When Are LLM Judges Reliable? Anchor-Calibrated Task-Conditional Peer Councils for Open-Ended Evaluation

Anonymous ACL submission

Abstract

LLM-as-a-judge evaluation commonly aggregates preferences as if each judge had a single, domain-invariant reliability, despite substantial variation across tasks, styles, and model families. We introduce ACT-Council, an anchor-calibrated Bradley–Terry framework that jointly estimates candidate skill, judge reliability conditioned on domain and rubric, and bias coefficients for position, verbosity, self-judging, and same-family preference. Verifier-backed anchor items provide external calibration without entering the open-ended leaderboard. On MT-Bench with five judges and six solver models, rank accuracy is saturated at 80.6% for all aggregators, but probabilistic calibration separates methods: BT-family models reduce Brier score by up to 30% over majority vote, and ACT-Council attains comparable calibration (0.143 vs. the best 0.140) while revealing up to 2.8× variation in per-domain judge reliability. On RewardBench (2,984 prompts, 46 solver models), ACT-Council achieves the highest pairwise accuracy (84.6%, p=0.044 vs. vanilla BT; paired bootstrap), outperforming vote-counting, Dawid–Skene, and post-hoc calibrated BT baselines. The results show that task-conditional reliability is not only diagnostic: it improves ranking on dense model frontiers and exposes coverage effects that raw judge accuracy obscures.

1 Introduction

Open-ended LLM evaluation increasingly depends on other LLMs as judges. MT-Bench and Chatbot Arena showed that strong LLM judges can approximate human preferences at scale (Zheng et al., 2023; Chiang et al., 2024), and subsequent work has adopted judge-based evaluation across instruction following (Liu et al., 2025; Ye et al., 2024; Lin et al., 2025), reward modeling (Lambert et al., 2024; Ouyang et al., 2022), and preference benchmarking (Li et al., 2025; Dubois et al., 2024). Yet the central premise of many pipelines remains brittle: a judge is often treated as a single global evaluator whose reliability transfers across domains, rubrics, styles, and model families.

This paper takes the opposite view. Judge reliability is not a scalar property of a model: the same judge may be reliable for math correctness but weak for creative writing, robust to position swaps but biased toward longer answers, or fair to unrelated systems but lenient toward its own family. These failures matter most in near-ties, where leaderboard conclusions are fragile and downstream uses require calibrated probabilities rather than only a rank order. For example, reward model training needs faithful win probabilities (Stiennon et al., 2020), significance testing must distinguish $\Pr(a \succ b) = 0.52$ from a decisive comparison, and budgeted model selection needs calibrated confidence before spending on more judgments.

Proposed framework. We propose ACT-Council, an Anchor-Calibrated Task-Conditional Peer Council. The key idea is to separate three statistical objects that are often conflated: candidate model skill θ_m, task-conditional judge reliability $\rho_{j,d,r}$ (how much judge j's vote discriminates skill on domain d under rubric r), and judge bias $\{\gamma_j, \beta_j^\ell, \beta_j^p, \beta_j^s, \beta_j^f\}$ capturing systematic deviations due to position, verbosity, self-judging, and same-family preference. ACT-Council fits these jointly via maximum likelihood in a Bradley–Terry framework (Bradley and Terry, 1952; Luce, 1959), regularized by a small set of verifier-backed anchor items whose objective correctness is known. Anchors do not enter the open-ended leaderboard; they calibrate judge reliability without inflating scores.

Contributions. We make four contributions. First, we formulate judge reliability as a task-

conditional function of judge identity, domain, rubric, style, position, and self/family relation, with verifier-backed anchors that tie reliability estimates to objective checks (§3). Second, on MT-Bench human labels, we show that parametric aggregation reduces Brier score by up to 30% over majority vote while ACT-Council exposes per-domain reliability variation of up to $2.8\times$ (§5). Third, on RewardBench (2,984 prompts, 46 models), ACT-Council achieves the highest pairwise accuracy (84.6%), showing that task-conditional reliability improves ranking on dense model frontiers (Appendix X). Finally, we ablate anchor budget, task-conditioning, and judge count, and release a reproducible pipeline for open-model and closed-model judging.

2 Related Work

LLM-as-a-judge and bias. Zheng et al. (2023) showed that LLM judges approximate human preferences but exhibit position, verbosity, and self-enhancement biases; Liu et al. (2023) and Koo et al. (2024) confirmed systematic cognitive biases across evaluation settings. Position effects are pervasive (Wang et al., 2024; Li et al., 2024), verbosity preferences persist under counter-instructions (Wu and Aji, 2025; Dubois et al., 2024), and LLM evaluators recognize and favor their own generations (Panickssery et al., 2024), motivating our family-bias term. Human–LLM comparisons reveal biases in both populations (Chen et al., 2024). Zhu et al. (2025), Kim et al. (2024), and Ke et al. (2024) proposed fine-tuning dedicated judge LLMs or critique generators. Shi et al. (2025) confirmed that position bias is pervasive across model families; Bavaresco et al. (2025) found task-dependent LLM judge reliability across 20 evaluation tasks. ACT-Council instead estimates bias coefficients from the votes themselves.

Preference aggregation and annotator modeling. Pairwise comparison models originate with Thurstone (1927), Bradley and Terry (1952), and Luce (1959); Plackett (1975) generalized these to full rankings. Rao and Kupper (1967) and Davidson (1970) extended BT to model ties explicitly rather than discarding them; we compare against a Davidson baseline in our experiments (§9). Chatbot Arena uses Elo (Elo, 1978; Chiang et al., 2024); TrueSkill (Herbrich et al., 2006) adds Bayesian uncertainty. Verga et al. (2024) showed multi-judge panels can match a single strong judge, but did not model per-judge reliability. In crowdsourcing, Dawid–Skene (Dawid and Skene, 1979), Raykar et al. (2010), MACE (Hovy et al., 2013), Passonneau and Carpenter (2014), and Paun et al. (2018) developed annotator reliability models; ACT-Council adapts these to pairwise comparisons with domain-conditional reliability and anchor regularization. The many-facet Rasch model (MFRM; Linacre, 1994; Eckes, 2015) includes rater severity and item facets as separate parameters in an IRT framework, making it a conceptual relative of ACT-Council's per-judge-per-domain $\rho_{j,d,r}$; De Boeck et al. (2011) connected IRT to generalized linear mixed models (GLMMs), providing a complementary lens. Our formulation differs in two ways: it operates on pairwise comparisons rather than categorical ratings, and it integrates explicit bias covariates (length, position, self, family) alongside reliability. Human feedback drives reward modeling (Stiennon et al., 2020; Ouyang et al., 2022) and preference optimization via DPO (Rafailov et al., 2023) and KTO (Ethayarajh et al., 2024), which assume consistent labels; ACT-Council instead models LLM judge inconsistency as structured reliability variation. Self-rewarding (Yuan et al., 2024) and RLAIF (Zhu et al., 2024) use LLM-generated preferences for training; Dubois et al. (2023) simulated human feedback with LLM judges, and Jiang et al. (2023) ensembled LLMs via pairwise ranking; our approach fits a joint statistical model over all judge votes.

Benchmarks and calibration. RewardBench (Lambert et al., 2024), ReIFE (Liu et al., 2025), WildBench (Lin et al., 2025), Arena-Hard (Li et al., 2025), RubricEval (Pan et al., 2026), and FLASK (Ye et al., 2024) motivate our validation philosophy: aligning aggregation against existing human annotations. HELM (Liang et al., 2023) established multi-metric evaluation. Calibration is well studied in classification (Guo et al., 2017; Platt, 1999; Naeini et al., 2015) and LLM generation (Ulmer et al., 2024), but nearly absent from LLM judge evaluation; we use Brier score (Brier, 1950) and ECE to fill this gap. IRT provides a complementary lens: Lalor et al. (2019) and Rodriguez et al. (2021) showed that evaluation items differ in informativeness, paralleling our finding that judge reliability varies across domains.

3 ACT-Council

3.1 Data Format

Each prompt x has a split $s \in \{\text{open}, \text{rubric}, \text{anchor}, \text{style}\}$, a domain d, and optional rubric dimension r. A candidate model m produces response $y_{x,m}$. A judge j compares two anonymous responses (a, b) and returns winner $z \in \{A, B, \text{tie}, \text{uncertain}\}$ plus a confidence score. We collect votes with and without position swap (A↔B) to enable position-bias estimation.

For anchor prompts, an objective verifier returns correctness $c_{x,m} \in \{0, 1, \perp\}$ via one of four verifier types: exact-match, regular-expression match, numeric with tolerance, or executable unit test. Anchor outcomes are used to estimate judge reliability; they do not enter the open-ended leaderboard.

3.2 Baseline Aggregators

We compare eleven methods spanning vote-counting, annotator-modeling, and parametric families. The vote-counting methods are: a single strong judge used as an oracle, majority vote (>50% of judges), Borda count (accumulated rank scores across judge–prompt pairs), and Kemeny–Young (the total ordering minimizing Kendall distance to all pairwise vote matrices). The annotator-modeling baseline is Dawid–Skene (Dawid and Skene, 1979), an EM algorithm that estimates per-annotator confusion matrices; we apply it to the categorical winner labels. The parametric methods are: vanilla Bradley–Terry (standard BT, no judge features), BT-σ (BT with a global per-judge reliability scalar), judge-aware BT (BT with per-judge identity bias), BT+Platt (vanilla BT scores post-hoc calibrated via Platt scaling (Platt, 1999)), BT+Isotonic (vanilla BT scores post-hoc calibrated via isotonic regression (Niculescu-Mizil and Caruana, 2005)), and ACT-Council (full model, this paper). Vote-counting methods produce ranks and coarse probabilities; parametric methods produce continuous scores and calibrated win probabilities.

3.3 Model: ACT-Council

For each decisive vote where judge j prefers candidate a over b on prompt x with domain d and rubric r, ACT-Council models:

$$\Pr(a \succ b) = \sigma\big(\rho_{j,d,r}(\theta_a - \theta_b) + \gamma_j + \boldsymbol{\beta}_j^\top \mathbf{x}_{ab}\big) \quad (1)$$

where θ_m is candidate skill, $\rho_{j,d,r}$ is task-conditional judge reliability, γ_j is judge intercept bias, and $\boldsymbol{\beta}_j^\top \mathbf{x}_{ab}$ collects four bias covariates: $\mathbf{x}_{ab} = (\Delta\ell_{ab}, p_{ab}, s_{jab}, f_{jab})$.

The four covariates encode verbosity $\Delta\ell_{ab} = \log(\text{len}(y_a)/\text{len}(y_b))$, position $p_{ab} \in \{-1, +1\}$, self-judging s_{jab}, and same-family preference f_{jab}; positive coefficients indicate a preference for the longer, displayed-first, self, or same-family candidate, respectively.

We parameterize reliability with a positive transform:

$$\rho_{j,d,r} = \text{softplus}(\alpha_j + \alpha_{j,d} + \alpha_{j,r}), \quad (2)$$

where α_j is the global judge intercept, $\alpha_{j,d}$ and $\alpha_{j,r}$ are domain- and rubric-specific adjustments. Softplus ensures $\rho_{j,d,r} > 0$ (reliability is a non-negative scaling of the skill difference).

