A CaLLM Judge is a Better Judge: Mitigating Bias in LLM-as-a-judge

Anonymous authors
Paper under double-blind review

Abstract

Large language models (LLMs) are increasingly used as judges to predict which of two candidate answers a human evaluator would prefer, providing a scalable alternative when human evaluation is too costly. Unfortunately, LLM judgments can exhibit systematic biases. Prior work has documented position, verbosity or model family of the candidate answers influencing the judge decision rather than their quality, while additional biases may remain undocumented. *Multicalibration* provides a natural framework to address this problem: each bias corresponds to miscalibration on an identifiable group, *e.g.*, position bias manifests as overconfidence in the answer shown first. The challenge is to fix miscalibration simultaneously across many, potentially unknown groups. To this end, we introduce `CaLLM`, a post hoc framework that multicalibrates an LLM judge, thereby bounding all such biases simultaneously. We prove that `CaLLM` bounds the bias of every group whose error structure is captured by an available representation, and that augmenting it with explicit bias groups further reduces the residual bias. We then derive theoretical conditions under which `CaLLM` improves the judge's accuracy, rather than trading it for calibration. Finally, we evaluate `CaLLM` across pairwise preference benchmarks, showing that it reduces the multicalibration error on both specified and unspecified bias groups, while improving predictive performance of the judge.

1 Introduction

When evaluating the outputs of generative models, obtaining human annotations at scale is often too slow and expensive (Li et al., 2024a). As a consequence, the LLM-as-a-judge framework has emerged as a practical alternative (Zheng et al., 2023): a strong language model is prompted to score or compare candidate responses, enabling evaluations at a scale that human raters cannot match. However, LLM judges can exhibit systematic biases that undermine the reliability of their evaluations (Gu et al., 2024; Tian et al., 2025). For example, Zheng et al. (2023) identify three such biases: *position bias*, *i.e.* the judge favours the answer shown first or second independently of its content, *verbosity bias*, *i.e.* the judge favours longer answers regardless of quality, and *family bias*, *i.e.* the judge favours answers produced by its own model family.

Although mitigation strategies have been proposed, controlling these biases with theoretical guarantees remains an open challenge (Schroeder and Wood-Doughty, 2024; Li et al., 2025a). We propose viewing systematic biases through the lens of group-wise miscalibration. From this perspective, each bias is a miscalibration of the judge's confidence within a particular group: *e.g.*, *verbosity bias* arises in pairwise evaluations where one answer is much longer than the other. To address this within-group miscalibration problem, we resort to *Multicalibration* (Hébert-Johnson et al., 2018), a theoretical framework that requires predictions to be calibrated simultaneously across identifiable groups. Specifically, for each group, the absolute expected discrepancy between the predictions and the label must be below a small threshold α. Problematically, the bias groups of interest are defined over the original input space, whereas LLM evaluation operates on unstructured text and a multicalibrator can only observe a learned representation of it. This raises the question of whether multicalibration in representation space controls biases defined in the original input space.

In this work, we address these challenges with Multi-Calibrated Large Language Models (`CaLLM`), a post hoc bias-mitigation framework that implements bias-correction guarantees in practice (see Figure 1). First, we cast the LLM-as-a-judge bias problem within a multicalibration framework,

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

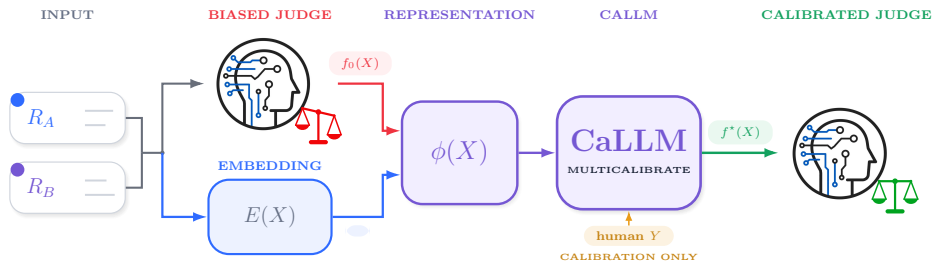


Figure 1: CaLLM's pipeline. We feed CaLLM with some representation $\phi(X)$ of the input (e.g., embeddings $E(X)$ and base scores $f_0(X)$). The final output is a multi-calibrated score $f^\star(X)$.

mapping biases to groups. We then prove that multicalibration in representation space controls bias groups in the original input space whenever their error structure is recoverable from that representation. Importantly, this does not require the groups to be explicitly specified to the calibrator. Finally, we show that, under suitable conditions, bias reduction need not come at the expense of predictive performance, rather CaLLM guarantees improvements of the judge's accuracy. Overall, CaLLM turns these calibration and performance guarantees into a practical post-hoc procedure for mitigating both explicitly specified and recoverable biases in LLM judges.

Our contributions. In a nutshell, our contributions are:

- (i) We formalize the LLM-as-a-Judge biases as a specific instance of a multicalibration problem;
- (ii) We propose CaLLM, a post-hoc bias-mitigation strategy that leverages multicalibration to bound identifiable judge's bias;
- (iii) We prove that multicalibration in representation space bounds recoverable input-space bias patterns and derive conditions under which it improves the judge's accuracy;
- (iv) We show on five benchmark datasets that CaLLM mitigates documented biases, while simultaneously improving the performance of the LLM-as-a-judge framework.

2 Preliminaries

LLM-as-a-judge. We focus on pairwise evaluation of candidate answers, which has emerged as a popular approach for assessing the quality of LLM outputs (Gu et al., 2024). In this setting, the judge receives a user prompt with two anonymized candidate responses and selects the best output, emulating human preference. More formally, a pairwise evaluation instance is a random variable $X = (Q, R_1, R_2) \in \mathcal{X}$, where Q is the prompt and R_1, R_2 are two candidate responses, respectively. The binary label space $\mathcal{Y} \in \{0, 1\}$ indicates human preference, with $Y = 1$ ($Y = 0$) denoting that the human prefers the first (second) answer. We assume access to a given dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ of input-output pairs drawn from an unknown joint distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$. The LLM judge induces a continuous scoring function $f : \mathcal{X} \to [0, 1]$, which estimates the true conditional probability $\mathbb{P}(Y = 1 \mid X)$. Thus, $f(X)$ is the judge's confidence in the first answer. We distinguish two instances of the judge's confidence: f_0 denotes the *base* score returned by the judge, and $f^\star$ the calibrated score. We write f whenever a statement holds for both. We obtain hard predictions by thresholding f at $1/2$ and define the judge's accuracy as $\mathrm{Acc}(f) := \mathbb{P}\big(\mathbb{I}_{\{f(X) \geq 1/2\}} = Y\big)$.

Bias groups. We represent each LLM bias through a collection of binary groups $\mathcal{G}$. Due to the inherently comparative nature of the task, to define groups we first impose a fixed ordering on the candidate responses. Let $\pi : \mathcal{X} \to \widetilde{\mathcal{X}}$ denote a deterministic reordering map such that $\pi(X) = (Q, R_A, R_B)$, where R_A and R_B are the original responses ordered according to a rule independent of Y and f[1], e.g., the lexicographical ordering of their source-model names. We represent group membership through a measurable attribute $a : \widetilde{\mathcal{X}} \to \mathcal{V}_a$, whose value indicates how

[1]This ensures that group membership remains unchanged after calibration.

R_A compares with R_B along the corresponding attribute dimension. Each value $v \in \mathcal{V}_a$ induces a binary group $\tilde{g}_{a,v}(\tilde{X}) := \mathbb{I}_{\{a(\tilde{X})=v\}}$. The corresponding group indicator on the original input space is defined by composition, $g_{a,v} := \tilde{g}_{a,v} \circ \pi$ so that $g_{a,v}(X) = \tilde{g}_{a,v}(\pi(X))$. For simplicity, we henceforth write a generic group as g, omitting the indices (a, v) when they are not needed explicitly.

Example (Documented-biases groups)**.** *Documented biases (Zheng et al., 2023) can be represented through the following attributes: (i)* ***Position****,* $a^{\text{pos}}(\tilde{X}) \in \{-1, +1\}$*, indicating whether* R_A *is shown second or first; (ii)* ***Verbosity****,* $a^{\text{len}}(\tilde{X}) \in \{-1, 0, +1\}$*, indicating whether* R_A *is shorter, the same or longer than* R_B*; (iii)* ***Family****,* $a^{\text{fam}}(\tilde{X}) \in \{-1, 0, +1, +2\}$*, indicating whether neither response belongs to the judge's model family, only* R_B *does, only* R_A *does, or both do. Each group is defined by the presence of a specific attribute value, or by a combination of values across multiple attributes.*

A *bias* is then a systematic discrepancy between the judge's confidence and the true human preference rate *within* these groups. However, measuring bias without accounting for the judge's confidence may hide discrimination that emerges only at particular confidence levels. Thus, we additionally condition on a finite collection of confidence bands $\mathcal{I}$, *e.g.*, sub-intervals of $[0, 1]$. Then, for each group, we compute the conditional error within each confidence band and define the judge's per-group bias as the largest of these errors (Tax et al., 2026):

Definition 1 (Per-group bias)**.** *Let* $\mathcal{G}$ *be a collection of measurable group indicators such that* $g : \mathcal{X} \to \{0, 1\}$ *and let* $\mathcal{I}$ *be a finite collection of sub-intervals of* $[0, 1]$*. The per-group bias of* f *is the worst conditional deviation across confidence bands:*

$$\text{bias}_{g,\mathcal{I}}(f) \;:=\; \max_{I \in \mathcal{I}} \left| \mathbb{E}\Big[\, Y - f(X) \;\Big|\; g(X) = 1,\; f(X) \in I \,\Big] \right|$$

We stress that (i) this is a *distributional* notion of bias rather than a bias-specific criterion such as counterfactual swap consistency (Shi et al., 2025); (ii) it applies uniformly to both documented and unspecified biases as it measures conditional deviations over arbitrary bias-induced groups.

Multicalibration. According to Definition 1 the judge is unbiased along g when its residuals $Y - f(X)$ average to zero within that group across every confidence band. This is a within-group calibration constraint, and controlling every group at once maps precisely to *multicalibration* (Hébert-Johnson et al., 2018), which requires f to satisfy the same approximate calibration guarantee simultaneously over a collection $\mathcal{G}$ of identifiable groups.

