq K - M < M$, so the vote at λ_2 cannot produce B . Therefore $C_{\lambda_2}^{\text{vote}}(x) \in \{A, \text{abstain}\}$.

Case 2: $C_{\lambda_1}^{\text{vote}}(x) = B$. By the symmetric argument, $C_{\lambda_2}^{\text{vote}}(x) \in \{B, \text{abstain}\}$.

Case 3: $C_{\lambda_1}^{\text{vote}}(x) = \text{abstain}$. Then $v_A(\lambda_1) < M$ and $v_B(\lambda_1) < M$. By Lemma 2, $v_A(\lambda_2), v_B(\lambda_2) < M$. So $C_{\lambda_2}^{\text{vote}}(x) = \text{abstain}$. $\square$

Lemma 4 (Loss monotonicity from no-flip). *Let $\lambda \mapsto C_{\lambda}(x)$ be any decision rule taking values in $\{A, B, \text{abstain}\}$ that satisfies the no-flip property of Lemma 3 (equivalently, of Lemma 2). Then for any (X, Y) , the loss $\lambda \mapsto \ell(C_{\lambda}(X), Y)$ is non-increasing in λ .*

Proof. Fix $\lambda_1 \leq \lambda_2$. If $\ell(C_{\lambda_1}(X), Y) = 0$, then $C_{\lambda_1}(X) \in \{Y, \text{abstain}\}$; by the no-flip property, $C_{\lambda_2}(X) \in \{C_{\lambda_1}(X), \text{abstain}\} \subseteq \{Y, \text{abstain}\}$, so $\ell(C_{\lambda_2}(X), Y) = 0$. If $\ell(C_{\lambda_1}(X), Y) = 1$, then $\ell(C_{\lambda_2}(X), Y) \in \{0, 1\} \leq 1$. $\square$

C.3 CRC Validity for Score Averaging

For Score Averaging, the scoring function $f_{\text{avg}}(x) = \frac{1}{K} \sum_{j=1}^K f_j(x)$ is a deterministic function of the data-independent $\{f_j\}_{j=1}^K$. Standard CRC is applied to f_{avg} , giving $L_i(\lambda) = \ell(C_{\lambda}(X_i), Y_i)$ with C_{λ} from Eq. 1.

(P1) Lemma 2, applied with f_{avg} in place of f_j , gives the no-flip property for $\lambda \mapsto C_{\lambda}(X_i)$; Lemma 4 then yields monotonicity of L_i in λ .

(P2) At any fixed λ , the map $(X_i, Y_i) \mapsto L_i(\lambda)$ is a single deterministic function; Lemma 1 gives exchangeability.

Both conditions of Theorem 1 hold; the CRC guarantee (12) applies.

C.4 CRC Validity for Decision Voting

Decision Voting uses a single shared threshold λ and aggregates via $C_{\lambda}^{\text{vote}} = \text{MajVote}(C_{\lambda}^{(1)}, \dots, C_{\lambda}^{(K)})$, with $\{f_j\}$ data-independent. The loss is $L_i(\lambda) = \ell(C_{\lambda}^{\text{vote}}(X_i), Y_i)$.

(P1) Lemma 3 gives the no-flip property for $\lambda \mapsto C_{\lambda}^{\text{vote}}(X_i)$; Lemma 4 yields monotonicity of L_i in λ .

(P2) At any fixed λ , $C_{\lambda}^{\text{vote}}$ is a deterministic function of $\{f_j\}_{j=1}^K$, so $(X_i, Y_i) \mapsto L_i(\lambda)$ is a single deterministic function; Lemma 1 gives exchangeability.

Both conditions of Theorem 1 hold; the CRC guarantee (12) applies.

C.5 CRC Validity for MC³

MC³ uses Phase 1 to compute per-expert ratios $\{\lambda_j^{(0)}\}_{j=1}^K \geq 0$ (Eq. 8). Phase 2 introduces a global scalar $t \geq 0$ and defines $C_t(x) = \text{MajVote}(C_{\lambda_1(t)}^{(1)}(x), \dots, C_{\lambda_K(t)}^{(K)}(x))$ with $\lambda_j(t) = t \cdot \lambda_j^{(0)}$ (Eq. 9). The loss is $L_i(t) = \ell(C_t(X_i), Y_i)$.

Theorem 2 (CRC validity for MC³). *Under the exchangeability of $(X_i, Y_i)_{i=1}^{n+1}$, the MC³ threshold $\hat{t}$ defined in Eq. 11 satisfies $\mathbb{E}[\ell(C_{\hat{t}}(X_{n+1}), Y_{n+1})] \leq \alpha$.*

Proof. (P1) Each $\lambda_j(t) = t \cdot \lambda_j^{(0)}$ is non-decreasing in t (since $\lambda_j^{(0)} \geq 0$), so Lemma 2 applies to each expert's threshold trajectory. Lemma 3 then gives the no-flip property for $t \mapsto C_t(X_i)$; Lemma 4 yields monotonicity of $L_i(t)$ in t .

(P2) Phase 1 produces a vector of per-expert ratios $(\lambda_1^{(0)}, \dots, \lambda_K^{(0)})$ that fixes the shape of the decision function. At any fixed $t \geq 0$ and given these ratios, the joint decision function

$$C_t(x) = \text{MajVote}(C_{t\lambda_1^{(0)}}^{(1)}(x), \dots, C_{t\lambda_K^{(0)}}^{(K)}(x))$$

is a deterministic function of x , so the map $(X_i, Y_i) \mapsto L_i(t) = \ell(C_t(X_i), Y_i)$ is a single deter-

ministic function. Lemma 1 gives exchangeability of $\{L_i(t)\}_{i=1}^{n+1}$.

Conclusion. Both (P1) and (P2) hold; the boundedness, right-continuity, and $\lambda_{\max}$ conditions are addressed in §C.1. Applying Theorem 1 with $\hat{t}$ as in Eq. 11 yields $\mathbb{E}[L_{n+1}(\hat{t})] \leq \alpha$. $\square$

C.6 Remarks

Why MC³ preserves exchangeability while *Test-time Voting* does not. In MC³, the per-expert thresholds $\lambda_j(t) = t \cdot \lambda_j^{(0)}$ define a single joint decision function C_t that is *used identically in calibration and at test*. By contrast, the *Test-time Voting* baseline of §4.1 calibrates each expert independently at level α to obtain $\lambda_j^{(0)}$, and aggregates them through a majority vote only at test time. The mismatch between the calibration-time decision (per-expert independent) and the test-time decision (collective vote) breaks the symmetry required by (P2), and the CRC bound no longer applies.

Generality. The proofs above use only the structure of the per-expert decisions in Eq. 1 and the majority-vote aggregation. They do not depend on the specific choice of conformity score f_j (any score satisfying the structure of Eq. 1 suffices), nor on the dimension or number of experts K , nor on the dataset.