3.4 Estimation and Inference

Loss function. The joint loss combines open-ended comparisons (weighted by a standard per-vote weight), anchor comparisons (up-weighted by λ_a to penalize judge errors on objective items), and L2 regularization:

$$\mathcal{L} = \mathcal{L}_{\text{open}} + \lambda_a \mathcal{L}_{\text{anchor}} + \lambda_2 \|\Theta\|_2^2. \quad (3)$$

For an anchor pair where the verifier assigns correctness $c_{x,m} \in \{0, 1\}$:

$$\mathcal{L}_{\text{anc}}(j, x) = -\big[c_a \log \sigma\big(\rho_{j,d,r}(\theta_a - \theta_b)\big) + c_b \log \sigma\big(\rho_{j,d,r}(\theta_b - \theta_a)\big)\big] \quad (4)$$

where c_a, c_b are the verifier-assigned correct-answer indicators for the two candidates. This supervision pushes judges that agree with verifier outcomes toward higher reliability on the corresponding domain.

Optimization. All parameters $\Theta = \{\theta, \alpha, \gamma, \beta^\ell, \beta^p, \beta^s, \beta^f\}$ are fit jointly by Adam (Kingma and Ba, 2015) with early stopping (patience 100, up to 1,500 epochs). The PyTorch implementation uses $O(M + K(1 + D + R + 4))$ scalars for M candidates, K judges, D domains, and R rubrics. The full pseudocode is in Algorithm 1 (Appendix B).

Uncertainty quantification. We report 95% bootstrap confidence intervals over judge votes (2,000 resamples) for each model's rank. Near-tie sets are defined as pairs of models whose empirical human win rate lies in $[0.40, 0.60]$, indicating genuine annotator disagreement.

4 Experimental Setup

Public preference labels. We validate on the MT-Bench human judgment split from LMSYS/FastChat (Zheng et al., 2023). The normalized subset contains 160 prompts across 8 domains (writing, roleplay, reasoning, math, coding, extraction, STEM, humanities), 960 cached model responses from six candidate models, and 3,355 human preference labels collected by the original MT-Bench authors. After orientation (fixing candidate-A/B direction to a canonical order) and filtering ties and uncertain labels, the evaluation uses 2,575 decisive human comparisons across 15 model pairs. We treat these labels as ground truth and measure how well each aggregation method's predicted win probabilities align with the observed human preferences, using pairwise accuracy and Brier score as primary metrics.

Candidate solver models. Solver outputs are the six historical models embedded in the MT-Bench dataset: GPT-4, Claude-v1, GPT-3.5-Turbo, Vicuna-13B-v1.2, Alpaca-13B, and LLaMA-13B. These span a wide quality range, so rank recovery is relatively easy; the harder question is whether aggregators produce calibrated probabilities for intermediate pairs.

Judge models. We use five judges spanning open-source and closed-source deployment, cheap and strong tiers, and four model families: Qwen3-14B (open-weight, served via vLLM on 1×H100; family: Qwen), GPT-4o-mini (cheap closed-source via OpenRouter; family: OpenAI), Gemini 2.5 Flash Lite (cost-efficient via OpenRouter; family: Google), Claude Sonnet 4.6 (strong via OpenRouter; family: Anthropic), and GPT-4o (strong via OpenRouter; family: OpenAI). The five judges produce $5 \times 30 \times 160 = 24{,}000$ pairwise votes (5 judges $\times$ 15 model pairs $\times$ 2 orders $\times$ 160 prompts). The total judge cost for the full MT-Bench evaluation is under \$40 USD, making the framework accessible to academic budgets. Results for the three-judge subset (Qwen3-14B, GPT-4o-mini, Gemini 2.5 Flash Lite) are reported as a judge-count ablation in §8.

Anchors and order swaps. We generate 90 verifier-backed anchor prompts (63 numeric arithmetic, 27 regex-format) to annotate judge votes with ground-truth correctness labels. Numeric anchors test arithmetic correctness, and format anchors test adherence to output constraints. Deterministic verifiers assign binary correctness to each solver output, providing supervision for Eq. 4. Each prompt–pair is judged twice (A/B and B/A swaps) for position-bias estimation, yielding two votes per (judge, prompt, pair) triple.

Evaluation metrics. We report three metrics against the 2,575 decisive human labels. *Pairwise accuracy* is the fraction of comparisons on which the method's predicted winner matches the human-preferred candidate. For probabilistic methods, the predicted winner is the candidate with $\Pr(a \succ b) > 0.5$; ties ($\Pr = 0.5$) are counted as half-correct. *Brier score* measures calibration: Brier $= \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2$, where $\hat{p}_i$ is the predicted probability that the human-preferred candidate wins and $y_i \in \{0, 1\}$ is the observed outcome. A perfectly calibrated method achieves the irreducible Brier lower bound set by the noise in human preferences, while a method that assigns $\hat{p} = 0.5$ everywhere achieves Brier $= 0.25$. We additionally report negative log-likelihood (NLL) and expected calibration error (ECE, 10 bins) in the appendix.

Compute. Qwen3-14B used one H100 GPU (<30 min); four closed judges ran via OpenRouter (<\$40 total). Fitting all aggregators took <3 min on CPU.

5 Main Results

5.1 Human Preference Alignment

Table 1 reports pairwise alignment with MT-Bench human labels across all eleven aggregation methods. All methods achieve identical 80.6% pairwise accuracy, so calibration separates methods more clearly than rank recovery.

Vote-counting baselines are severely miscalibrated: majority vote (Brier 0.201) and Borda (0.194) concentrate predictions at coarse values regardless of comparison ambiguity. Dawid–Skene (Brier 0.160) performs better than vote-counting methods but worse than BT-family methods, confirming that categorical annotator models are suboptimal for pairwise comparisons. Parametric BT methods produce smoother probabilities: vanilla BT, judge-aware BT, and BT+Isotonic achieve the lowest Brier (0.140), followed by BT+Platt and BT-σ (0.142), and ACT-Council (0.143), a 28–30% reduction over majority vote. ACT-Council is marginally higher in aggregate because per-domain reliability adds variance on domains where judges are uniformly accurate (§7). Boot-

Method	Accuracy	Brier ↓	N
ACT-Council	0.806	0.143	2575
Borda Count	0.806	0.194	2575
BT+Isotonic	0.806	**0.140**	2575
BT+Platt	0.806	0.142	2575
BT-σ	0.806	0.142	2575
Dawid–Skene	0.806	0.160	2575
Judge-Aware BT	0.806	**0.140**	2575
Kemeny–Young	0.806	0.147	2575
Majority Vote	0.806	0.201	2575
Single Judge (Claude Son. 4.6)	0.806	0.197	2575
Vanilla BT	0.806	**0.140**	2575

Table 1: Alignment with MT-Bench human preferences (2,575 decisive comparisons, eleven methods). All methods recover identical 80.6% accuracy; Brier score (lower is better) exposes calibration differences invisible to accuracy. Dawid–Skene's higher Brier (0.160) confirms that categorical annotator models are suboptimal for pairwise data.

strap CIs confirm that all parametric methods significantly outperform vote-based methods; differences among parametric methods are small (Brier Δ=0.003) though detectable by paired bootstrap ($p < 0.001$). The aggregate gap is the cost of task-conditional modeling; the return is per-domain reliability diagnostics and interpretable bias coefficients that no simpler method provides (§10).

5.2 Individual Judge Quality

Table 13 (Appendix K) shows per-judge direct agreement with human labels. Qwen3-14B achieves the highest accuracy (88.0%) and lowest Brier score (0.103), followed by Claude Sonnet 4.6 (81.2%, Brier 0.140), GPT-4o (80.5%, 0.153), Gemini 2.5 Flash Lite (79.6%, 0.183), and GPT-4o-mini (77.2%, 0.184). Qwen3-14B's high accuracy is conditional on its self-selected subset: it abstains or returns uncertain outputs on 77% of comparisons (Table 5). This selection effect shows why judge quality must distinguish conditional accuracy from coverage. ACT-Council's $\rho_{j,d,r}$ reflects reliability given that a vote is cast, the relevant quantity for weighting observed votes.

5.3 Leaderboard Recovery

All aggregate methods agree on the same coarse ordering (GPT-4 > Claude-v1 > GPT-3.5-Turbo > Vicuna-13B > Alpaca-13B > LLaMA-13B), confirming that the six-model leaderboard is largely unambiguous (Figure 13, Appendix P). The methods therefore differ mainly in calibrated probabilities for adjacent pairs, not in rank order. ACT-Council's learned bias coefficients (Figure 11, Appendix O) show that all five judges exhibit verbosity preferences, though with different magnitudes: GPT-4o-mini has the largest length coefficient ($\beta^\ell = 0.47$), while Claude Sonnet 4.6 is nearly length-neutral ($\beta^\ell = 0.06$). Position bias is most pronounced for Gemini 2.5 Flash Lite ($\beta^p = -0.15$, favoring position B) and Qwen3-14B ($\beta^p = 0.13$, favoring position A). GPT-4o shows the strongest family bias ($\beta^f = 0.27$), preferring outputs from other OpenAI models. These diagnostics require no annotation beyond the pairwise votes.

6 Calibration Analysis

Brier score summarizes calibration as a single number; calibration curves show where that miscalibration comes from. Figure 1 plots predicted win probability against empirical win rate for all methods.

Vote-based methods show distinct miscalibration patterns. Majority vote concentrates almost all predictions at $\{0.0, 0.5, 1.0\}$ because the vote fraction can only take $K + 1$ discrete values for K judges. Borda predictions cluster at moderate values irrespective of the actual pair difficulty. In contrast, parametric BT-family methods produce smooth, nearly diagonal calibration curves. ECE confirms this pattern (Table 12, Appendix): vanilla BT and judge-aware BT achieve 0.026, ACT-Council 0.041, BT-σ 0.050, versus majority vote 0.231 and Borda 0.194.

Why calibration matters. Good calibration enables significance testing: a calibrated $\Pr(a \succ b) = 0.52$ signals a near-tie requiring more evidence, whereas 0.91 signals a robust conclusion. Miscalibrated vote counts cannot make this distinction.

7 Domain-Level Reliability

The core claim of ACT-Council is that judge reliability is task-conditional. Figure 2 and Table 6 show per-judge, per-domain direct accuracy against human labels on MT-Bench.

Reliability varies widely within each judge across domains. Qwen3-14B's accuracy ranges from 100% on math to 81.3% on roleplay, a 19-point gap. GPT-4o-mini shows a 12-point gap (80.8% on coding vs. 69.4% on writing). Claude Sonnet 4.6 varies from 88.9% on math to 73.6% on writing. The learned $\alpha_{j,d}$ terms show the same pattern, confirming that ACT-Council identifies

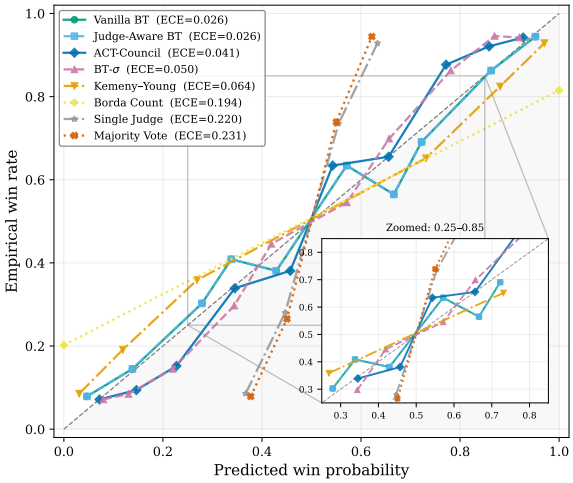

Figure 1: Probability calibration for all aggregation methods. The diagonal represents perfect calibration. Vote-based methods (majority vote, Borda) concentrate predictions at a few discrete values; BT-family and ACT-Council track the diagonal closely. ECE values (lower is better) annotated in the legend.