Definition 2 (Multicalibration)**.** *Let* $\mathcal{G}$ *be a collection of measurable group indicators such that* $g : \mathcal{X} \to \{0, 1\}$ *and let* $\mathcal{I}$ *be a finite collection of sub-intervals of* $[0, 1]$*. For* $g \in \mathcal{G}$ *and* $I \in \mathcal{I}$*, define the multicalibration deviation and the noise scale as* $\Delta_{g,I}(f) := \mathbb{E}\big[\,\big(Y - f(X)\big)\, g(X)\, \mathbb{I}_{\{f(X)\in I\}}\,\big]$*,* $\tau_g(f)^2 := \mathbb{E}\big[\, g(X)\, f(X)\big(1 - f(X)\big)\,\big]$*. Given* $\alpha \geq 0$*, the predictor* f *is* $(\mathcal{G}, \mathcal{I}, \alpha)$-multicalibrated *if*

$$\big|\Delta_{g,I}(f)\big| \;\leq\; \alpha\, \tau_g(f), \qquad \forall\, g \in \mathcal{G},\; \forall\, I \in \mathcal{I}$$

In other terms, $\Delta_{g,I}$ is the expected amount by which the judge over- or under-states its confidence on the instances such that $g(\mathbf{x}) = 1$ and $f(\mathbf{x}) \in I$, while τ_g captures the corresponding scale of sampling variability. We quantify how far a judge is from this requirement through the Multicalibration Error (MCE) (Guy et al., 2025) *i.e.*, the largest group-level Expected Cumulative Calibration Error range (ECCE-R) (Arrieta-Ibarra et al., 2022). ECCE-R orders examples by predicted score, cumulatively sums the residuals $y_i - f(\mathbf{x}_i)$, and measures the range spanned by this cumulative error curve.

Definition 3 (Multicalibration Error)**.** *Consider a collection* $\mathcal{G}$ *of measurable group indicators* $g : \mathcal{X} \to \{0, 1\}$*. For each* $g \in \mathcal{G}$*, let* $\mathcal{D}_g := \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_g}$ *denote the sample restricted to instances satisfying* $g(\mathbf{x}_i) = 1$*, and let* ECCE-R_g *be the ECCE-R computed restricting the sample to* $\mathcal{D}_g$*. Under the null hypothesis of perfect calibration, define the corresponding noise scale as* $\tau_g(f) := \frac{1}{n_g}\sqrt{\sum_{i=1}^{n_g} f(x_i)\big(1 - f(x_i)\big)}$*. The empirical Multicalibration Error of* f *over* $\mathcal{G}$ *is*

$$\text{MCE} := \max_{g \in \mathcal{G}} \frac{\text{ECCE-R}_g(f)}{\tau_g(f)}$$

We report MCE both relative to group prevalence ($\text{MCE}_{\%}$) and in units of the sampling variation expected under perfect calibration (MCE_{σ}). We also report accuracy together with log loss, which is a strictly proper scoring rule (Gneiting and Raftery, 2007). We provide further details in Appendix D.

3 MULTICALIBRATED LARGE LANGUAGE MODELS

Our work tackles the following problem:

Given: a calibration set $\mathcal{D}^{\text{cal}} \subseteq \mathcal{D}$ and, for each pairwise evaluation, a judge confidence score $f_0(X)$ that R_1 is preferred and a representation $\phi : \mathcal{X} \to \mathbb{R}^d$;

Do: Multicalibrate $f_0(X)$, producing a confidence score $f^\star(X)$ that controls the bias of every $g \in \mathcal{G}$ whose error structure is recoverable from $\phi(X)$, *i.e.*, $f^\star$ is $(\mathcal{G}, \mathcal{I}, \alpha)$ multicalibrated.

The core idea of `CaLLM` is to start with uncalibrated confidence scores and apply a post-hoc multicalibration procedure to uncover and correct systematic biases. This has three challenges. **First**, standard multicalibration is typically formulated over features from which relevant groups can be directly identified, whereas we operate on a latent representation in which bias groups may be only implicitly encoded. We show that multicalibration nevertheless bounds the residual bias of any group whose error structure can be recovered from this representation (Section 3.1). **Second**, some bias groups may be known a priori, yet relying solely on the implicit encoding in the representation of their error patterns can weaken the bias-mitigation guarantee. We show that providing these groups directly to the multicalibrator can further reduce their residual error (Section 3.2). **Third**, correcting bias can change the judge's decisions, raising the concern that it may reduce accuracy. We show that accuracy improves whenever the confidence gained on corrected decisions is large enough to outweigh the residual calibration error permitted by approximate multicalibration (Section 3.3).

3.1 BIAS MITIGATION GUARANTEES FOR MULTICALIBRATED LLM JUDGES

Classical multicalibration is typically formulated over an input space in which relevant groups can be directly identified through interpretable covariates. In our setting, however, the inputs are unstructured natural-language responses, and the multicalibrator instead operates on a learned representation $\phi(X) \in \mathbb{R}^d$, which can consist, *e.g.*, of embeddings $E(X)$ and optionally base scores $f_0(X)$. Consequently, the representation need not preserve group membership itself, but may still preserve the systematic error structure associated with that group. This creates a gap between the groups on which biases are defined and the feature space available to the multicalibrator. We therefore characterize how the information preserved by $\phi(X)$ interacts with the judge's systematic prediction error. To this end, we derive an interpretable decomposition of the residual bias in an arbitrary input-space group g. Consider a class $\mathcal{H}$ of functions $h : \mathbb{R}^d \to [0, 1]$, where $h(\phi(X))$ represents a possibly soft representation-space group available to the multicalibrator. Unlike prior work on proxy-group multicalibration (Bharti et al., 2025), our decomposition separates missing label information from missing group information, so a group can be protected when the representation captures its error structure, even if it cannot recover the group. Formally, let us denote with $B_I = \{f(X) \in I\}$ confidence-band membership and let us define the following quantities *i.e.*,

$$\lambda_I^2 = \mathbb{E}\Big[\big(\mathbb{E}[Y \mid X] - f(X) - \mathbb{E}[\mathbb{E}[Y \mid X] - f(X) \mid \phi(X),\, B_I]\big)^2 \mid B_I\Big],$$

$$\varphi_I^2 = \mathbb{E}\Big[\big(g(X) - \mathbb{E}[g(X) \mid \phi(X),\, B_I]\big)^2 \mid B_I\Big]$$

which measure, respectively, the judge error and group information not captured by $\phi(X)$ within band I. Moreover, we define two representation-level components of the judge's error:

$$\chi_I^2 = \mathbb{E}\Big[\big(\mathbb{E}[\mathbb{E}[Y \mid X] - f(X) \mid \phi(X),\, B_I]\big)^2 \mid B_I\Big],$$

$$\psi_I^2 = \mathbb{E}\Big[\big(\mathbb{E}[g(X) - h(\phi(X)) \mid \phi(X),\, B_I]\big)^2 \mid B_I\Big]$$

which capture, respectively, the systematic judge error visible from the representation and the representation-level group-approximation error. We now show how these components bound the bias:

Theorem 1 (Residual bias decomposition)**.** *Let $g \in \mathcal{G}$ and fix $h \in \mathcal{H}$. Assume that f is measurable with respect to $\phi(X)$, and define $\tau_h(f)^2 := \mathbb{E}\big[h(\phi(X))^2 f(X)(1 - f(X))\big]$. Let f be multicalibrated over $(\mathcal{H}, \mathcal{I})$, i.e., for every $h \in \mathcal{H}$ and $I \in \mathcal{I}$, $\big|\mathbb{E}\big[(Y - f(X))h(\phi(X))\mathbb{I}_{\{B_I\}}\big]\big| \le \alpha\tau_h(f)$, with $B_I = \{f(X) \in I\}$ such that $\mathbb{P}(B_I) > 0$. Then, for the fixed $h \in \mathcal{H}$,*

$$\text{bias}_{g,\mathcal{I}}(f) \le \max_{I \in \mathcal{I}} \frac{\alpha\tau_h(f) + \mathbb{P}(B_I)\,(\lambda_I\varphi_I + \chi_I\psi_I)}{\mathbb{P}(g(X) = 1, B_I)}$$

Proof. We provide the proof in Appendix A.1. □

The bound separates three sources of residual bias. The first term, $\alpha\tau_h(f)$, is the calibration violation that remains along the representation-space group h. This is controlled directly by multicalibration. Classical multicalibration procedures iteratively identify residual violations over a function class and update the predictor along them, with each non-trivial correction yielding progress in squared prediction error (Hébert-Johnson et al., 2018). Viewing $\mathcal{H}$ as the multicalibrator's function class over $\phi(X)$, the residual-orthogonality assumption therefore captures the multicalibration machinery. The second term, $\lambda_I\varphi_I$, couples the portions of the judge's error and group membership that are not captured by the representation. Finally, $\chi_I\psi_I$ couples the systematic error visible in representation space with the corresponding group-approximation error.

Importantly, the latter two contributions depend on *products*. Hence, accurate recovery of group membership is not necessary for bias control: $\lambda_I\varphi_I$ can be small if either the relevant error structure or the group structure is well captured by the representation, while $\chi_I\psi_I$ can be small if either little systematic error remains visible in representation space or h accurately captures the relevant group component. Thus, poor input-space group recovery alone does not imply poor bias mitigation, what matters is how the information preserved in the representation interacts with the judge's error.

3.2 The Value of Explicit Bias Specification

Perhaps surprisingly, eliminating the multicalibration term need not mitigate the bias of g. Even under exact residual orthogonality to h, the interaction between the unrecovered error and group components may remain non-zero. The following proposition shows that this term is unavoidable.