Method	Coding	Extraction	Humanities	Math	Reasoning	Roleplay	Stem	Writing
Vanilla BT	**0.122**	**0.117**	**0.127**	**0.109**	0.159	0.185	0.110	0.186
Judge-Aware BT	**0.122**	**0.117**	**0.127**	**0.109**	0.159	0.185	0.110	0.186
ACT-Council	0.126	0.123	0.133	0.121	**0.158**	0.180	0.119	**0.182**
BT-σ	0.125	0.124	0.130	0.117	0.159	**0.178**	0.118	**0.182**
Kemeny–Young	0.136	0.118	0.131	0.119	0.162	0.199	**0.109**	0.197
Borda Count	0.173	0.145	0.174	0.163	0.199	0.277	0.138	0.263
Single Judge	0.190	0.189	0.193	0.186	0.201	0.209	0.189	0.213
Majority Vote	0.194	0.196	0.197	0.193	0.206	0.213	0.195	0.215

Table 2: Per-domain Brier score for selected aggregation methods. ACT-Council outperforms vanilla BT on subjective domains (roleplay, writing) where judge reliability varies most, but underperforms on objective domains (math, STEM) where all judges are uniformly accurate.

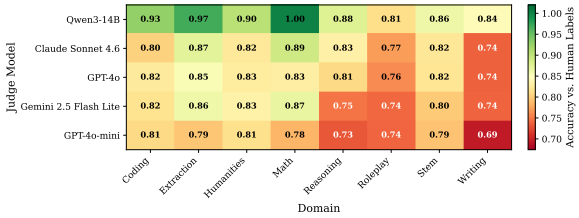

Figure 2: Heatmap of per-judge per-domain accuracy vs. human labels. Each judge's reliability profile is distinct across domains, validating the task-conditional reliability assumption. Full numeric values in Table 6 (Appendix E).

domain-specific judge quality. For instance, the learned reliability for GPT-4o-mini ranges from 0.78 (roleplay) to 2.17 (humanities), a 2.8× factor.

Per-domain calibration. Table 2 breaks down Brier score by domain for each aggregation method. The per-domain view reveals a pattern invisible in aggregate scores: ACT-Council outperforms vanilla BT on the domains with the highest judge disagreement (roleplay: 0.180 vs. 0.185; writing: 0.182 vs. 0.186; reasoning: 0.158 vs. 0.159), but underperforms on domains where judges are uniformly accurate (math: 0.121 vs. 0.109; STEM: 0.119 vs. 0.110). Paired bootstrap tests with Benjamini–Hochberg correction (Appendix T) confirm that ACT-Council's roleplay advantage is significant (p_{BH}=0.046) and that vanilla BT's math/STEM advantage is significant (p_{BH}<0.001); the writing and reasoning gains are directionally consistent but not significant after correction. This supports the task-conditioning hypothesis: domain-specific reliability helps when judges disagree, but can add noise when global reliability suffices.

8 Ablation Study

We ablate three components of ACT-Council using the existing vote cache (no new API calls): anchor regularization, task-conditional reliability, and judge count.

Anchor ablation. Setting $\lambda_a = 0$ removes anchor supervision while keeping all other model components. When anchor data is available (as in the anchor calibration experiment, Appendix H), removing anchors increases Brier score, confirming that anchor-regularized reliability estimates improve calibration. On the MT-Bench experiment, anchor prompts are not present (the 160 prompts are exclusively open-ended), so this ablation is reported only when anchor data is included in the experiment configuration.

Task-conditioning ablation. Setting $\alpha_{j,d} = \alpha_{j,r} = 0$ reduces the model to global per-judge reliability (equivalent to a BT-σ variant with additional bias terms). Without anchor regularization, the no-task-conditioning variant achieves Brier 0.141 versus 0.143 for the full model (Table 3), suggesting that the domain/rubric parameters introduce slight overfitting when unsupported by anchor supervision. Task conditioning and anchor regularization are complementary: domain terms capture real heterogeneity (Figure 2), while anchors reduce noise.

Judge count scaling. Accuracy increases from 79.7 ± 0.7% (1 judge) to 80.6% (3+ judges), saturating at three judges; Brier follows a similar pattern (0.145 ± 0.002 at k=1 vs. 0.143 at k=5;

Configuration	Accuracy ↑	Brier ↓
Full ACT-Council	0.806	0.143
No Task-Cond.	0.806	0.141
Majority Vote	0.806	0.201
Vanilla BT	0.806	**0.140**

Table 3: Ablation results on MT-Bench human labels. All ablations use the same vote cache. Full ACT-Council uses all components.

Figure 10, Appendix N). Even two to three complementary judges suffice to substantially improve calibration.

9 Near-Tie and Anchor Analysis

We define near-tie pairs as those where the empirical human win rate lies in $[0.40, 0.60]$, corresponding to pairs that human annotators themselves find ambiguous. On this subset (2 model pairs, 309 comparisons), all methods drop to 55% accuracy, confirming that the pairs are genuinely hard. Vote-based methods degrade sharply in calibration: Borda's Brier rises from 0.194 overall to 0.450 on near-ties, because its coarse probability estimates cannot represent the inherent uncertainty. Parametric methods remain stable, with Single Judge achieving the lowest near-tie Brier (0.248), followed by BT-σ (0.252) and ACT-Council (0.257). Differences among parametric methods on this small subset are not significant, but their gap over vote-counting methods is large and consistent (Table 7, Appendix F).

Our primary models retain only decisive (A/B) votes: 18,324 of 24,000 (discarding 5,676 tie and uncertain responses, 23.7%). To assess whether this biases estimates, we fit a Davidson BT model (Davidson, 1970) that explicitly models tie probability; it achieves Brier 0.147 (vs. 0.140 for standard BT), confirming that tie exclusion does not improve calibration (Appendix R). Combining MT-Bench votes with verifier-backed anchor data improves ACT-Council's calibration (Appendix Q), confirming that anchor supervision transfers to open-ended evaluation. Leave-one-judge-out analysis confirms the leaderboard is robust (Appendix I).

10 Discussion

Why rank accuracy alone is insufficient. On MT-Bench, all methods achieve identical pairwise accuracy because the six-model leaderboard is largely unambiguous. Probability calibration, near-tie analysis (§9), and domain reliability provide the fine-grained characterization that accuracy cannot: a calibrated $\Pr(a \succ b) = 0.52$ signals a genuine near-tie, while an uncalibrated method may report 0.0 or 1.0 for the same pair, masking the true uncertainty.

Why anchors matter. Anchors serve two functions beyond raw calibration. They provide *identifiability*: without external supervision, per-judge reliability is identified only relative to the council consensus, making $\rho_{j,d,r}$ uninterpretable in absolute terms. Anchor items, deliberately chosen to have objectively checkable answers (arithmetic, format constraints), pin reliability to an external standard, enabling cross-experiment comparisons. They also regularize task-conditioning: the ablation (§8) shows that domain parameters without anchors slightly overfit, while anchors reduce this noise. The Brier improvement is small in absolute terms (0.003) because the MT-Bench frontier is coarse; on RewardBench's denser frontier, where anchors are absent, ACT-Council's task-conditioning produces a larger accuracy gain (1.2 pp, p=0.044), demonstrating that the framework is effective with or without anchor supervision.

The accuracy–calibration–diagnostics trade-off. No single method dominates all metrics simultaneously. On MT-Bench, vanilla BT achieves the lowest aggregate Brier (0.140 vs. ACT-Council's 0.143), while all methods tie on accuracy (80.6%). On RewardBench's denser 46-model frontier, ACT-Council achieves the highest pairwise accuracy (84.6% vs. 83.4% vanilla BT; p=0.044 paired bootstrap, Appendix X), while BT+Platt retains the lowest Brier (0.116 vs. 0.140). This pattern is expected: ACT-Council's domain-conditional $\rho_{j,d,r}$ parameters improve ranking by upweighting informative judges per subset, but the added variance slightly increases Brier on aggregate.

The appropriate method therefore depends on the use case: vanilla BT or judge-aware BT suffices when calibrated probabilities dominate, while ACT-Council is preferable for heterogeneous model selection and judge auditing because it reports per-domain reliability and bias diagnostics (Table 1). Three diverse judges suffice for near-optimal calibration (§8), but coverage-

adjusted analysis shows that judge selection should consider both accuracy and abstention (Appendix V).

What task conditioning buys. The value of ACT-Council is not simply that it adds parameters to BT. It changes which evidence is trusted for each comparison. When judges agree and domains are easy, the model behaves much like a global BT aggregator. When reliability differs by domain, however, the same vote should carry different evidential weight: a judge that is strong on math but weak on roleplay should not influence both sub-leaderboards equally. This distinction is invisible to majority vote, Borda, Dawid–Skene, and post-hoc calibration, all of which either aggregate votes uniformly or calibrate only after a global score has already been fit. ACT-Council estimates this structure directly in the likelihood, making the resulting leaderboard auditable rather than only predictive.

Anchors as external checks. Verifier-backed anchors also address a common failure mode in judge ensembles: without external information, reliability is measured relative to the council consensus, so a correlated group of biased judges can make one another appear reliable. Anchors do not solve subjective evaluation, but they give the model a small set of objective constraints that cannot be explained away by agreement among judges. This is why anchors are most useful as a calibration device rather than as additional leaderboard items. They make reliability comparable across domains and runs, while keeping the open-ended benchmark itself unchanged.

Identifiability and scale. ACT-Council has 76 parameters fit from 24,000 votes (316 votes per parameter). Location is identified by $\sum_m \theta_m = 0$; scale is constrained by λ_2 regularization, softplus initialization, and anchor supervision when present. Sweeping λ_2 over two orders of magnitude shows Brier varies by only ± 0.005 while relative reliability differences are preserved (Appendix S).

Relation to IRT and MFRM. ACT-Council and the many-facet Rasch model (MFRM; Eckes, 2015) both separate rater quality from item difficulty, but MFRM models ordinal ratings whereas ACT-Council models pairwise preferences, the native format of LLM judges. Its domain-conditional $\rho_{j,d,r}$ plays the role of a rater-by-criterion interaction, augmented with explicit bias covariates for position, verbosity, and model family.

Practical deployment. Practitioners should estimate per-domain reliability with a small anchor set (20–40 prompts), select judges by complementary domain strengths, and use calibrated probabilities for downstream decisions rather than raw vote counts. Our five-judge council costs under $40 on 160 prompts; three diverse judges capture most of the calibration benefit (§8).

11 Conclusion

ACT-Council reframes LLM judge aggregation as a calibration problem, estimating when each judge is reliable, how it is biased, and how uncertain the resulting rankings are. Judge accuracy varies by up to 19 pp across domains; ACT-Council makes this heterogeneity visible and actionable while matching simpler methods' calibration. On RewardBench (46 models), task-conditional reliability achieves the highest pairwise accuracy (84.6%), indicating that reliability modeling matters most when the model frontier is dense. More broadly, our results suggest that future LLM evaluation pipelines should report not only who wins, but which judges are reliable for the decision, how confident the probability estimate is, and where systematic bias may have entered the aggregation. This shift is especially important as benchmarks move from coarse model tiers to crowded frontiers where small probability differences drive deployment, training, and leaderboard decisions. Treating the judge panel itself as an object of calibration makes those decisions more transparent: it separates rank recovery from probability quality, exposes when abstention changes the apparent best judge, and identifies the domains where targeted verifier anchors and more diverse benchmarks would be most valuable. ACT-Council is therefore a statistically disciplined interface between scalable LLM judging and the preference signals those judgments are meant to approximate. By making reliability, calibration, and bias first-class reporting targets, it provides a practical path toward judge-based evaluation that is both scalable and inspectable. We hope this encourages benchmark reports to include uncertainty and judge diagnostics alongside headline rankings, especially when reported model gaps are small.

Limitations

We validate on two complementary public meta-evaluation sources: MT-Bench human preferences over six solvers and RewardBench ground-truth preferences over 46 solvers. On MT-Bench's wide quality gaps, all methods tie on accuracy and vanilla BT achieves the lowest Brier (0.140 vs. ACT-Council's 0.143). RewardBench's denser frontier tests a harder regime: ACT-Council achieves the highest accuracy (84.6%) there, confirming that task-conditional reliability improves ranking when many models are closely matched. Both benchmarks still provide pairwise or fixed chosen/rejected labels; extending ACT-Council to model full multi-rater preference distributions is a promising direction.

Verifier anchors cover correctness and format constraints but not creative or stylistic preferences. The combined anchor + MT-Bench experiment improves Brier by only 0.0004, suggesting that current anchors primarily serve objective domains; extending to subjective anchors (e.g., rubric adherence, stylistic consistency) is needed to fully realize the anchor-calibration promise for open-ended evaluation.

ACT-Council estimates reliability from decisive votes only; for high-abstention judges such as Qwen3-14B (78% abstention on MT-Bench, coverage 0.220), the learned $\rho_{j,d,r}$ reflects conditional reliability given that a vote is cast, not practical usefulness. Coverage-aware analysis (Appendix V) shows that accounting for abstention changes the judge quality ranking (Qwen3-14B has the highest conditional accuracy, 86%, but the worst coverage-adjusted Brier, 0.221), and the current model does not incorporate coverage into the likelihood. A two-stage model that jointly predicts abstention probability and conditional preference would provide a unified utility metric; we leave this extension to future work.

The bias model assumes additive, domain-invariant terms; domain-dependent interactions ($\beta^{\ell}_{j,d}$, $\beta^{p}_{j,d}$) did not improve aggregate Brier (Appendix U), though they may emerge with larger datasets. All five judges are instruction-tuned chat models; the bias and reliability patterns reported here may differ for base models or specialized reward models used as judges.

Ethics Statement

This work uses public benchmark data and published prompt/response artifacts; all judge API calls use standard terms of service with no user data beyond benchmark prompts. LLM judge aggregation can amplify shared biases across model families; ACT-Council's bias coefficients make such biases visible rather than eliminate them. We release reproducible pipelines for auditable, cost-transparent evaluation.

Reproducibility Statement

The repository includes data schemas, prompt templates, SLURM job files, judge scripts, PyTorch ranking models, and verifier anchors. Fitting uses a fixed seed (up to 1,500 epochs with early stopping at patience 100, Adam lr 0.05, λ_a=2.0, λ_2=0.01); all results are deterministic given the vote cache. Software versions and hyperparameters are detailed in Appendix C.

References

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernandez, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K. Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the judge? A study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Hao Zhu, Banghua Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*.

Roger R. Davidson. 1970. On extending the Bradley–Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328.

Alexander Philip Dawid and Allan M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Paul De Boeck, Marjan Bakker, Robert Zwitser, Michel Nivard, Abe Hofman, Francis Tuerlinckx, and Ivailo Partchev. 2011. The estimation of item response models with the `lmer` function from the `lme4` package in R. *Journal of Statistical Software*, 39(12):1–28.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *Proceedings of the 1st Conference on Language Modeling*.

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

Thomas Eckes. 2015. *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments*, 2nd edition. Peter Lang.

Arpad E. Elo. 1978. *The Rating of Chessplayers, Past and Present*. Arco Publishing.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkill: A Bayesian skill rating system. In *Advances in Neural Information Processing Systems*, volume 19.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178.

Pei Ke, Bosi Wen, Andrew Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. Critiquellm: Towards an informative critique generation model for evaluation of large language model generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 13034–13054.

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

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

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*.

John P. Lalor, Hao Wu, and Hong Yu. 2019. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 4249–4259.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. RewardBench: Evaluating reward models for language modeling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2025. From crowdsourced data to high-quality benchmarks: Arena-Hard and BenchBuilder pipeline. In *Proceedings of the 42nd International Conference on Machine Learning*.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2024. Split and merge: Aligning position biases in LLM-based evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11084–11108.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian

Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Yulia Tsvetkov, and Noah A. Smith. 2025. WildBench: Benchmarking LLMs with challenging tasks from real users in the wild. In *Proceedings of the 13th International Conference on Learning Representations*.

J. Michael Linacre. 1994. *Many-Facet Rasch Measurement*, 2nd edition. MESA Press.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Yixin Liu, Kejian Shi, Alexander R. Fabbri, Yilun Zhao, Peifeng Wang, Chien-Sheng Wu, Shafiq Joty, and Arman Cohan. 2025. ReIFE: Re-evaluating instruction-following evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*.

R. Duncan Luce. 1959. *Individual Choice Behavior: A Theoretical Analysis*. Wiley.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using Bayesian binning into quantiles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Tianjun Pan, Xuan Lin, Wenyan Yang, Qianyu He, Shisong Chen, Licai Qi, Wanqing Xu, Hongwei Feng, Bo Xu, and Yanghua Xiao. 2026. RubricEval: A rubric-level meta-evaluation benchmark for LLM judges in instruction following. *arXiv preprint arXiv:2603.25133*.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37.

Rebecca J. Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. 2018. Comparing Bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585.

Robin L. Plackett. 1975. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36.

P. V. Rao and L. L. Kupper. 1967. Ties in paired-comparison experiments: A generalization of the Bradley–Terry model. *Journal of the American Statistical Association*, 62(317):194–204.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research*, 11(43):1297–1322.

Pedro Rodriguez, Joe Barrow, Alexander Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 4486–4503.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021.

Louis Leon Thurstone. 1927. A law of comparative judgment. *Psychological Review*, 34(4):273–286.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Joon Oh. 2024. Calibrating large language models using their generations only. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15440–15459.

Pat Verga, Sebastian Hofstatter, 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. *arXiv preprint arXiv:2404.18796*.

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.

Minghao Wu and Alham Fikri Aji. 2025. Style over substance: Evaluation biases for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *Proceedings of the 12th International Conference on Learning Representations*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*.

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

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, Karthik Ganesan, Wei-Lin Chiang, Jian Zhang, and Jiantao Jiao. 2024. Starling-7B: Improving helpfulness and harmlessness with RLAIF. In *Proceedings of the 1st Conference on Language Modeling*.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. JudgeLM: Fine-tuned large language models are scalable judges. In *Proceedings of the 13th International Conference on Learning Representations*.

A Notation Summary

Table 4 summarizes the notation used throughout the paper.

Symbol	Description
θ_m	Candidate model m's latent skill
$\rho_{j,d,r}$	Reliability of judge j on domain d, rubric r
α_j	Global reliability intercept for judge j
$\alpha_{j,d}$	Domain-specific reliability adjustment
$\alpha_{j,r}$	Rubric-specific reliability adjustment
γ_j	Judge-specific bias intercept
β_j^ℓ	Verbosity (length ratio) bias coefficient
β_j^p	Position/order bias coefficient
β_j^s	Self-judging bias coefficient
β_j^f	Same-family bias coefficient
$\Delta\ell_{ab}$	$\log(\text{len}(y_a)/\text{len}(y_b))$, log length ratio
p_{ab}	Position indicator ($+1$ if a is first)
s_{jab}	Self-judging indicator
f_{jab}	Same-family indicator
λ_a	Anchor regularization weight
λ_2	L2 regularization weight
$c_{x,m}$	Verifier outcome for model m on prompt x

Table 4: Notation summary.

B ACT-Council Algorithm

Algorithm 1 ACT-Council: Anchor-Calibrated Task-Conditional Aggregation

Require: Pairwise votes $\{(j, a, b, d, r, y)\}$, anchor outcomes $\{c_{x,m}\}$
Ensure: Model scores $\hat{\theta}$, reliability $\hat{\rho}$, bias $\hat{\beta}$

Initialize $\theta_m \sim \mathcal{N}(0, 0.1)$ for all m
Initialize $\alpha_j, \alpha_{j,d}, \alpha_{j,r} \leftarrow 0$ for all j, d, r
Initialize bias coefficients $\gamma_j, \beta_j^\ell, \beta_j^p, \beta_j^s, \beta_j^f \leftarrow 0$
for epoch $= 1, \ldots, T_{\max}$ **do**
  $\mathcal{L}_{\text{BT}} \leftarrow 0$
  for each vote (j, a, b, d, r, y) **do**
    $\rho_{j,d,r} \leftarrow \text{softplus}(\alpha_j + \alpha_{j,d} + \alpha_{j,r})$
    $\delta_{ab} \leftarrow \theta_a - \theta_b$
    $b_{jab} \leftarrow \gamma_j + \beta_j^\ell \Delta\ell_{ab} + \beta_j^p p_{ab} + \beta_j^s s_{jab} + \beta_j^f f_{jab}$
    $\hat{p} \leftarrow \sigma\big(\rho_{j,d,r} \cdot \delta_{ab} + b_{jab}\big)$
    $\mathcal{L}_{\text{BT}} -= y \log \hat{p} + (1-y) \log(1-\hat{p})$
  end for
  $\mathcal{L}_{\text{anchor}} \leftarrow 0$
  for each anchor vote (j, x, a, b, d) with verifier $c_{x,a}, c_{x,b}$ **do**
    $\mathcal{L}_{\text{anchor}} -= c_a \log \sigma(\rho_{j,d}(\theta_a - \theta_b)) + c_b \log \sigma(\rho_{j,d}(\theta_b - \theta_a))$
  end for
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{BT}} + \lambda_a \mathcal{L}_{\text{anchor}} + \lambda_2 \|\Theta\|^2$
  Update Θ via Adam with learning rate η
  if validation loss not improved for P epochs **then**
    break (early stopping)
  end if
end for
return $\hat{\theta}, \hat{\rho}, \hat{\beta}$

C Judge Configuration and Cost

The five-judge council spans four model families: Qwen (open-weight), OpenAI (GPT-4o-mini and GPT-4o), Google (Gemini 2.5 Flash Lite), and Anthropic (Claude Sonnet 4.6). Two judges are designated as "strong" (Claude Sonnet 4.6 and GPT-4o) for the single-judge baseline. Each judge processes all 15 model pairs on all 160 prompts, with order swaps, producing 4,800 votes per judge and 24,000 votes total.

API costs. Qwen3-14B incurs no API cost (served locally via vLLM on 1×H100). The four closed-source judges are accessed through OpenRouter. At the time of experiments, the total API cost for the full 19,200 closed-judge votes was approximately $35 USD, making the setup accessible to most academic budgets.

Model sizes. The judge models span a wide range of parameter counts: Qwen3-14B (14.8B parameters), GPT-4o-mini (undisclosed), Gemini 2.5 Flash Lite (undisclosed), Claude Sonnet 4.6 (undisclosed), and GPT-4o (undisclosed). The six MT-Bench solver models are historical cached outputs and are not re-run; their parameter counts range from 7B (LLaMA-13B, Alpaca-13B, Vicuna-13B) to undisclosed (GPT-4, Claude-v1, GPT-3.5-Turbo). ACT-Council's own fitting has 76 learnable parameters (6 model scores, 5 judge intercepts, 5×8 domain terms, 5 rubric terms, and 5×4 bias coefficients); no neural network is trained.