Proposition 1 (Necessity and tightness of the unrecoverable interaction)**.** *Assume that $\mathcal{I}$ contains a confidence band including $1/2$, that $\mathcal{H}$ contains the constant function $h \equiv 1/2$, and that $\mathrm{bias}_{g,\mathcal{I}}(f)$ is evaluated only over $I \in \mathcal{I}$ such that $\mathbb{P}(g(X) = 1, f(X) \in I) > 0$. Then, for every $p_g \in (0, 1)$ and every $0 < \rho \leq \frac{1}{2}\min\{p_g, 1 - p_g\}$, there exist a distribution, a group g with $\mathbb{P}(g(X) = 1) = p_g$, a representation $\phi(X)$ and a predictor $f \equiv 1/2$ such that f is perfectly calibrated and the residual is orthogonal to h on every $I \in \mathcal{I}$. Moreover, for every $I \in \mathcal{I}$, $\lambda_I\varphi_I = \rho$, $\chi_I\psi_I = 0$, while*

$$\mathrm{bias}_{g,\mathcal{I}}(f) = \frac{\lambda_I\varphi_I}{p_g} = \frac{\rho}{p_g}$$

Proof. We provide the proof in Appendix A.2. □

Consequently, the term $\lambda_I\varphi_I$ cannot in general be removed from the bound of Theorem 1. However, when a documented bias such as position, verbosity, or family is known, its group indicator can be supplied directly to the multicalibrator. Augmenting the representation with $g(X)$ makes the group exactly identifiable and eliminates both group-side approximation terms, φ_I and ψ_I.

Corollary 1 (What naming a bias is worth)**.** *Let $g \in \mathcal{G}$ be a bias group and consider $\tilde{\phi}(X) = [\phi(X), g(X)]$. Assume that $\mathcal{H}$ contains the coordinate projection $h_g(\tilde{\phi}(X)) = g(X)$. Then, under the setting and assumptions of Theorem 1, $\varphi_I = \psi_I = 0$ for all $I \in \mathcal{I}$ and therefore:*

$$\mathrm{bias}_{g,\mathcal{I}}(f) \leq \max_{I \in \mathcal{I}} \frac{\alpha\tau_g(f)}{\mathbb{P}(g(X) = 1, B_I)}$$

Proof. We provide the proof in Appendix A.3. □

The corollary isolates *what naming a bias is worth*: once $g(X)$ is provided explicitly, both interaction terms in Theorem 1 vanish, leaving only the multicalibration term. Thus, explicit bias specification removes the representation-induced approximation error for that group while preserving the broader protection afforded by the learned representation.

3.3 Predictive Performance Guarantees for Multicalibrated LLM-Judges

Multicalibration should not come at the expense of predictive performance. Decision-theoretic analyses of multicalibrated predictors have previously established general guarantees on downstream decision quality (Baldeschi et al., 2025). In the binary classification setting considered here, we derive an explicit accuracy-improvement bound that separates the gain induced by changed decisions from the residual slack permitted by approximate multicalibration. Our analysis begins denoting the two sets of instances whose decision *changes* after multicalibration with $\mathcal{C}^{\uparrow} := \{\mathbf{x} : f_0(\mathbf{x}) < \frac{1}{2} \leq f^\star(\mathbf{x})\}$ and $\mathcal{C}^{\downarrow} := \{\mathbf{x} : f^\star(\mathbf{x}) < \frac{1}{2} \leq f_0(\mathbf{x})\}$. We denote with $\mathcal{C} := \mathcal{C}^{\uparrow} \cup \mathcal{C}^{\downarrow}$ their union. This is the subset of instances such that the two judges f_0 and $f^\star$ disagree.

Any result on the net accuracy change between the two judges is restricted to this subset of instances. We decompose the predictive performance variation within $\mathcal{C}$ in two interpretable terms. The first term (Γ) is the predicted-utility advantage of the decision rule induced by $f^\star$ over that induced by f_0. It equals the decision margin of instances whose confidences cross the threshold. The second term (ζ) instead accounts for the residual discrepancy between predicted and true utility permitted by approximate multicalibration. Whenever $\Gamma > \zeta$, the net accuracy change is strictly positive.

Theorem 2 (Accuracy improvement under multicalibration)**.** *Suppose that $f^\star$ is $(\mathcal{G}^\star, \mathcal{I}, \alpha)$-multicalibrated, where $\mathcal{I}$ contains the intervals $[0, \frac{1}{2})$ and $[\frac{1}{2}, 1]$. Let $g_-(X) = \mathbb{I}_{\{f_0(X) < \frac{1}{2}\}}$ and $g_+(X) = \mathbb{I}_{\{f_0(X) \geq \frac{1}{2}\}}$ denote the two groups induced by the base judge's hard prediction and let $\{g_-, g_+\} \subseteq \mathcal{G}^\star$. Let $\tau^{\uparrow}$ and $\tau^{\downarrow}$ denote the corresponding multicalibration scale parameters. Then:*

$$\mathrm{Acc}(f^\star) - \mathrm{Acc}(f_0) \;\geq\; \underbrace{\mathbb{E}\left[\left|2f^\star(X) - 1\right| \mathbb{I}_{\{X \in \mathcal{C}\}}\right]}_{\Gamma} - \underbrace{2\,\alpha\left(\tau^{\uparrow} + \tau^{\downarrow}\right)}_{\zeta}$$

Proof. We refer readers to Appendix A.4 for a complete proof. □

The bound reads as a contest between two quantities: the magnitude of the confidence margin for new decisions (Γ), against how much room multicalibration leaves it to be wrong (ζ). Accuracy strictly improves whenever the first exceeds the second, and, if $\Gamma > 0$, in the perfectly multicalibrated limit where $\alpha \to 0$ *any* decision change is a strict gain.

4 Experimental Evaluation

In the following, we address three research questions[2]:

- **Q1.** Does `CaLLM` reduce the biases of the LLM-as-a-judge over existing mitigation strategies without having direct access to the bias groups?
- **Q2.** Does `CaLLM`'s bias mitigation further improve when given access to the bias groups?
- **Q3.** Does `CaLLM` improve over the baselines' predictive performance metrics?

4.1 Experimental Settings

Datasets. We consider five pairwise evaluation datasets from the literature. `RewardBench` (`RWB`) consists of 3k prompt-chosen-rejected trios (Lambert et al., 2025). `MTBench` (`MTB`) consists of 709 pairwise comparisons over 80 questions (Zheng et al., 2023). Each comparison has LLM-generated answers that are judged pairwise by one or multiple human experts. `ArenaExpl` (`AEX`) is a corpus of 53.8k pairwise comparisons extracted from conversational data (Chiang et al., 2024). Here users write the prompt, receive two LLM-generated answers, and act as the judge by expressing a preference. Similarly, `PPE` contains 10.2k prompts in different languages, with users casting a preference vote for one of two options generated by two LLMs (Frick et al., 2025). Finally, `PKU` consists of 73.9k pairwise comparisons over safety-oriented prompts. These combine human demonstrations with LLMs (Ji et al., 2025). We use the more helpful answer as ground truth. More details in Appendix C.

[2]The code is available in the supplementary material and will be published upon acceptance.

Implementation details. We rely on `MCGrad` (Tax et al., 2026) as the default multicalibrator for `CaLLM`, because it can automatically search the supplied feature space and thereby uncover undocumented, latent sources of bias. We ablate the choice of multicalibrator and compare with popular calibration techniques in Appendix E.4. To exploit `MCGrad`'s ability to discover groups automatically, we encode both candidate responses using the open-source `Muse-Glimmer-30B` (Meta, 2026). We provide results with `Qwen3-0.6B` (Team, 2025) in Appendix E.3. We denote the resulting embeddings with $e_1, e_2 \in \mathbb{R}^d$. Following previous work, we represent each comparison through the Principal Component Analysis (PCA) of the difference embedding (Zhang et al., 2026), denoted as $E(X) = \text{PCA}(e_1 - e_2)$. We then feed $\phi(X) = [E(X), f_0(X)]$ to `MCGrad`. The former suppresses components shared by the two responses and emphasizes their relative attributes, which are more directly aligned with the structure of pairwise biases. The latter allows to discover miscalibration concealed within confidence bands $\mathcal{I}$. We provide further details in Appendix B.

Baselines. We compare `CaLLM` against five baselines: (i) the raw verbalized confidence (f_0), which serves as our reference level; (ii) `BPC` (Wang et al., 2024), which queries the judge with both response orderings and averages the two confidences; (iii) `PORTIA` (Li et al., 2024b), which interleaves the two responses in a single prompt such that neither of them occupies a privileged position; (iv) `CalibraEval` (Li et al., 2025b), which queries the judge under swapped orderings and learns a non-parametric, order-preserving mapping that removes the preference for one position; (v) `LenControl` (Dubois et al., 2024), which models the judge's preference as a function of response-length difference and then predicts as if both responses had equal length.

Metrics. For **Q1** and **Q2**, we report the maximum error among the groups induced by the bias-inducing attribute a, which can be seen as a per-documented-bias Multicalibration Error ($\text{MCE}_{\%}$). We report MCE_{σ} in Appendix E.6. We compute and report this quantity separately for each documented bias. For *verbosity* and *family* biases, we set π as the original ordering in which the candidate responses are presented to the judge. This is not suitable for *position* bias, since every comparison necessarily contains a first and a second response. Thus, we define π using the alphabetical ordering of the responses source-model names, enabling meaningful position-based groups. Following prior work (Tax et al., 2026), we report the $\text{MCE}_{\%}$ over random partitions of the representation of X in Appendix E.1. This shows if the multicalibrator can uncover and mitigate errors on unspecified groups. For **Q3**, we report accuracy (ACC) together with the log-loss ($\mathcal{L}_{\log}$) of the judges. For completeness, we report the full results for **Q1**, **Q2**, and **Q3** in Appendix E.6.

Setup. We use `Muse-Glimmer-30B`, a recent open-weight causal language model that achieves competitive performance among models of comparable size across a broad range of reasoning and agentic benchmarks (Meta Superintelligence Lab, 2026), as a judge. We ablate the judge choice with `Qwen 2.5-7B` (Team, 2024) in Appendix E.2. Each instance $X = (Q, R_1, R_2)$ is judged once, extracting the judge's verbalized confidence (Tian et al., 2023) $f_0(X)$ that R_1 is the preferred answer. This choice allows our framework to extend to the black-box setting, where one lacks access to internal model weights. We then either multicalibrate the base scores $f_0(X)$ or apply a baseline method. Concretely, we perform 5-fold cross validation, making sure that different answer pairs for the same prompt always belong to the same fold.

4.2 Experimental Results

Q1: `CaLLM` shows no significant miscalibration in 12 out of 14 evaluations. We analyze whether `CaLLM` reduces the *Position*, *Verbosity* and *Family* biases. Fig. 2 reports the average per-group $MCE_{\%}$ across datasets (see Table 9 for full results). We additionally validate our choice of `MCGrad` as main multicalibration method and compare it with global calibration approaches that also leverage labelled calibration data in Appendix E.4. `CaLLM` outperforms global calibration methods on all 14 dataset-groups pairs, and `MCGrad` is the best multicalibration technique on 11 of 14 of them. In Appendix E.5, we assess whether `CaLLM` can recover groups from the available representation.