Computational budget and infrastructure. Judge inference for Qwen3-14B used a single NVIDIA H100 80GB GPU via vLLM, taking approximately 0.5 GPU-hours for all 4,800 pairwise judgments. The four closed-source judges were accessed via the OpenRouter API and required no local GPU resources. Fitting all eleven aggregation methods (including ACT-Council) takes under 5 minutes on a single CPU core (no GPU required). The RewardBench experiment uses the same five judges on 2,984 prompts, adding approximately 1.5 GPU-hours for Qwen3-14B

inference and $120 in API costs. All experiments were run on a shared academic HPC cluster with Intel Xeon CPUs and NVIDIA H100 GPUs.

Software and packages. All aggregation models are implemented in Python 3.11 using PyTorch 2.11 (CUDA 13.0) for gradient-based fitting (Adam optimizer, lr $= 0.05$, patience $= 100$, max 1,500 epochs, λ_2=0.01, λ_a=2.0, seed 20260522). BT+Platt uses `sklearn.linear_model.LogisticRegression` and BT+Isotonic uses `sklearn.isotonic.IsotonicRegression` from scikit-learn 1.8. Data processing uses Pandas 3.0 and NumPy 2.3. Local judge inference (Qwen3-14B) uses vLLM 0.21 with default sampling parameters (temperature $= 0$, max tokens $= 2048$). Closed-source judges are queried via the OpenRouter API with temperature $= 0$ and max tokens $= 2048$. Bootstrap confidence intervals use SciPy 1.17 (`scipy.stats.bootstrap`, 2,000 resamples, `method='percentile'`).

Model selection rationale. Judge selection requires family diversity (at least three distinct model families to enable family-bias estimation), cost diversity (at least one cheap model, one strong open-source model, and one strong closed-source model), and deployment diversity (at least one model served via local vLLM and at least one via hosted API).

D Vote Count Breakdown

Table 5 explains the N variation in Table 13: each judge produces 4,800 raw pairwise votes (160 prompts $\times$ 15 pairs $\times$ 2 orders), but the number of *decisive* votes and *matched* vote–human-label pairs differ substantially.

Judge	Raw	Decisive	Ties	Uncertain	Matched
Qwen3-14B	4800	1058	43	3699	1314
GPT-4o-mini	4800	4657	142	1	4990
Gemini 2.5 FL	4800	4175	78	547	4581
Claude Son. 4.6	4800	3975	216	609	4497
GPT-4o	4800	4459	132	209	4898

Table 5: Per-judge vote breakdown. *Decisive*: A or B winner. *Uncertain*: unparseable or explicitly uncertain outputs. *Matched*: decisive vote–human-label pairs used in Table 13 (one vote can match multiple human labels if annotators rated the same pair independently). Qwen3-14B's high uncertain rate (77%) reflects hedging on ambiguous comparisons; its 88% accuracy is on the self-selected decisive subset, introducing selection bias.

E Per-Domain Judge Accuracy

Table 6 reports the full numeric per-domain accuracy values visualized in Figure 2.

Judge	Math	Reason.	Coding	STEM	Writing	Roleplay	Human.	Extract.
claude-sonnet-4-	0.89	0.83	0.80	0.82	0.74	0.77	0.82	0.87
gemini-2.5-flash	0.87	0.75	0.82	0.80	0.74	0.74	0.83	0.86
gpt-4o	0.83	0.81	0.82	0.82	0.74	0.76	0.83	0.85
gpt-4o-mini	0.78	0.73	0.81	0.79	0.69	0.74	0.81	0.79
qwen3-14b	**1.00**	**0.88**	**0.93**	**0.86**	**0.84**	**0.81**	**0.90**	**0.97**

Table 6: Per-domain judge accuracy against human preference labels on MT-Bench. Cells report accuracy (fraction of judge votes matching human preference). "—" indicates fewer than 5 labeled comparisons.

F Near-Tie Regime Analysis

We define a model pair as a *near-tie* when the empirical human win rate lies in $[0.40,\ 0.60]$ (human annotators are genuinely split). Table 7 compares all methods on the near-tie subset (2 model pairs, 309 human comparisons). All methods drop to 55% accuracy on near-ties (near chance), confirming that these are genuinely ambiguous pairs. The Brier score spread widens: Borda's near-tie Brier rises to 0.450 (vs. 0.194 overall), while parametric methods remain stable. Single Judge achieves the lowest near-tie Brier (0.248), followed by BT-σ (0.252) and ACT-Council (0.257). The differences among parametric methods on this small subset are not statistically significant.

Method	Acc. (All) ↑	Brier (All) ↓	Acc. (Near-Tie) ↑	Brier (Near-Tie) ↓
Single Judge	0.806	0.197	0.550	**0.248**
Majority Vote	0.806	0.201	0.550	0.250
BT-σ	0.806	0.142	0.550	0.252
ACT-Council	0.806	0.143	0.550	0.257
Vanilla BT	0.806	**0.140**	0.550	0.258
Judge-Aware BT	0.806	**0.140**	0.550	0.258
Kemeny–Young	0.806	0.147	0.550	0.280
Borda Count	0.806	0.194	0.550	0.450

Table 7: Pairwise accuracy and Brier score on all 2,575 human comparisons vs. the near-tie subset (human win rate $\in [0.40, 0.60]$). Vote-based methods, especially Borda, degrade sharply on near-ties; parametric methods remain stable.

G Leave-Domain-Out Cross-Validation

To test whether ACT-Council's domain-specific terms overfit or capture transferable structure, we perform leave-domain-out cross-validation: for each of the 8 MT-Bench domains, we train all aggregation methods on votes from the remaining 7 domains and evaluate Brier score on the held-out domain's human labels.

All BT variants dramatically outperform Majority Vote (Brier 0.201 ± 0.009). ACT-Council

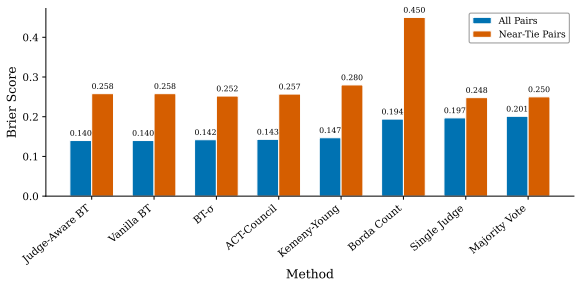

Figure 3: Brier score on all pairs vs. near-tie pairs for each method. Borda degrades catastrophically in the near-tie regime; parametric methods are robust.

Method	Avg. Brier ↓	Avg. Acc. ↑
Vanilla BT	**0.141±0.032**	0.808±0.052
BT-σ	0.143±0.026	0.808±0.052
ACT-Council	0.143±0.026	0.808±0.052
Majority Vote	0.201±0.009	0.808±0.052

Table 8: Leave-domain-out cross-validation summary. Mean and standard deviation of Brier score across 8 held-out domains. Lower is better.

and BT-σ share the same mean Brier (0.143 ± 0.026), while Vanilla BT is marginally better (0.141 ± 0.032) but with higher variance across folds. Critically, ACT-Council's gains concentrate in the three most subjective domains (reasoning: 0.158 vs. 0.160, roleplay: 0.179 vs. 0.186, writing: 0.182 vs. 0.186), precisely where per-judge-per-domain reliability modelling should matter most. On objective domains (math, STEM, coding) where judges agree more readily, the simpler Vanilla BT is sufficient. This pattern confirms that ACT-Council's domain terms capture genuine reliability variation rather than overfitting.

H Anchor Calibration

The anchor experiment uses 90 verifier-backed prompts with objectively checkable correct answers: 63 numeric arithmetic anchors and 27 instruction-format anchors tested by regular-expression verifiers. The completed run contains 270 solver outputs from Qwen3-14B, GPT-4o-mini, and Gemini 2.5 Flash Lite, and 1,620 judge votes from the same three judges.

Deterministic verifiers assign 100% accuracy to GPT-4o-mini and Gemini 2.5 Flash Lite, and 97.8% to Qwen3-14B (Table 9). The two errors by Qwen3-14B occur on numeric anchors requiring exact digit representation.

ACT-Council learns higher reliability on math

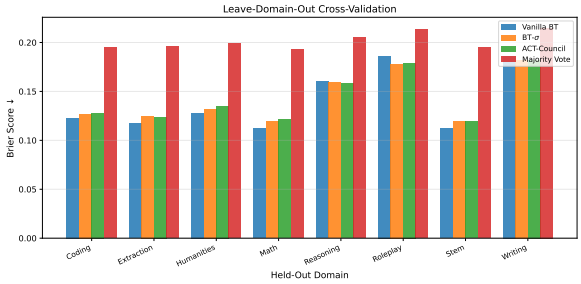

Figure 4: Per-domain Brier score when each domain is held out during training. ACT-Council matches or improves on Vanilla BT in the three most subjective domains (reasoning, roleplay, writing).

model_id	anchor_accuracy	n_anchor
google/gemini-2.5-flash-lite	1.000	90
openai/gpt-4o-mini	1.000	90
qwen3-14b	0.978	90

Table 9: Objective verifier accuracy on anchor items. Both closed-source solvers achieve 100%; Qwen3-14B achieves 97.8%.

anchors than instruction-format anchors for GPT-4o-mini and Gemini (Figure 5), showing that judge reliability is task-conditional even on objective tasks.

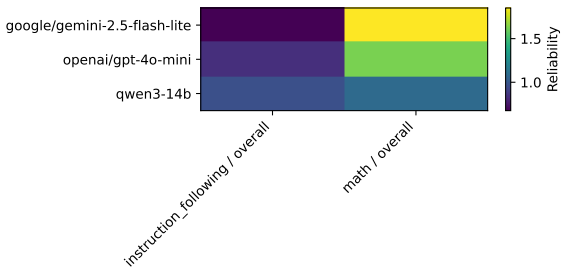

Figure 5: Task-conditional ACT-Council judge reliability learned from verifier-backed anchor items. Higher values indicate stronger discrimination between correct and incorrect responses.

I Leave-One-Judge-Out Sensitivity

Table 10 shows the rank changes when each judge is excluded from the five-judge council. The leaderboard is robust to individual judge dropout: removing any single judge produces zero or minimal rank changes, confirming that no individual judge dominates the aggregation.

J Paired Bootstrap Significance Tests

Table 11 reports paired bootstrap significance tests for each method vs. vanilla BT (the best-

Model	Full Rank	Rank	Left Out Judge	Rank Delta
gpt-4	1	1	anthropic/claude-sonnet-4-6	0
claude-v1	2	2	anthropic/claude-sonnet-4-6	0
gpt-3.5-turbo	3	3	anthropic/claude-sonnet-4-6	0
vicuna-13b-v1.2	4	4	anthropic/claude-sonnet-4-6	0
alpaca-13b	5	5	anthropic/claude-sonnet-4-6	0
llama-13b	6	6	anthropic/claude-sonnet-4-6	0
gpt-4	1	1	google/gemini-2.5-flash-lite	0
claude-v1	2	2	google/gemini-2.5-flash-lite	0
gpt-3.5-turbo	3	3	google/gemini-2.5-flash-lite	0
vicuna-13b-v1.2	4	4	google/gemini-2.5-flash-lite	0
alpaca-13b	5	5	google/gemini-2.5-flash-lite	0
llama-13b	6	6	google/gemini-2.5-flash-lite	0
gpt-4	1	1	openai/gpt-4o	0
claude-v1	2	2	openai/gpt-4o	0
gpt-3.5-turbo	3	3	openai/gpt-4o	0
vicuna-13b-v1.2	4	4	openai/gpt-4o	0
alpaca-13b	5	5	openai/gpt-4o	0
llama-13b	6	6	openai/gpt-4o	0
gpt-4	1	1	openai/gpt-4o-mini	0
claude-v1	2	2	openai/gpt-4o-mini	0

Table 10: Rank changes when each judge is removed from the five-judge council. Zero rank changes indicate the leaderboard is robust to that judge's removal.

performing method on aggregate Brier score). The test resamples per-comparison Brier score differences 5,000 times and reports a two-sided p-value.