We observe that `CaLLM` achieves the lowest error in 11 out of 14 datasets and is the only method that substantially reduces all three biases, with a median reduction with respect to the uncalibrated model f_0 of $\approx 64\%$, $\approx 60\%$, and $\approx 46\%$ for position, verbosity, and family, vs. at most $\approx 36\%$, $\approx 22\%$, and $\approx 31\%$ for the other baselines. Interestingly, we observe the largest gains on larger datasets: *e.g.* on `PKU`, `CaLLM` achieves $\text{MCE}_{\%} \approx 2.8 \pm .2$ for *Position* (second-best `CalibraEval` with

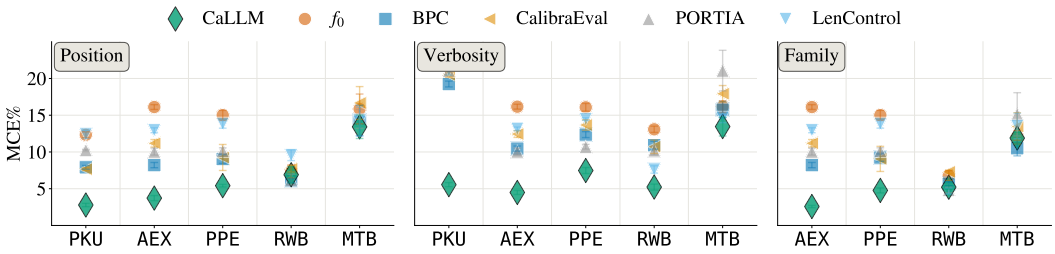


Figure 2: Average per-axis $\text{MCE}_{\%}$ ($\pm$ s.e.) for each method across datasets. PKU doesn't contain the information regarding the LLMs that generated the responses, so we do not report *Family* bias.

Table 1: Comparison between CaLLM and CaLLM+g in terms of $MCE_{\%}$ for documented biases. PKU doesn't contain the information regarding the LLMs that generated the responses, so we do not report *Family* bias.

	PKU		AEX		PPE		RWB		MTB	
	CaLLM	CaLLM +g	CaLLM	CaLLM +g	CaLLM	CaLLM +g	CaLLM	CaLLM +g	CaLLM	CaLLM +g
Position	2.8 ± 0.2	$\mathbf{2.6 \pm 0.1}$	3.7 ± 0.3	$\mathbf{2.5 \pm 0.3}$	5.4 ± 0.2	$\mathbf{3.7 \pm 0.3}$	6.9 ± 1.1	$\mathbf{4.7 \pm 0.5}$	13.4 ± 1.6	$\mathbf{12.5 \pm 1.5}$
Verbosity	5.6 ± 0.3	$\mathbf{2.8 \pm 0.2}$	4.5 ± 0.3	$\mathbf{2.5 \pm 0.2}$	7.5 ± 0.4	$\mathbf{4.0 \pm 0.2}$	5.2 ± 0.4	$\mathbf{4.7 \pm 0.4}$	13.5 ± 1.3	$\mathbf{12.3 \pm 0.9}$
Family	–	–	2.6 ± 0.2	$\mathbf{2.5 \pm 0.2}$	4.8 ± 0.3	$\mathbf{4.6 \pm 0.4}$	5.2 ± 0.3	$\mathbf{4.6 \pm 0.1}$	11.9 ± 0.8	$\mathbf{11.6 \pm 0.2}$

$\text{MCE}_{\%} \approx 7.7 \pm .2$) and $\text{MCE}_{\%} \approx 5.6 \pm .3$ for *Verbosity* (second-best BPC with $\text{MCE}_{\%} \approx 19.2 \pm .4$). Conversely, the three exceptions occur on the two smallest datasets: on RWB, where f_0 shows no strong evidence of *Position* bias ($\text{MCE}_{\sigma} = 3.4$), and two of the three position-specific baselines (BPC and PORTIA) outperform CaLLM on position. Notably, explicitly providing CaLLM with the position groups reverses this ranking (see **Q2**); for *Family* bias, CaLLM is within one standard error of the best method on both RWB and MTB. Notably, CaLLM leaves no strong evidence of miscalibration ($\text{MCE}_{\sigma} < 4.71$) [3] in 12 out of 14 dataset pairs (reference level was 36% for f_0), with the only exceptions occurring for *Verbosity* on PKU and AEX, the two largest datasets, where even small deviations are detectable (see Table 9).

Q2: Naming the groups always lowers the per-bias $\text{MCE}_{\%}$. We quantify *what naming a bias is worth* by comparing CaLLM with an augmented version (CaLLM+g) that adds to $\phi(X)$ the group indicators resulting from the *Position*, *Verbosity* and *Family* biases (see Table 1).

More precisely, we can see that for *Position* bias, CaLLM+g further reduces $\text{MCE}_{\%}$ on all five datasets, with the largest gains on AEX ($\text{MCE}_{\%}$ passes from $\approx 3.7 \pm 0.3$ to $\approx 2.5 \pm 0.3$, a drop of $\approx 33\%$) and RWB ($\text{MCE}_{\%}$ passes from $\approx 6.9 \pm 1.1$ to $\approx 4.7 \pm 0.5$, a drop of $\approx 32\%$).

For *Verbosity* bias, CaLLM+g exhibits even stronger gains, with an average reduction of $\approx 32\%$ across the five datasets. In particular, we observe the largest improvements on PKU ($\approx 49.4\%$), PPE ($\approx 46.7\%$) and AEX ($\approx 43.2\%$), the three largest datasets of our collection. For *Family* bias, we observe smaller improvements, likely due to the fact that CaLLM already performed quite well across all datasets. Still, we can observe some gains on smaller datasets: *e.g.* on RWB $\text{MCE}_{\%}$ passes from $\approx 5.2 \pm 0.3$ to $\approx 4.6 \pm 0.1$. Overall, CaLLM+g completely removes evidence of miscalibration on all datasets by reducing the MCE_{σ} to < 4.71 (compared to 12/14 for CaLLM, see Table 10).

Q3: A CaLLM judge is also a better judge. Table 2 displays the results in terms of predictive performance, reporting both ACC and $\mathcal{L}_{\text{log}}$ for all the methods. We can see that CaLLM achieves the highest accuracy on four out of five datasets (matching the best baseline on the fifth) and the lowest log-loss on all five datasets.

Compared with f_0, ACC increases by ≈ 4.6 percentage points on average, while $\mathcal{L}_{\text{log}}$ decreases by $\approx 24\%$. Consistently with previous results, largest gains are obtained on largest data: on PKU, CaLLM passes from $\approx .63 \pm .01$ to $\approx .69 \pm .01$, a $\approx 9.8\%$ increase; on AEX, CaLLM reaches $\approx .68 \pm .01$ starting from a base model of $\approx .63$, an improvement of $\approx 7\%$. We observe similar

[3] $\text{MCE}_{\sigma} = 4.71$ equals a conservative p-value of 10^{-5} under the null hypothesis of perfect calibration.

Table 2: Average ACC and log-loss $\mathcal{L}_{\log}$ ($\pm$ s.e.) across datasets. CaLLM always matches or improves both metrics over all baselines, showing that multicalibration also improves predictive performance.

	PKU		AEX		PPE		RWB		MTB	
	ACC	$\mathcal{L}_{\log}$	ACC	$\mathcal{L}_{\log}$	ACC	$\mathcal{L}_{\log}$	ACC	$\mathcal{L}_{\log}$	ACC	$\mathcal{L}_{\log}$
CaLLM	**.69 ± .01**	**.57 ± .01**	**.68 ± .01**	**.60 ± .01**	**.67 ± .01**	**.61 ± .01**	**.93 ± .01**	**.21 ± .01**	**.81 ± .01**	**.41 ± .02**
f_0	.63 ± .01	.73 ± .01	.63 ± .01	.76 ± .01	.64 ± .01	.74 ± .01	.86 ± .01	.36 ± .02	.79 ± .01	.49 ± .02
BPC	.64 ± .01	.67 ± .01	.65 ± .01	.69 ± .01	.65 ± .01	.68 ± .01	.88 ± .01	.29 ± .01	.80 ± .01	.43 ± .02
CalibraEval	.64 ± .01	.65 ± .01	.65 ± .01	.69 ± .01	.65 ± .01	.66 ± .01	.88 ± .01	.30 ± .01	.81 ± .01	.43 ± .02
PORTIA	.61 ± .01	.72 ± .01	.65 ± .01	.72 ± .01	.66 ± .01	.70 ± .01	.87 ± .01	.31 ± .02	.77 ± .02	.55 ± .05
LenControl	.64 ± .01	.67 ± .01	.65 ± .01	.69 ± .01	.65 ± .01	.68 ± .01	.88 ± .01	.31 ± .01	.80 ± .01	.43 ± .02

patterns also for $\mathcal{L}_{\log}$, with CaLLM consistently outperforming competitors. The largest gain is on RWB, where $L_{\log}$ drops by 42%, passing from the base level of $\approx .36 \pm .02$ to $\approx .21 \pm .01$. Overall, these results suggest that CaLLM effectively enforces predictive performance gains.

5 Related Work

LLM judges and bias mitigation. LLM judges provide a scalable alternative to human evaluation through pointwise, pairwise, and reference-guided protocols (Liu et al., 2023; Liusie et al., 2024). Their judgments, however, are affected by position, verbosity, model family, and other incidental attributes (Wang et al., 2024; Panickssery et al., 2024; Koo et al., 2024; Chen et al., 2024; Ye et al., 2025). Existing mitigation strategies modify inference through response swapping, interleaving, panels, or decomposed criteria (Wang et al., 2024; Li et al., 2024b; Verga et al., 2024; Chan et al., 2024; Yang et al., 2026); fine-tune evaluators (Kim et al., 2024; Vu et al., 2024; Zhu et al., 2025); or post-process confidence scores (Dubois et al., 2024; Li et al., 2025b). These methods target predefined biases and may require additional inference or annotated data (Li et al., 2025a). Conversely, CaLLM is post-hoc and controls any bias whose error pattern is recoverable from the supplied representation.

Calibration. Classical parametric and non-parametric calibration methods learn a global correction (Vaicenavicius et al., 2019; Platt, 1999; Guo et al., 2017; Kull et al., 2017; Zadrozny and Elkan, 2002; 2001), with extensions to multiclass scores (Kull et al., 2019; Berta et al., 2026). Local calibration refines this incorporating a notion of distance between inputs (Luo et al., 2022; Barbera et al., 2026a; Lin et al., 2023; Valk and Kull, 2023; Barbera et al., 2026b). Multicalibration instead requires calibration simultaneously across groups identified by a hypothesis class (Hébert-Johnson et al., 2018; Jung et al., 2021; Baldeschi et al., 2025; Jin et al., 2025; Tax et al., 2026). This group-wise formulation matches how LLM-judge biases arise and allows CaLLM to provide guarantees beyond a predefined list of biases (Globus-Harris et al., 2023; Noarov and Roth, 2023).

6 Conclusion

We tackled bias in LLM-as-a-judge by casting it as a multicalibration problem. We introduced CaLLM, a post-hoc framework that uses the representation of each comparison to identify and correct groups with systematic residual error. We showed that CaLLM bounds every bias group whose residual structure is recoverable from the representation and that explicitly providing a documented group further improves the corresponding guarantees. We showed that multicalibration improves accuracy when changed-decision margins exceed residual slack, which is always true under perfect multicalibration. Across five benchmark data, CaLLM reduces bias on documented biases while improving accuracy and log loss over the raw judge and bias-specific baselines.

Limitations and future work. Our guarantees require the pattern of target biases to be recoverable from functions of the chosen representation. Biases whose residual structure is not captured may therefore remain uncontrolled. Moreover, our guarantees are population-level and become weaker for rare groups, where calibration errors are also harder to estimate from finite data. Our notion of bias is distributional: CaLLM corrects systematic group-wise discrepancies but does not identify whether the corresponding attributes causally influence the judge, suggesting a complementary connection with counterfactual notions of bias. Finally, we focus on binary pairwise evaluation. We leave extensions to pointwise scoring, multi-way comparisons, and ranking-based LLM evaluation as future work.

AI USE STATEMENT

We used generative AI tools to review and improve the writing and presentation of the mathematical proofs, provide feedback on the experimental methodology, and assist with the implementation of the experiment code. We also used these tools to help prepare Figure 1. We reviewed all AI-assisted text, proofs, code, and figures, verified the resulting arguments and experimental outputs, and take full responsibility for the final content of this work.

ETHICS STATEMENT

This work aims to make LLM-based evaluation less sensitive to systematic biases. However, lower group-wise calibration error should not be interpreted as a guarantee of fairness: `CaLLM` may not protect groups that are absent from the calibration data or whose error structure cannot be recovered from the chosen representation, and it inherits limitations in the human preference labels used for calibration. The method should therefore complement, rather than replace, human oversight in high-stakes settings.

REPRODUCIBILITY STATEMENT

The main text describes the datasets, judge, confidence-elicitation procedure, baselines, metrics, representation, and five-fold evaluation protocol used in our experiments. Complete proofs are provided in the appendix, together with additional experimental results. The implementation and experiment scripts used to produce our results are attached with the submission.

REFERENCES

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*, 2023.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. A survey on llm-as-a-judge. *CoRR*, abs/2411.15594, 2024.

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

Kayla Schroeder and Zach Wood-Doughty. Can you trust LLM judgments? reliability of llm-as-a-judge. *CoRR*, abs/2412.12509, 2024.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *EMNLP*, pages 2757–2791. Association for Computational Linguistics, 2025a.

Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 1944–1953. PMLR, 2018.

Niek Tax, Lorenzo Perini, Fridolin Linder, Daniel Haimovich, Dima Karamshuk, Nastaran Okati, Milan Vojnovic, and Pavlos Athanasios Apostolopoulos. Mcgrad: Multicalibration at web scale. In *KDD (1)*, pages 2470–2481. ACM, 2026.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in llm-as-a-judge. In *IJCNLP-AACL (long papers)*, pages 292–314. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, 2025.

Ido Guy, Daniel Haimovich, Fridolin Linder, Nastaran Okati, Lorenzo Perini, Niek Tax, and Mark Tygert. Measuring multi-calibration. *CoRR*, abs/2506.11251, 2025.

Imanol Arrieta-Ibarra, Paman Gujral, Jonathan Tannen, Mark Tygert, and Cherie Xu. Metrics of calibration for probabilistic predictions. *Journal of Machine Learning Research*, 23(351):1–54, 2022. URL https://arxiv.org/abs/2205.09680.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Beepul Bharti, Mary Versa Clemens-Sewall, Paul H. Yi, and Jeremias Sulam. Multiaccuracy and multicalibration via proxy groups. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Riccardo Colini Baldeschi, Simone Di Gregorio, Simone Fioravanti, Federico Fusco, Ido Guy, Daniel Haimovich, Stefano Leonardi, Fridolin Linder, Lorenzo Perini, Matteo Russo, et al. Multicalibration yields better matchings. *arXiv preprint arXiv:2511.11413*, 2025.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, Lester James V. Miranda, Bill Yuchen Lin, Khyathi Raghavi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, 2025.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning, ICML 2024*, 2024.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios Nikolas Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. How to evaluate reward models for RLHF. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*, 2025.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Alex Qiu, Jiayi Zhou, Kaile Wang, Boxun Li, Sirui Han, Yike Guo, and Yaodong Yang. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, ACL 2025*, 2025.

Meta. Muse spark safety & preparedness report. *CoRR*, abs/2606.12429, 2026.

Qwen Team. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025.

Zilong Zhang, Yi-Ting Hung, Lei Ding, and Chi-Kuang Yeh. Quantifying and auditing LLM evaluation via positive-unlabeled learning. *CoRR*, abs/2606.19057, 2026.

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

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. Split and merge: Aligning position biases in llm-based evaluators. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 11084–11108. Association for Computational Linguistics, 2024b. doi: 10.18653/V1/2024.EMNLP-MAIN.621.

Haitao Li, Junjie Chen, Qingyao Ai, Zhumin Chu, Yujia Zhou, Qian Dong, and Yiqun Liu. Calibraeval: Calibrating prediction distribution to mitigate selection bias in llms-as-judges. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 16537–16552. Association for Computational Linguistics, 2025b. doi: 10.18653/V1/2025.ACL-LONG.808. URL https://doi.org/10.18653/v1/2025.acl-long.808.

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

Meta Superintelligence Lab. Muse-glimmer-30b. Hugging Face model repository, 2026. URL https://huggingface.co/meta-models/Muse-Glimmer-30B. Model card.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *EMNLP*, pages 5433–5442. Association for Computational Linguistics, 2023.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In *EMNLP*, pages 2511–2522. Association for Computational Linguistics, 2023.

Adian Liusie, Potsawee Manakul, and Mark J. F. Gales. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *EACL (1)*, pages 139–151. Association for Computational Linguistics, 2024.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pages 517–545. Association for Computational Linguistics, 2024.

Guiming Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or llms as the judge? A study on judgement bias. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8301–8327. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.474. URL https://doi.org/10.18653/v1/2024.emnlp-main.474.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. Justice or prejudice? quantifying biases in llm-as-a-judge. In *International Conference on Learning Representations*, 2025.

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

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Jinming Yang, Zheng Hu, Chuxian Qiu, Zhenyu Deng, Xinshan Jiao, and Tao Zhou. Quantifying and mitigating self-preference bias of llm judges. *arXiv preprint arXiv:2604.22891*, 2026.

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

Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, and Yun-Hsuan Sung. Foundational autoraters: Taming large language models for better automatic evaluation. In *EMNLP*, pages 17086–17105. Association for Computational Linguistics, 2024.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025.

Juozas Vaicenavicius, David Widmann, Carl R. Andersson, Fredrik Lindsten, Jacob Roll, and Thomas B. Schön. Evaluating model calibration in classification. In *AISTATS*, volume 89 of *Proceedings of Machine Learning Research*, pages 3459–3467. PMLR, 2019.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017.

Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 623–631, 2017.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *KDD*, pages 694–699. ACM, 2002.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *ICML*, pages 609–616. Morgan Kaufmann, 2001.

Meelis Kull, Miquel Perelló-Nieto, Markus Kängsepp, Telmo de Menezes e Silva Filho, Hao Song, and Peter A. Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *NeurIPS*, pages 12295–12305, 2019.

Eugene Berta, David Holzmüller, Michael I. Jordan, and Francis Bach. Structured Matrix Scaling for Multi-Class Calibration. In *AISTATS*, 2026.

Rachel Luo, Aadyot Bhatnagar, Yu Bai, Shengjia Zhao, Huan Wang, Caiming Xiong, Silvio Savarese, Stefano Ermon, Edward Schmerling, and Marco Pavone. Local calibration: metrics and recalibration. In *UAI*, volume 180 of *Proceedings of Machine Learning Research*, pages 1286–1295. PMLR, 2022.

Cesare Barbera, Lorenzo Perini, Giovanni De Toni, Andrea Passerini, and Andrea Pugnana. Multiclass local calibration with the jensen-shannon distance. In *AISTATS*, Proceedings of Machine Learning Research. PMLR, 2026a.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Taking a step back with kcal: Multi-class kernel-based calibration for deep neural networks. In *ICLR*. OpenReview.net, 2023.

Kaspar Valk and Meelis Kull. Assuming locally equal calibration errors for non-parametric multiclass calibration. *Trans. Mach. Learn. Res.*, 2023, 2023.

Cesare Barbera, Lorenzo Perini, Giovanni De Toni, Andrea Passerini, and Andrea Pugnana. Divide et calibra: Multiclass local calibration via vector quantization. *arXiv preprint arXiv:2605.21060*, 2026b.

Christopher Jung, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. Moment multicalibration for uncertainty estimation. In *Conference on Learning Theory*, pages 2634–2678. PMLR, 2021.

Hongyi Henry Jin, Zijun Ding, Dung Daniel Ngo, and Zhiwei Steven Wu. Discretization-free multicalibration through loss minimization over tree ensembles. *arXiv preprint arXiv:2505.17435*, 2025.

Ira Globus-Harris, Varun Gupta, Christopher Jung, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Multicalibrated regression for downstream fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 259–286, 2023.