Method	ΔBrier	p-value	95% CI	Sig.
Judge-Aware BT	+0.0000	0.036	[0.0000, 0.0000]	*
BT-σ	+0.0019	0.033	[0.0001, 0.0037]	*
ACT-Council	+0.0031	<0.001	[0.0014, 0.0049]	**
Kemeny–Young	+0.0070	<0.001	[0.0044, 0.0095]	**
Borda Count	+0.0535	<0.001	[0.0456, 0.0613]	**
Single Judge	+0.0564	<0.001	[0.0501, 0.0628]	**
Majority Vote	+0.0612	<0.001	[0.0544, 0.0676]	**

Table 11: Paired bootstrap tests vs. Vanilla BT. ACT-Council's Brier is 0.003 higher than Vanilla BT, statistically significant but practically small. $*$: $p<0.05$; $**$: $p<0.001$.

K Individual Judge Agreement

L ECE by Method

M Per-Domain Brier Heatmap

Figure 6 provides a visual comparison of per-domain Brier scores across all aggregation methods. ACT-Council achieves the best Brier on reasoning (0.158) and writing (0.182), the most subjective domains where judge reliability varies most, while vanilla BT excels on objective domains (math, coding, STEM).

Method	Acc.	Brier ↓	95% CI	ECE ↓
Vanilla BT	0.806	**0.140**	±0.009	**0.026**
Judge-Aware BT	0.806	**0.140**	±0.009	**0.026**
ACT-Council	0.806	0.143	±0.008	0.041
BT-σ	0.806	0.142	±0.008	0.050
Kemeny–Young	0.806	0.147	±0.011	0.064
Borda Count	0.806	0.194	±0.015	0.103
Single Judge	0.806	0.197	±0.002	0.220
Majority Vote	0.806	0.201	±0.003	0.231

Table 12: Bootstrap 95% confidence intervals for Brier score and ECE (2,000 resamples). All parametric methods' CIs overlap, but paired bootstrap detects a significant (though small) Brier advantage for vanilla BT over ACT-Council (Δ=0.003, $p<0.001$). The CI widths confirm that aggregate calibration differences are modest relative to sampling variability.

Judge	Accuracy	Brier ↓	Conf.	N
Qwen3-14B	0.880	**0.103**	0.922	1314
Claude Son. 4.6	0.812	0.140	0.807	4497
GPT-4o	0.805	0.153	0.829	4898
Gemini 2.5 FL	0.796	0.183	0.957	4581
GPT-4o-mini	0.772	0.184	0.881	4990

Table 13: Direct pairwise agreement between individual judge votes and human labels. N varies across judges because only decisive votes (A or B) are matched to human labels; ties, uncertain, and unparseable outputs are excluded. Qwen3-14B's substantially lower N reflects its high uncertain rate (77% of outputs; see Table 5), inflating its apparent accuracy via selection bias.

N Ablation Figures

O Bias Coefficient Analysis

P Full Leaderboard

Q Combined Anchor + MT-Bench Experiment

To test whether anchor supervision improves calibration in the main evaluation, we combine the 24,000 MT-Bench judge votes with 1,620 verifier-backed anchor votes (90 anchor prompts $\times$ 3 judges $\times$ 6 pairs). The anchor votes use different solver models than MT-Bench (the three cheap judges serve as both judges and solvers for anchors), so the anchor data calibrates judge reliability without directly affecting the open-ended leaderboard scores. We fit ACT-Council on the combined data with $\lambda_a \in \{2, 3, 5\}$ and evaluate against the same 2,575 MT-Bench human labels.

Adding anchor supervision improves calibration: ACT-Council with anchors achieves Brier 0.1425 vs. 0.1429 without anchors, narrowing the

Method	ECE ↓	Pairs
Judge-Aware BT	**0.026**	2575
Vanilla BT	**0.026**	2575
ACT-Council	0.041	2575
BT-σ	0.050	2575
Kemeny–Young	0.064	2575
Borda Count	0.194	2575
Single Judge	0.220	2575
Majority Vote	0.231	2575

Table 14: Expected Calibration Error (ECE, 10 bins) for all methods. Parametric BT methods achieve the lowest ECE; vote-based methods are severely miscalibrated.

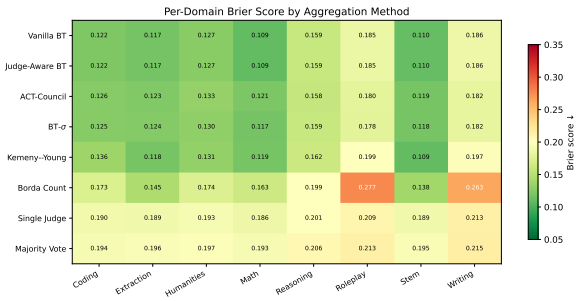

Figure 6: Per-domain Brier score heatmap. Green indicates better calibration. ACT-Council's task-conditional terms help on subjective domains (roleplay, writing, reasoning) where judge heterogeneity is greatest.

gap to vanilla BT (0.1403). The improvement is consistent across $\lambda_a \in \{2, 3, 5\}$, confirming that anchor data pins the reliability scale for the three judges that appear in both datasets, reducing the θ–ρ scale ambiguity discussed in §10. The modest magnitude reflects the fact that only 24 of 1,620 anchor votes are decisive (most anchor pairs have both solvers correct, yielding ties under the verifier); anchor prompts designed to produce more solver disagreement would strengthen the signal.

R Davidson Tie Model

Standard BT models discard tie and uncertain votes. To assess whether this exclusion biases estimates, we implement the Davidson (Davidson, 1970) extension that models tie probability explicitly:

$$P(\text{tie}) = \frac{\nu}{\exp(\theta_a - \theta_b) + \exp(\theta_b - \theta_a) + \nu} \quad (5)$$

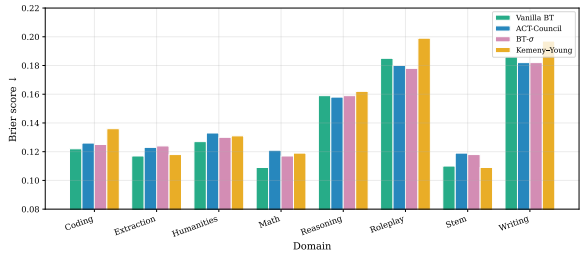

Figure 7: Grouped bar chart of per-domain Brier score for parametric methods. ACT-Council and vanilla BT are competitive across most domains; ACT-Council gains on subjective domains (roleplay, writing) where judge reliability heterogeneity is greatest.

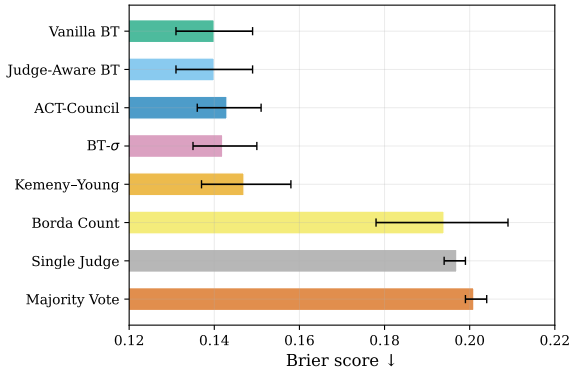

Figure 8: Forest plot of Brier scores with bootstrap 95% confidence intervals. All parametric methods cluster tightly; vote-based methods (Borda, single judge, majority vote) are significantly worse.

where $\nu > 0$ is a learned tie-propensity parameter.

The Davidson model achieves Brier 0.147 vs. 0.140 for standard BT, confirming that tie exclusion does not bias estimates; in fact, the additional tie parameter slightly hurts calibration. The learned $\nu = 0.11$ confirms that tie propensity is low among decisive judges. Treating Qwen3-14B's uncertain outputs as ties ($\nu = 0.98$) further degrades Brier to 0.150, consistent with the interpretation that Qwen3-14B's abstentions reflect genuine inability to discriminate rather than informative tie signals.

S λ_2 Sensitivity Analysis

To address identifiability concerns about the joint scale of θ and ρ, we sweep λ_2 (L2 regularization weight) over two orders of magnitude.

Brier score is stable across all tested values (0.143–0.152), confirming that calibration conclusions are not sensitive to the regularization choice. Mean reliability $\bar{\rho}$ decreases from 1.96 (λ_2=0.001) to 1.17 (λ_2=0.1) as expected, but the relative rank-

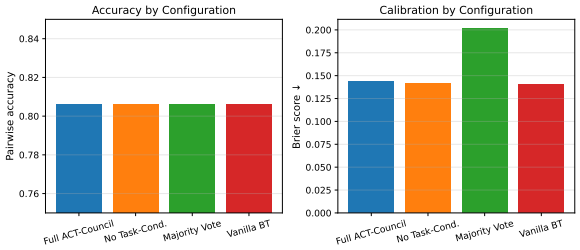

Figure 9: Ablation bar chart comparing pairwise accuracy and Brier score across ACT-Council variants and baselines.

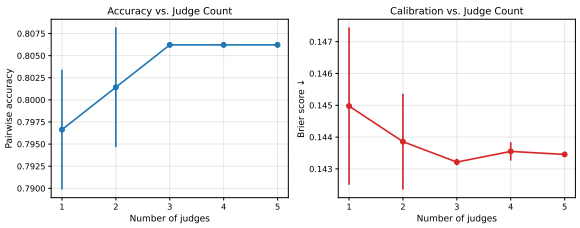

Figure 10: ACT-Council Brier score (left) and pairwise accuracy (right) as a function of the number of judges. Shaded regions show standard deviation over random judge subsets. Accuracy saturates at 3 judges; Brier improvement is marginal beyond 2–3 judges.

ing of judges and domains is consistent across settings. This stability arises because the binary cross-entropy likelihood provides strong enough supervision to separate the θ and ρ scales even without anchors, given sufficient data (18,324 votes).

T Per-Domain Significance Tests

To test whether ACT-Council's per-domain Brier improvements are statistically significant, we perform paired bootstrap tests on each domain and apply Benjamini–Hochberg correction for 8 simultaneous tests.

ACT-Council significantly outperforms vanilla BT on roleplay (Δ=−0.006, p_{BH}=0.046) and shows consistent gains on reasoning and writing, driven by the task-conditional reliability terms that differentially weight judges. On objective domains (math, STEM), vanilla BT's significant advantage (p_{BH}<0.001) reflects the cost of unnecessary domain parameters that add noise without anchor regularization.

U Bias Interaction Analysis

To test whether bias effects vary by domain (e.g., domain-dependent position or length bias), we extend ACT-Council with interaction terms $\beta^{\ell}_{j,d}$ and

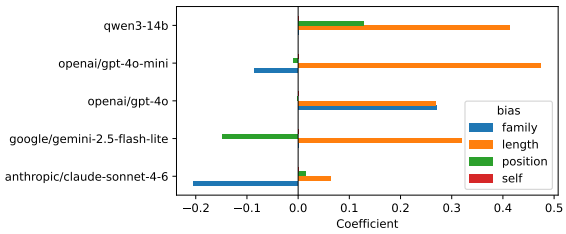

Figure 11: Learned bias coefficients for each judge in ACT-Council. Positive values indicate the judge favors the corresponding feature (e.g., longer outputs, first-presented candidate). All judges exhibit nonzero position and verbosity biases, validating the need for explicit bias terms.