Georgy Noarov and Aaron Roth. The statistical scope of multicalibration. In *International Conference on Machine Learning*, pages 26283–26310. PMLR, 2023.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *KDD*, pages 2623–2631. ACM, 2019.

Parikshit Gopalan, Michael P. Kim, Mihir Singhal, and Shengjia Zhao. Low-degree multicalibration. In Po-Ling Loh and Maxim Raginsky, editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 3193–3234. PMLR, 2022. URL https://proceedings.mlr.press/v178/gopalan22a.html.

Gianluca Detommaso, Martin Bertran, Riccardo Fogliato, and Aaron Roth. Multicalibration for confidence scoring in llms. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 10624–10641, 2024.

A PROOFS

A.1 PROOF OF THEOREM 1

Theorem 1 (Residual bias decomposition)**.** *Let $g \in \mathcal{G}$ and fix $h \in \mathcal{H}$. Assume that f is measurable with respect to $\phi(X)$, and define $\tau_h(f)^2 := \mathbb{E}\big[h(\phi(X))^2 f(X)(1-f(X))\big]$. Let f be multicalibrated over $(\mathcal{H},\mathcal{I})$, i.e., for every $h \in \mathcal{H}$ and $I \in \mathcal{I}$, $\big|\mathbb{E}\big[(Y-f(X))h(\phi(X))\mathbb{I}_{\{B_I\}}\big]\big| \le \alpha\tau_h(f)$, with $B_I = \{f(X) \in I\}$ such that $\mathbb{P}(B_I) > 0$. Then, for the fixed $h \in \mathcal{H}$,*

$$\mathrm{bias}_{g,\mathcal{I}}(f) \le \max_{I\in\mathcal{I}} \frac{\alpha\tau_h(f) + \mathbb{P}(B_I)\,(\lambda_I\varphi_I + \chi_I\psi_I)}{\mathbb{P}(g(X)=1, B_I)}$$

Proof. Fix $I \in \mathcal{I}$ such that $p_{g,I} := \mathbb{P}(g(X)=1, f(X)\in I) > 0$, and let $B_I := \{f(X)\in I\}$. For compactness, define

$$\eta_X := \mathbb{E}[Y \mid X], \tag{1}$$

and, within the band B_I,

$$\mu_I(\phi(X)) := \mathbb{E}[\eta_X - f(X) \mid \phi(X), B_I], \qquad \gamma_I(\phi(X)) := \mathbb{E}[g(X) \mid \phi(X), B_I]. \tag{2}$$

Since $g(X)$, $f(X)$, and $\mathbb{I}_{\{B_I\}}$ are X-measurable, the tower property gives

$$\mathbb{E}\big[(Y-f(X))g(X)\mathbb{I}_{\{B_I\}}\big] = \mathbb{E}\big[(\eta_X - f(X))g(X)\mathbb{I}_{\{B_I\}}\big]. \tag{3}$$

Similarly, since $h(\phi(X))$ is X-measurable,

$$\mathbb{E}\big[(Y-f(X))h(\phi(X))\mathbb{I}_{\{B_I\}}\big] = \mathbb{E}\big[(\eta_X - f(X))h(\phi(X))\mathbb{I}_{\{B_I\}}\big]. \tag{4}$$

Hence,

$$\mathbb{E}\big[(Y-f(X))\big(g(X)-h(\phi(X))\big)\mathbb{I}_{\{B_I\}}\big] = \mathbb{E}\big[(\eta_X - f(X))\big(g(X)-h(\phi(X))\big)\mathbb{I}_{\{B_I\}}\big].$$

Conditioning on B_I and adding and subtracting $\mu_I(\phi(X))$ and $\gamma_I(\phi(X))$, we obtain

$$\begin{aligned}
&\mathbb{E}\big[(\eta_X - f(X))\big(g(X)-h(\phi(X))\big) \mid B_I\big] \\
&= \mathbb{E}\big[\big(\eta_X - f(X) - \mu_I(\phi(X))\big)\big(g(X)-\gamma_I(\phi(X))\big) \mid B_I\big] \\
&+ \mathbb{E}\big[\mu_I(\phi(X))\big(\gamma_I(\phi(X)) - h(\phi(X))\big) \mid B_I\big].
\end{aligned}$$

Indeed, the remaining cross terms vanish by the tower property. First,

$$\begin{aligned}
&\mathbb{E}\big[\big(\eta_X - f(X) - \mu_I(\phi(X))\big)\big(\gamma_I(\phi(X)) - h(\phi(X))\big) \mid B_I\big] \\
&= \mathbb{E}\Big[\big(\gamma_I(\phi(X)) - h(\phi(X))\big)\mathbb{E}[\eta_X - f(X) - \mu_I(\phi(X)) \mid \phi(X), B_I] \;\Big|\; B_I\Big] = 0,
\end{aligned}$$

because

$$\mu_I(\phi(X)) = \mathbb{E}[\eta_X - f(X) \mid \phi(X), B_I]. \tag{5}$$

Likewise,

$$\begin{aligned}
&\mathbb{E}\big[\mu_I(\phi(X))\big(g(X)-\gamma_I(\phi(X))\big) \mid B_I\big] \\
&= \mathbb{E}\Big[\mu_I(\phi(X))\mathbb{E}[g(X) - \gamma_I(\phi(X)) \mid \phi(X), B_I] \;\Big|\; B_I\Big] = 0.
\end{aligned}$$

because

$$\gamma_I(\phi(X)) = \mathbb{E}[g(X) \mid \phi(X), B_I]. \tag{6}$$

Multiplying by $\mathbb{P}(B_I)$ therefore yields

$$\begin{aligned}
&\mathbb{E}\big[(Y-f(X))\big(g(X)-h(\phi(X))\big)\mathbb{I}_{\{B_I\}}\big] \\
&= \mathbb{P}(B_I)\mathbb{E}\big[\big(\eta_X - f(X) - \mu_I(\phi(X))\big)\big(g(X)-\gamma_I(\phi(X))\big) \mid B_I\big] \\
&\quad + \mathbb{P}(B_I)\mathbb{E}\big[\mu_I(\phi(X))\big(\gamma_I(\phi(X)) - h(\phi(X))\big) \mid B_I\big].
\end{aligned}$$

Apply the triangle inequality:

$$\begin{aligned}&\left|\mathbb{E}\Big[(Y-f(X))\big(g(X)-h(\phi(X))\big)\mathbb{I}_{\{B_I\}}\Big]\right| \\ &\leq \mathbb{P}(B_I)\left|\mathbb{E}\Big[\big(\eta_X - f(X) - \mu_I(\phi(X))\big)\big(g(X)-\gamma_I(\phi(X))\big) \mid B_I\Big]\right| \\ &\quad + \mathbb{P}(B_I)\left|\mathbb{E}\Big[\mu_I(\phi(X))\big(\gamma_I(\phi(X))-h(\phi(X))\big) \mid B_I\Big]\right|.\end{aligned}$$

For the first term Cauchy-Schwarz gives:

$$\begin{aligned}&\left|\mathbb{E}\Big[\big(\eta_X - f(X) - \mu_I(\phi(X))\big)\big(g(X)-\gamma_I(\phi(X))\big) \mid B_I\Big]\right| \\ &\leq \sqrt{\mathbb{E}\Big[\big(\eta_X - f(X) - \mu_I(\phi(X))\big)^2 \mid B_I\Big]}\cdot\sqrt{\mathbb{E}\Big[\big(g(X)-\gamma_I(\phi(X))\big)^2 \mid B_I\Big]} = \lambda_I\varphi_I\end{aligned}$$

For the second term, Cauchy–Schwarz gives:

$$\begin{aligned}&\left|\mathbb{E}\Big[\mu_I(\phi(X))\big(\gamma_I(\phi(X))-h(\phi(X))\big) \mid B_I\Big]\right| \\ &\leq \sqrt{\mathbb{E}[\mu_I(\phi(X))^2 \mid B_I]}\cdot\sqrt{\mathbb{E}\Big[\big(\gamma_I(\phi(X))-h(\phi(X))\big)^2 \mid B_I\Big]} \\ &= \chi_I\psi_I.\end{aligned}$$

Combining the two bounds yields:

$$\left|\mathbb{E}\Big[(Y-f(X))\big(g(X)-h(\phi(X))\big)\mathbb{I}_{\{B_I\}}\Big]\right| \leq \mathbb{P}(B_I)\left(\lambda_I\varphi_I + \chi_I\psi_I\right).$$

Now consider the bias argument Definition 1

$$\begin{aligned}&\mathbb{E}\big[(Y-f(X))g(X)\mathbb{I}_{\{B_I\}}\big] \\ &= \mathbb{E}\big[(Y-f(X))h(\phi(X))\mathbb{I}_{\{B_I\}}\big] \\ &\quad + \mathbb{E}\big[(Y-f(X))\big(g(X)-h(\phi(X))\big)\mathbb{I}_{\{B_I\}}\big]\end{aligned}$$

Using the assumed multicalibration condition and the previous bound,

$$\left|\mathbb{E}\big[(Y-f(X))g(X)\mathbb{I}_{\{B_I\}}\big]\right| \leq \alpha\tau_h(f) + \mathbb{P}(B_I)\left(\lambda_I\varphi_I + \chi_I\psi_I\right) \tag{7}$$

Finally, since g is an indicator,

$$\begin{aligned}&\left|\mathbb{E}\Big[Y - f(X) \mid g(X)=1, B_I\Big]\right| \\ &= \frac{\left|\mathbb{E}\Big[(Y-f(X))g(X)\mathbb{I}_{\{B_I\}}\Big]\right|}{p_{g,I}} \\ &\leq \frac{\alpha\tau_h(f) + \mathbb{P}(B_I)\left(\lambda_I\varphi_I + \chi_I\psi_I\right)}{p_{g,I}}\end{aligned}$$

Taking the maximum over all $I \in \mathcal{I}$ such that $p_{g,I} > 0$ yields:

$$\mathrm{bias}_{g,\mathcal{I}}(f) \leq \max_{I\in\mathcal{I}} \frac{\alpha\tau_h(f) + \mathbb{P}(B_I)\big(\lambda_I\varphi_I + \chi_I\psi_I\big)}{\mathbb{P}(g(X)=1, f(X)\in I)} \tag{8}$$

This proves the claim. □

A.2 Proof of Proposition 1

Proposition 1 (Necessity and tightness of the unrecoverable interaction)**.** *Assume that $\mathcal{I}$ contains a confidence band including $1/2$, that $\mathcal{H}$ contains the constant function $h \equiv 1/2$, and that $\mathrm{bias}_{g,\mathcal{I}}(f)$ is evaluated only over $I \in \mathcal{I}$ such that $\mathbb{P}(g(X) = 1, f(X) \in I) > 0$. Then, for every $p_g \in (0, 1)$ and every $0 < \rho \leq \frac{1}{2}\min\{p_g, 1 - p_g\}$, there exist a distribution, a group g with $\mathbb{P}(g(X) = 1) = p_g$, a representation $\phi(X)$ and a predictor $f \equiv 1/2$ such that f is perfectly calibrated and the residual is orthogonal to h on every $I \in \mathcal{I}$. Moreover, for every $I \in \mathcal{I}$, $\lambda_I \varphi_I = \rho$, $\chi_I \psi_I = 0$, while*

$$\mathrm{bias}_{g,\mathcal{I}}(f) = \frac{\lambda_I \varphi_I}{p_g} = \frac{\rho}{p_g}$$

Proof. Let

$$X \sim \mathrm{Bernoulli}(p_g), \qquad g(X) = X, \qquad \phi(X) = k, \quad \text{with } k \text{ constant.} \tag{9}$$

Set

$$c = \frac{\rho}{p_g(1 - p_g)} \tag{10}$$

and define

$$Y \mid X \sim \mathrm{Bernoulli}\left(\frac{1}{2} + c(X - p_g)\right), \qquad f(X) = \frac{1}{2}, \qquad h(\phi(X)) = \frac{1}{2} \tag{11}$$

The condition

$$\rho \leq \frac{1}{2}\min\{p_g, 1 - p_g\} \tag{12}$$

ensures that the conditional success probability lies in $[0, 1]$. Indeed, it equals

$$\frac{1}{2} + \frac{\rho}{p_g} \quad \text{when } X = 1, \qquad \frac{1}{2} - \frac{\rho}{1 - p_g} \quad \text{when } X = 0 \tag{13}$$

Since $\mathbb{E}[X] = p_g$,

$$\mathbb{E}[Y] = \frac{1}{2} + c\mathbb{E}[X - p_g] = \frac{1}{2} \tag{14}$$

Thus, $f \equiv 1/2$ is perfectly calibrated.

Now fix $I \in \mathcal{I}$. If $\mathbb{P}(f(X) \in I) = 0$, residual orthogonality holds trivially. Otherwise, I contains $1/2$, so $\mathbb{I}_{\{f(X) \in I\}} = 1$ almost surely. Therefore,

$$\mathbb{E}\big[(Y - f(X))h(\phi(X))\mathbb{I}_{\{f(X)\in I\}}\big] = \frac{1}{2}\mathbb{E}\left[Y - \frac{1}{2}\right] = 0$$

For every such positive-probability I,

$$\mathbb{E}[Y - f(X) \mid g(X) = 1, f(X) \in I] = \mathbb{E}[Y \mid X = 1] - \frac{1}{2} = c(1 - p_g) = \frac{\rho}{p_g}$$

Hence,

$$\mathrm{bias}_{g,\mathcal{I}}(f) = \frac{\rho}{p_g} \tag{15}$$

It remains to evaluate the terms in Theorem 1. We have

$$\mathbb{E}[Y \mid X] - f(X) = c(X - p_g) \tag{16}$$

For any I with $\mathbb{P}(f(X) \in I) > 0$, the event $B_I = \{f(X) \in I\}$ occurs almost surely. Since $\phi(X)$ is constant,

$$\mathbb{E}[\mathbb{E}[Y \mid X] - f(X) \mid \phi(X), B_I] = 0, \qquad \mathbb{E}[g(X) \mid \phi(X), B_I] = p_g \tag{17}$$

Consequently,

$$\begin{aligned}
\lambda_I^2 &= c^2\mathbb{E}[(X - p_g)^2] = c^2 p_g(1 - p_g), \\
\varphi_I^2 &= \mathbb{E}[(X - p_g)^2] = p_g(1 - p_g), \\
\chi_I^2 &= 0
\end{aligned}$$

It follows that

$$\lambda_I \varphi_I = c\, p_g(1 - p_g) = \rho, \qquad \chi_I \psi_I = 0 \tag{18}$$

Finally,

$$\mathbb{P}(B_I) = 1, \qquad \mathbb{P}(g(X) = 1, f(X) \in I) = p_g \tag{19}$$

and the multicalibration term is zero by residual orthogonality Therefore, Theorem 1 gives

$$\mathrm{bias}_{g,\mathcal{I}}(f) \leq \frac{\lambda_I \varphi_I}{p_g} = \frac{\rho}{p_g} \tag{20}$$

Since the actual bias equals ρ/p_g, the bound is attained with equality. This concludes the proof. □

A.3 Proof of Corollary 1

Corollary 1 (What naming a bias is worth)**.** *Let $g \in \mathcal{G}$ be a bias group and consider $\tilde{\phi}(X) = [\phi(X), g(X)]$. Assume that $\mathcal{H}$ contains the coordinate projection $h_g(\tilde{\phi}(X)) = g(X)$. Then, under the setting and assumptions of Theorem 1, $\varphi_I = \psi_I = 0$ for all $I \in \mathcal{I}$ and therefore:*

$$\mathrm{bias}_{g,\mathcal{I}}(f) \leq \max_{I \in \mathcal{I}} \frac{\alpha \tau_g(f)}{\mathbb{P}(g(X) = 1, B_I)}$$

Proof. Since $g(X)$ is a coordinate of $\tilde{\phi}(X)$,

$$\mathbb{E}[g(X) \mid \tilde{\phi}(X), B_I] = g(X), \tag{21}$$

so $\varphi_I = 0$. Moreover, choosing the corresponding coordinate projection h_g gives

$$h_g(\tilde{\phi}(X)) = g(X), \tag{22}$$

hence $\psi_I = 0$. Since g is a binary map:

$$h_g(\tilde{\phi}(X))^2 = g(X)^2 = g(X), \tag{23}$$

and therefore:

$$\tau_{h_g}(f)^2 = \mathbb{E}\Big[h_g(\tilde{\phi}(X))^2 f(X)\big(1 - f(X)\big)\Big] = \mathbb{E}\big[g(X) f(X)\big(1 - f(X)\big)\big] = \tau_g(f)^2 \tag{24}$$

Therefore both interaction terms in Theorem 1 vanish, yielding

$$\mathrm{bias}_{g,\mathcal{I}}(f) \leq \max_{I \in \mathcal{I}} \frac{\alpha \tau_g(f)}{\mathbb{P}(g(X) = 1, f(X) \in I)} \tag{25}$$

□

A.4 Proof of Theorem 2

Theorem 2 (Accuracy improvement under multicalibration)**.** *Suppose that $f^\star$ is $(\mathcal{G}^\star, \mathcal{I}, \alpha)$-multicalibrated, where $\mathcal{I}$ contains the intervals $[0, \frac{1}{2})$ and $[\frac{1}{2}, 1]$. Let $g_-(X) = \mathbb{I}_{\{f_0(X) < \frac{1}{2}\}}$ and $g_+(X) = \mathbb{I}_{\{f_0(X) \geq \frac{1}{2}\}}$ denote the two groups induced by the base judge's hard prediction and let $\{g_-, g_+\} \subseteq \mathcal{G}^\star$. Let $\tau^\uparrow$ and $\tau^\downarrow$ denote the corresponding multicalibration scale parameters. Then:*

$$\mathrm{Acc}(f^\star) - \mathrm{Acc}(f_0) \;\geq\; \underbrace{\mathbb{E}\Big[\Big|2f^\star(X) - 1\Big|\, \mathbb{I}_{\{X \in \mathcal{C}\}}\Big]}_{\Gamma} - \underbrace{2\,\alpha\,\Big(\tau^\uparrow + \tau^\downarrow\Big)}_{\zeta}$$

Proof. The proof of Theorem 2 relies on two Lemmas illustrated in what follows. First, we characterize the net predictive performance variation between the uncalibrated and the calibrated model.

Lemma 1. *The variation in predictive performance between the calibrated judge and the base judge, which depends only on the number of instances inside the set $\mathcal{C}$, is:*

$$\begin{aligned}\mathrm{Acc}(f^\star) - \mathrm{Acc}(f_0) = \mathbb{P}(X \in \mathcal{C}^\uparrow)(2\mathbb{P}(Y=1 \mid X \in \mathcal{C}^\uparrow) - 1) \\ + \mathbb{P}(X \in \mathcal{C}^\downarrow)(1 - 2\mathbb{P}(Y=1 \mid X \in \mathcal{C}^\downarrow))\end{aligned} \tag{26}$$

Proof. Given that the correction is bounded, only instances within the set $\mathcal{C}$ can actually change their prediction. Specifically, a multicalibrated judge $f^\star$ gains accuracy for an instance in $\mathcal{C}^\uparrow$ if $Y=1$ and loses it if $Y=0$, while the opposite applies for $\mathcal{C}^\downarrow$:

$$\begin{aligned}\mathrm{Acc}(f^\star) - \mathrm{Acc}(f_0) &= \mathbb{P}(X \in \mathcal{C}^\uparrow)\left(\mathbb{P}(Y=1 \mid X \in \mathcal{C}^\uparrow) - (1 - \mathbb{P}(Y=1 \mid X \in \mathcal{C}^\uparrow))\right) \\ &\quad + \mathbb{P}(X \in \mathcal{C}^\downarrow)\left((1 - \mathbb{P}(Y=1 \mid X \in \mathcal{C}^\downarrow)) - \mathbb{P}(Y=1 \mid X \in \mathcal{C}^\downarrow)\right) \\ &= \mathbb{P}(X \in \mathcal{C}^\uparrow)(2\mathbb{P}(Y=1 \mid X \in \mathcal{C}^\uparrow) - 1) \\ &\quad + \mathbb{P}(X \in \mathcal{C}^\downarrow)(1 - 2\mathbb{P}(Y=1 \mid X \in \mathcal{C}^\downarrow))\end{aligned}$$

□

Notably, the flip rates $\mathbb{P}(Y=1 \mid X \in \mathcal{C}^\uparrow)$ and $\mathbb{P}(Y=1 \mid X \in \mathcal{C}^\downarrow)$ are controlled by the multicalibration property of $f^\star$ which promotes alignment with the true label. We characterize this in the following Lemma which establishes two bounds on the population error of the predictions flips.