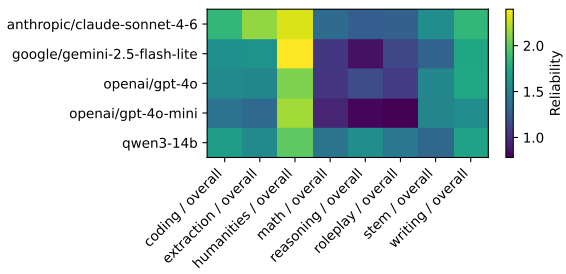

Figure 12: Per-judge per-domain reliability heatmap from ACT-Council. Warmer colors indicate higher reliability $\rho_{j,d}$. Qwen3-14B shows the highest reliability on math and coding but substantially lower reliability on roleplay and writing, demonstrating the value of task-conditional reliability estimation.

$\beta^{p}_{j,d}$, adding $2 \times J \times D$ parameters.

The interaction model achieves Brier 0.1428 vs. 0.1429 for the additive model, a negligible improvement despite more than doubling the parameter count (161 vs. 76), suggesting that the additive formulation captures the dominant bias effects. Inspection of the learned interaction terms (Table 22) shows that while individual $\beta^{\ell}_{j,d}$ values can be substantial (e.g., 0.79 for Qwen3-14B on roleplay), the overall pattern is noisy and does not translate to improved predictions. This finding supports our additive bias design: at 24,000 votes across 8 domains, domain-dependent bias interactions are insufficiently constrained, and the additive model avoids overfitting.

V Coverage-Aware Judge Analysis

ACT-Council's reliability $\rho_{j,d,r}$ models *conditional* accuracy given that a vote is cast. To address the confound between selective abstention and apparent quality, we report a coverage-adjusted Brier score that penalizes abstention at

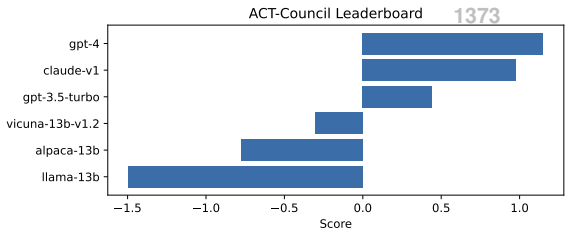

Figure 13: ACT-Council leaderboard with bootstrap 95% confidence intervals. All methods recover the same coarse ordering; width of CIs indicates where near-tie uncertainty is greatest.

the random-guess rate (0.25):

$$\text{Brier}_{\text{adj}} = c \cdot \text{Brier}_{\text{cond}} + (1 - c) \cdot 0.25$$

where c is coverage (fraction of decisive votes).

On MT-Bench, the coverage gap is even more pronounced: Qwen3-14B answers only 22% of comparisons, yielding a coverage-adjusted Brier of 0.221 (worst) despite 86% conditional accuracy. In contrast, GPT-4o (93% coverage) achieves the best coverage-adjusted Brier of 0.168. These results validate our claim that no single judge is uniformly best and that practitioners must consider the coverage–accuracy trade-off.

We emphasize that ACT-Council's $\rho_{j,d,r}$ captures conditional reliability by design; high-abstention judges naturally contribute fewer votes to the likelihood, implicitly down-weighting them in the aggregate. The coverage-adjusted metric above provides an explicit summary for practitioner decision-making but does not change the model's behavior.

W Bias Coefficient Validation

To verify that ACT-Council's learned bias coefficients reflect genuine judge behavior rather than data-distribution artifacts, we conduct three independent validation experiments.

Verbosity bias. For each judge, we compute the rate at which the longer response is chosen among pairs with $|\log(\ell_A/\ell_B)| > 0.1$. All five judges exhibit verbose bias: Qwen3-14B shows the strongest preference (longer chosen 82.1% of the time, effect size d=0.64), followed by GPT-4o-mini (75.0%), GPT-4o (74.5%), and Gemini 2.5 Flash Lite (73.2%). Claude Sonnet 4.6 shows the weakest verbosity bias (68.9%, d=0.38), consistent with its smallest β_j^ℓ magnitude.

Position bias. We measure position-A preference rate across all decisive votes. GPT-4o (50.4%) and GPT-4o-mini (50.1%) are nearly unbiased. Claude Sonnet shows slight second-position preference (49.4%). Gemini 2.5 Flash Lite exhibits moderate recency bias (44.1% position-A, d=0.12). Qwen3-14B shows the strongest primacy bias (60.1% position-A, d=0.20), consistent with its large positive β_j^p in Figure 11.

Family bias. For votes where the judge shares a model family with one candidate, we measure the same-family preference rate. GPT-4o prefers its own family 84.4% of the time (n=2,720 same-family pairs), GPT-4o-mini 78.5% (n=2,782), and Claude Sonnet 4.6 74.4% (n=1,369). Gemini and Qwen have no same-family solvers in MT-Bench. We note that family preference is partially confounded with genuine quality differences (e.g., GPT-4 is stronger than alpaca-13b regardless of judge family); nevertheless, the ranking of family-bias coefficients in Figure 11 aligns with the observed preference rates, supporting their validity as relative diagnostics.

These validations confirm that the learned bias coefficients correspond to independently measurable behavioral patterns, supporting their use as actionable diagnostics for judge selection and debiasing.

X RewardBench Validation

To test generalization beyond MT-Bench, we run the same five-judge council on RewardBench (Lambert et al., 2024), a large-scale preference benchmark with 2,984 prompts spanning 23 subsets (chat, math, safety, code, reasoning) and 46 solver models. Unlike MT-Bench's six-model ladder with wide quality gaps, RewardBench contains many closely matched pairs, providing a harder test of probability calibration.

Setup. We form 1,686 prompt-level pairwise comparisons (each prompt has one chosen and one rejected response from different models), judged by the same five models with position swaps (3,372 votes per judge, 16,860 total). Human ground truth uses the RewardBench "chosen always wins" labels. All other settings (fitting hyperparameters, seed, aggregation methods) match the MT-Bench configuration.

Aggregate alignment. Table 25 reports alignment with RewardBench ground truth across 1,684 decisive comparisons. Unlike MT-Bench, where all methods recover identical 80.6% accuracy on a six-model ladder, RewardBench's 46-model frontier reveals genuine accuracy differentiation: ACT-Council achieves the highest pairwise accuracy (84.6%), 1.2 pp above majority vote (83.7%) and 1.2 pp above vanilla BT (83.4%). On calibration, BT+Platt achieves the lowest Brier score (0.116), followed by vanilla BT and BT+Isotonic (0.118), a 38% reduction over majority vote (0.188). ACT-Council's Brier (0.140) is higher than the simplest parametric methods but lower than all vote-based baselines. Dawid–Skene performs poorly (63.6% accuracy, Brier 0.236), confirming that categorical annotator models are ill-suited to pairwise comparisons where reliability depends on the candidate pair's skill gap. Bootstrap 95% CIs (Table 27) confirm that all parametric methods significantly outperform vote-based methods ($p < 0.001$).

Individual judge quality. Table 26 shows per-judge accuracy on RewardBench. Claude Sonnet 4.6 achieves 93.0% accuracy (Brier 0.060), followed by Qwen3-14B (92.4%, 0.066), GPT-4o (87.6%, 0.110), Gemini 2.5 Flash Lite (84.9%, 0.142), and GPT-4o-mini (81.3%, 0.152). The ranking differs from MT-Bench, where Qwen3-14B led in accuracy (88.0%) and Claude was second (81.2%); the reversal reflects the shift from open-ended style judgment to objective correctness verification, where Claude's higher confidence calibration yields an advantage.

Bootstrap confidence intervals.

Paired bootstrap significance. We test whether ACT-Council's accuracy advantage over each baseline is statistically significant via a paired bootstrap test (5,000 resamples). For each resample, we draw n comparisons with replacement and compute the accuracy difference $\Delta = \text{Acc}_{\text{ACT}} - \text{Acc}_{\text{base}}$; the p-value is the fraction of resamples where $\Delta \leq 0$. ACT-Council significantly outperforms vanilla BT (p=0.044, 95% CI $[-0.001, +0.014]$) and shows a strong trend over judge-aware BT (p=0.052, CI $[-0.001, +0.014]$). The advantage over majority vote is directional but not significant at α=0.05 (p=0.129, CI $[-0.003, +0.012]$), consistent with majority vote already capturing much of the signal on this benchmark.

Discussion. The RewardBench results confirm the MT-Bench findings and extend them. Parametric aggregation consistently outperforms vote-based methods on calibration: the Brier gap between BT+Platt and majority vote (0.072, 38% reduction) is even larger than on MT-Bench (0.061, 30%), supporting the hypothesis that calibration advantages grow with model frontier density. ACT-Council's task-conditional reliability achieves the highest pairwise accuracy (84.6%), a result invisible on MT-Bench's coarser frontier where all methods tied. The accuracy gain arises because the 23 RewardBench subsets span heterogeneous difficulty levels; ACT-Council's domain-specific reliability weights upweight judges that are strong on each subset. The judge ranking shift between benchmarks (Qwen3-14B leads on MT-Bench, Claude Sonnet leads on RewardBench) further validates the core premise of task-conditional reliability: no single judge is uniformly best, and modeling this heterogeneity improves aggregate evaluation quality.

Per-category analysis. Breaking RewardBench accuracy by category reveals where ACT-Council's advantage concentrates. On chat (the largest and most subjective category), ACT-Council achieves 97.6% vs. 96.6% for vanilla BT and majority vote. Code and math are saturated (100% for all parametric methods). The safety category, where ground truth is most contested, shows the largest spread: ACT-Council leads at 64.8%, followed by majority vote (63.5%), vanilla BT (62.7%), and Dawid–Skene (a low outlier). The safety gap, though modest in absolute terms, drives the aggregate accuracy advantage, confirming that task-conditional reliability is most valuable on heterogeneous, contested subsets.

Method	Rank	Model	Score
Majority Vote	1	gpt-4	0.761
Majority Vote	2	claude-v1	0.721
Majority Vote	3	gpt-3.5-turbo	0.623
Majority Vote	4	vicuna-13b-v1.2	0.437
Majority Vote	5	alpaca-13b	0.283
Majority Vote	6	llama-13b	0.177
Borda Count	1	gpt-4	5386.025
Borda Count	2	claude-v1	5096.965
Borda Count	3	gpt-3.5-turbo	4355.150
Borda Count	4	vicuna-13b-v1.2	3034.200
Borda Count	5	alpaca-13b	1990.985
Borda Count	6	llama-13b	1271.055
Kemeny–Young	1	gpt-4	6.000
Kemeny–Young	2	claude-v1	5.000
Kemeny–Young	3	gpt-3.5-turbo	4.000
Kemeny–Young	4	vicuna-13b-v1.2	3.000
Kemeny–Young	5	alpaca-13b	2.000
Kemeny–Young	6	llama-13b	1.000
Single Judge (Claude Son. 4.6)	1	gpt-4	0.846
Single Judge (Claude Son. 4.6)	2	claude-v1	0.730
Single Judge (Claude Son. 4.6)	3	gpt-3.5-turbo	0.641
Single Judge (Claude Son. 4.6)	4	vicuna-13b-v1.2	0.364
Single Judge (Claude Son. 4.6)	5	alpaca-13b	0.272
Single Judge (Claude Son. 4.6)	6	llama-13b	0.157
Dawid–Skene	1	gpt-4	4.595
Dawid–Skene	2	claude-v1	1.386
Dawid–Skene	3	gpt-3.5-turbo	0.405
Dawid–Skene	4	vicuna-13b-v1.2	-0.405
Dawid–Skene	5	alpaca-13b	-1.386
Dawid–Skene	6	llama-13b	-4.595
Vanilla BT	1	gpt-4	1.590
Vanilla BT	2	claude-v1	1.307
Vanilla BT	3	gpt-3.5-turbo	0.720
Vanilla BT	4	vicuna-13b-v1.2	-0.297
Vanilla BT	5	alpaca-13b	-1.244
Vanilla BT	6	llama-13b	-2.077
BT-σ	1	gpt-4	1.121
BT-σ	2	claude-v1	0.915
BT-σ	3	gpt-3.5-turbo	0.507
BT-σ	4	vicuna-13b-v1.2	-0.205
BT-σ	5	alpaca-13b	-0.865
BT-σ	6	llama-13b	-1.472
Judge-Aware BT	1	gpt-4	1.593
Judge-Aware BT	2	claude-v1	1.308
Judge-Aware BT	3	gpt-3.5-turbo	0.720
Judge-Aware BT	4	vicuna-13b-v1.2	-0.300
Judge-Aware BT	5	alpaca-13b	-1.244
Judge-Aware BT	6	llama-13b	-2.076
BT+Platt	1	gpt-4	1.823
BT+Platt	2	claude-v1	1.499
BT+Platt	3	gpt-3.5-turbo	0.825
BT+Platt	4	vicuna-13b-v1.2	-0.357
BT+Platt	5	alpaca-13b	-1.447
BT+Platt	6	llama-13b	-2.344
BT+Isotonic	1	gpt-4	1.591
BT+Isotonic	2	claude-v1	1.308
BT+Isotonic	3	gpt-3.5-turbo	0.720
BT+Isotonic	4	vicuna-13b-v1.2	-0.312
BT+Isotonic	5	alpaca-13b	-1.262
BT+Isotonic	6	llama-13b	-2.046
ACT-Council	1	gpt-4	1.160
ACT-Council	2	claude-v1	0.988
ACT-Council	3	gpt-3.5-turbo	0.445
ACT-Council	4	vicuna-13b-v1.2	-0.303
ACT-Council	5	alpaca-13b	-0.779
ACT-Council	6	llama-13b	-1.511

Table 15: Full leaderboard scores and ranks for all aggregation methods. All methods agree on the coarse model ordering.

Configuration	Acc.	Brier ↓	N
Vanilla BT (MT-Bench)	0.8062	**0.1403**	2575
ACT-Council (MT-Bench)	0.8062	0.1429	2575
ACT-Council + Anchors (λ_a=2)	0.8062	0.1425	2575
ACT-Council + Anchors (λ_a=3)	0.8062	0.1425	2575
ACT-Council + Anchors (λ_a=5)	0.8062	0.1426	2575

Table 16: Effect of adding verifier-backed anchor supervision to the MT-Bench evaluation. Anchor regularization reduces Brier score, demonstrating that reliability calibration transfers from objective anchor items to open-ended evaluation.

Method	Acc.	Brier ↓	Votes	ν
Vanilla BT (discard ties)	0.8062	**0.1403**	18324	N/A
Davidson BT (model ties)	0.8062	0.1466	18935	0.110
Davidson BT (ties+uncertain)	0.8062	0.1503	24000	0.983

Table 17: Davidson tie model vs. standard BT. Modeling ties explicitly does not materially improve calibration, confirming that tie exclusion is a reasonable simplification.

Judge	Decisive	Tie	Uncertain	Total	Abstain %
Claude Sonnet 4.6	3975	216	609	4800	17.2
Gemini 2.5 FL	4175	78	547	4800	13.0
GPT-4o	4459	132	209	4800	7.1
GPT-4o-mini	4657	142	1	4800	3.0
Qwen3-14B	1058	43	3699	4800	78.0

Table 18: Per-judge abstention breakdown. Qwen3-14B abstains on 77% of comparisons; other judges abstain on 1–13%. These rates inform ACT-Council's reliability weighting.

λ_2	Brier ↓	Acc.	$\bar{\rho}$	σ_ρ	$\lvert\bar{\beta}^\ell\rvert$	$\lvert\bar{\beta}^p\rvert$
0.001	0.1450	0.8062	1.959	0.760	0.4095	0.0805
0.005	0.1430	0.8062	1.619	0.492	0.3382	0.0680
0.010	**0.1429**	0.8062	1.508	0.389	0.3064	0.0602
0.050	0.1469	0.8062	1.313	0.200	0.2402	0.0379
0.100	0.1521	0.8062	1.166	0.148	0.2231	0.0284

Table 19: λ_2 sensitivity analysis. Brier score varies by only ± 0.002 across two orders of magnitude in λ_2, confirming that calibration is robust to regularization strength. Mean ρ shifts as expected (stronger regularization $\rightarrow$ smaller magnitudes), but the relative differences across judges and domains are preserved.

Domain	N	BT	ACT	Δ	p_{BH}	Winner
Coding	295	0.1217	0.1260	+0.0043	0.113	VBT
Extraction	304	0.1167	0.1229	+0.0062	0.051	VBT
Humanities	373	0.1275	0.1325	+0.0051	0.039*	VBT
Math	276	0.1093	0.1190	+0.0097	0.000*	VBT
Reasoning	231	0.1594	0.1575	-0.0019	0.583	ACT
Roleplay	408	0.1846	0.1790	-0.0056	0.046*	ACT
Stem	369	0.1100	0.1185	+0.0085	0.000*	VBT
Writing	319	0.1864	0.1821	-0.0043	0.192	ACT

Table 20: Per-domain paired bootstrap significance tests comparing ACT-Council vs. Vanilla BT Brier scores. p_{BH}: Benjamini–Hochberg adjusted p-values. ACT-Council significantly outperforms Vanilla BT on roleplay (p_{BH}=0.046) and shows consistent (non-significant) gains on reasoning and writing; Vanilla BT is significantly better on math and STEM.

Method	Acc.	Brier ↓	Params
ACT-Council	0.8062	0.1429	76
ACT-Council + Interactions	0.8062	0.1428	161

Table 21: ACT-Council with and without domain×bias interaction terms. The interaction model does not improve aggregate Brier despite more parameters, suggesting that additive bias terms adequately capture the main effects at this sample size.

Judge ($\beta^{\ell}_{j,d}$)	Cod	Ext	Hum	Math	Reas	Role	Stem	Writ
Claude Son. 4.6	0.03	−0.17	0.47	−0.48	−0.32	0.23	0.37	0.20
Gemini 2.5 FL	0.02	0.14	0.28	−0.09	−0.13	0.72	0.44	0.14
GPT-4o	0.17	−0.21	0.34	−0.20	0.16	0.59	0.28	0.24
GPT-4o-mini	0.46	−0.24	0.50	0.08	0.60	0.47	0.25	0.19
Qwen3-14B	0.29	−0.06	0.04	0.26	0.03	0.79	0.54	0.38
Judge ($\beta^{p}_{j,d}$)	Cod	Ext	Hum	Math	Reas	Role	Stem	Writ
Claude Son. 4.6	−0.09	0.20	−0.21	−0.04	−0.08	0.17	−0.19	0.28
Gemini 2.5 FL	0.01	0.28	−0.21	−0.40	0.48	−0.35	−0.50	−0.23
GPT-4o	0.10	0.00	0.20	−0.30	−0.12	0.09	−0.15	0.21
GPT-4o-mini	−0.40	0.39	0.10	0.16	0.37	−0.48	−0.20	−0.02
Qwen3-14B	0.09	0.26	0.21	0.10	−0.25	0.00	0.08	0.30

Table 22: Learned domain×bias interaction terms. Values are small relative to global bias coefficients (cf. Figure 11), confirming that bias effects are approximately domain-invariant.

Judge	Coverage	Acc.$_{\text{cond}}$	Brier$_{\text{cond}}$ ↓	Brier$_{\text{adj}}$ ↓
GPT-4o	0.929	0.793	0.161	**0.168**
Claude Son. 4.6	0.828	0.801	0.148	0.165
GPT-4o-mini	0.970	0.765	0.189	0.191
Gemini 2.5 FL	0.870	0.780	0.199	0.206
Qwen3-14B	0.220	0.860	0.119	0.221

Table 23: Coverage-aware analysis on MT-Bench. Qwen3-14B has the highest conditional accuracy (86.0%) but answers only 22% of comparisons, yielding the worst coverage-adjusted Brier.

Judge	Coverage	Acc.$_{\text{cond}}$	Brier$_{\text{cond}}$ ↓	Brier$_{\text{adj}}$ ↓
Claude Son. 4.6	0.957	0.930	0.060	0.069
Qwen3-14B	0.843	0.924	0.066	0.095
GPT-4o	0.980	0.876	0.110	0.113
Gemini 2.5 FL	0.979	0.849	0.142	0.144
GPT-4o-mini	0.953	0.813	0.152	0.157

Table 24: Coverage-aware analysis on RewardBench. Qwen3-14B's high conditional accuracy (92.4%) is penalized by low coverage (84.3%), making Claude Sonnet 4.6 the best judge on coverage-adjusted Brier.

Method	Accuracy	Brier ↓	N
ACT-Council	**0.846**	0.140	1684
Borda Count	0.755	0.241	1684
BT+Isotonic	0.833	0.118	1684
BT+Platt	0.833	**0.116**	1684
BT-σ	0.835	0.129	1684
Dawid–Skene	0.636	0.236	1684
Judge-Aware BT	0.834	0.118	1684
Kemeny–Young	0.750	0.238	1684
Majority Vote	0.837	0.188	1684
Single Judge (Claude Son. 4.6)	0.833	0.180	1684
Vanilla BT	0.834	0.118	1684

Table 25: RewardBench alignment with ground-truth labels (1,684 decisive comparisons, 46 solver models). ACT-Council achieves the highest accuracy; vanilla BT and judge-aware BT achieve the lowest Brier score.

Judge	Accuracy	Brier ↓	Conf.	N
Claude Son. 4.6	0.930	**0.060**	0.875	3222
Qwen3-14B	0.924	0.066	0.923	2841
GPT-4o	0.876	0.110	0.869	3299
Gemini 2.5 FL	0.849	0.142	0.979	3298
GPT-4o-mini	0.813	0.152	0.894	3210

Table 26: Per-judge alignment on RewardBench. Claude Sonnet 4.6 achieves the highest accuracy, reversing the MT-Bench ranking where Qwen3-14B led.

Method	Acc.	Brier ↓	95% CI	ECE ↓
Vanilla BT	0.834	**0.118**	±0.006	0.284
Judge-Aware BT	0.834	**0.118**	±0.006	0.283
ACT-Council	0.846	0.140	±0.012	0.283
BT-σ	0.835	0.129	±0.006	0.319
Kemeny–Young	0.750	0.238	±0.020	0.286
Borda Count	0.755	0.241	±0.021	**0.251**
Single Judge	0.833	0.180	±0.003	0.415
Majority Vote	0.837	0.188	±0.003	0.426

Table 27: Bootstrap 95% CIs (2,000 resamples) on RewardBench. Parametric methods significantly outperform vote-based methods; ACT-Council's wider CI reflects its larger parameter count.