Lemma 2. *Suppose that $f^\star$ is $(\mathcal{G}^\star, \mathcal{I}, \alpha)$-multicalibrated, where $\mathcal{I}$ contains the intervals $[0, \frac{1}{2})$ and $[\frac{1}{2}, 1]$. Let $g_-(X) = \mathbb{I}_{\{f_0(X) < \frac{1}{2}\}}$ and $g_+(X) = \mathbb{I}_{\{f_0(X) \geq \frac{1}{2}\}}$ denote the two groups induced by the base judge's hard prediction and let $\{g_-, g_+\} \subseteq \mathcal{G}^\star$. Let $\tau^\uparrow$ and $\tau^\downarrow$ denote the corresponding multicalibration scale parameters. Then:*

$$\mathbb{P}(Y=1 \mid X \in \mathcal{C}^\uparrow) \geq \tfrac{1}{2} - \frac{\alpha\tau^\uparrow}{\mathbb{P}(X \in \mathcal{C}^\uparrow)}, \qquad \mathbb{P}(Y=1 \mid X \in \mathcal{C}^\downarrow) \leq \tfrac{1}{2} + \frac{\alpha\tau^\downarrow}{\mathbb{P}(X \in \mathcal{C}^\downarrow)} \tag{27}$$

Proof. We prove the lower bound on $\mathbb{P}(Y=1 \mid X \in \mathcal{C}^\uparrow)$ in full. The upper bound on $\mathbb{P}(Y=1 \mid X \in \mathcal{C}^\downarrow)$ is symmetric.

By definition $\mathcal{C}^\uparrow = \{\mathbf{x} : f_0(\mathbf{x}) < \frac{1}{2} \leq f^\star(\mathbf{x})\}$. Let $I_+ := \mathbb{I}_{\{f^\star(X) \geq \frac{1}{2}\}}$. Since $g_- := \mathbb{I}_{\{f_0(X) < \frac{1}{2}\}}$:

$$\mathbb{I}_{\{X \in \mathcal{C}^\uparrow\}} \;=\; \mathbb{I}_{\{f_0(X) < \frac{1}{2}\}}\, \mathbb{I}_{\{f^\star(X) \geq \frac{1}{2}\}} \;=\; g_-\, I_+. \tag{28}$$

By assumption $g_- \in \mathcal{G}^\star$ and $I_+ \in \mathcal{I}$, so $\mathbb{I}_{\{X \in \mathcal{C}^\uparrow\}}$ is an admissible group–level-set combination.

Because $f^\star$ is $(\mathcal{G}^\star, \alpha)$-multicalibrated with respect to $\mathcal{I}$, the joint residual over $\mathcal{C}^\uparrow$ is bounded by the scale $\tau^\uparrow$:

$$\left|\mathbb{E}\Big[(Y - f^\star(X))\, \mathbb{I}_{\{X \in \mathcal{C}^\uparrow\}}\Big]\right| \;=\; \left|\mathbb{E}\Big[(Y - f^\star(X))\, g_-\, I_+\Big]\right| \;\leq\; \alpha\, \tau^\uparrow. \tag{29}$$

Considering the lower side of the absolute value, we obtain

$$\mathbb{E}[Y\, \mathbb{I}_{\{X \in \mathcal{C}^\uparrow\}}] - \mathbb{E}[f^\star(X)\, \mathbb{I}_{\{X \in \mathcal{C}^\uparrow\}}] \;=\; \mathbb{E}\big[(Y - f^\star(X))\, \mathbb{I}_{\{X \in \mathcal{C}^\uparrow\}}\big] \;\geq\; -\alpha\, \tau^\uparrow, \tag{30}$$

Since $f^\star(\mathbf{x}) \geq \frac{1}{2}$ over the set $\mathcal{C}^\uparrow$, hence $f^\star(X)\, \mathbb{I}_{\{X \in \mathcal{C}^\uparrow\}} \geq \frac{1}{2}\, \mathbb{I}_{\{X \in \mathcal{C}^\uparrow\}}$. Thus,

$$\mathbb{E}[Y\, \mathbb{I}_{\{X \in \mathcal{C}^\uparrow\}}] \geq \mathbb{E}[f^\star(X)\, \mathbb{I}_{\{X \in \mathcal{C}^\uparrow\}}] - \alpha\, \tau^\uparrow \;\geq\; \tfrac{1}{2}\, \mathbb{P}(X \in \mathcal{C}^\uparrow) - \alpha\, \tau^\uparrow. \tag{31}$$

Assuming $\mathbb{P}(X \in \mathcal{C}^\uparrow) > 0$ and

$$\mathbb{P}(Y=1 \mid X \in \mathcal{C}^\uparrow) = \frac{\mathbb{P}(Y=1, X \in \mathcal{C}^\uparrow)}{\mathbb{P}(X \in \mathcal{C}^\uparrow)} = \frac{\mathbb{E}[Y\, \mathbb{I}_{\{X \in \mathcal{C}^\uparrow\}}]}{\mathbb{P}(X \in \mathcal{C}^\uparrow)} \tag{32}$$

Then, dividing the previous Eq. (31) by $\mathbb{P}(X \in \mathcal{C}^\uparrow)$ yields

$$\mathbb{P}(Y=1 \mid X \in \mathcal{C}^\uparrow) \;\geq\; \tfrac{1}{2} - \frac{\alpha\, \tau^\uparrow}{\mathbb{P}(X \in \mathcal{C}^\uparrow)}, \tag{33}$$

This concludes the proof for the lower bound of $\mathbb{P}(Y = 1 \mid X \in \mathcal{C}^{\uparrow})$. For the upper bound of $\mathbb{P}(Y = 1 \mid X \in \mathcal{C}^{\downarrow})$, we apply the same procedure considering $\mathcal{C}^{\downarrow} = \{\mathbf{x} : f^{\star}(\mathbf{x}) < \frac{1}{2} \leq f_0(\mathbf{x})\}$, $g_+ := \mathbb{I}_{\{f_0(X) \geq \frac{1}{2}\}} \in \mathcal{G}$ and the level set $I_- = \mathbb{I}_{\{f^{\star}(X) < \frac{1}{2}\}} \in \mathcal{I}$, giving $\mathbb{I}_{\{X \in \mathcal{C}^{\downarrow}\}} = g_+ I_-$. □

We can now conclude the proof of the main statement. From Lemma 2, multicalibration on the up-flip set gives:

$$\mathbb{E}[Y\,\mathbb{I}_{\{X \in \mathcal{C}^{\uparrow}\}}] \geq \mathbb{E}[f^{\star}(X)\,\mathbb{I}_{\{X \in \mathcal{C}^{\uparrow}\}}] - \alpha\,\tau^{\uparrow}, \qquad \mathbb{E}[Y\,\mathbb{I}_{\{X \in \mathcal{C}^{\downarrow}\}}] \leq \mathbb{E}[f^{\star}\,\mathbb{I}_{\{X \in \mathcal{C}^{\downarrow}\}}] + \alpha\,\tau^{\downarrow} \tag{34}$$

Using $\mathbb{E}[Y\,\mathbb{I}_{\{X \in \mathcal{C}^{\uparrow}\}}] = \mathbb{P}(X \in \mathcal{C}^{\uparrow})\mathbb{P}(Y = 1 \mid X \in \mathcal{C}^{\uparrow})$ and $\mathbb{E}[\mathbb{I}_{\{X \in \mathcal{C}^{\uparrow}\}}] = \mathbb{P}(X \in \mathcal{C}^{\uparrow})$, the first term of Lemma 1 can be rewritten as:

$$\begin{aligned}
\mathbb{P}(X \in \mathcal{C}^{\uparrow})(2\mathbb{P}(Y = 1 \mid X \in \mathcal{C}^{\uparrow}) - 1) &= 2\,\mathbb{E}[Y\,\mathbb{I}_{\{X \in \mathcal{C}^{\uparrow}\}}] - \mathbb{P}(X \in \mathcal{C}^{\uparrow}) \\
&\geq 2\,\mathbb{E}[f^{\star}(X)\,\mathbb{I}_{\{X \in \mathcal{C}^{\uparrow}\}}] - \mathbb{P}(X \in \mathcal{C}^{\uparrow}) - 2\alpha\,\tau^{\uparrow} \\
&= \mathbb{E}\big[(2f^{\star}(X) - 1)\,\mathbb{I}_{\{X \in \mathcal{C}^{\uparrow}\}}\big] - 2\alpha\,\tau^{\uparrow}.
\end{aligned} \tag{35}$$

Symmetrically, using $\mathbb{E}[Y\,\mathbb{I}_{\{X \in \mathcal{C}^{\downarrow}\}}] = \mathbb{P}(X \in \mathcal{C}^{\downarrow})\mathbb{P}(Y = 1 \mid X \in \mathcal{C}^{\downarrow})$ and $\mathbb{E}[\mathbb{I}_{\{X \in \mathcal{C}^{\downarrow}\}}] = \mathbb{P}(X \in \mathcal{C}^{\downarrow})$, the second term is:

$$\begin{aligned}
\mathbb{P}(X \in \mathcal{C}^{\downarrow})(1 - 2\mathbb{P}(Y = 1 \mid X \in \mathcal{C}^{\downarrow})) &= \mathbb{P}(X \in \mathcal{C}^{\downarrow}) - 2\,\mathbb{E}[Y\,\mathbb{I}_{\{X \in \mathcal{C}^{\downarrow}\}}] \\
&\geq \mathbb{P}(X \in \mathcal{C}^{\downarrow}) - 2\,\mathbb{E}[f^{\star}(X)\,\mathbb{I}_{\{X \in \mathcal{C}^{\downarrow}\}}] - 2\alpha\,\tau^{\downarrow} \\
&= \mathbb{E}\big[(1 - 2f^{\star}(X))\,\mathbb{I}_{\{X \in \mathcal{C}^{\downarrow}\}}\big] - 2\alpha\,\tau^{\downarrow}.
\end{aligned} \tag{36}$$

Substituting these quantities into Lemma 1,

$$\mathrm{Acc}(f^{\star}(X)) - \mathrm{Acc}(f_0) \;\geq\; \mathbb{E}\big[(2f^{\star}(X) - 1)\,\mathbb{I}_{\{X \in \mathcal{C}^{\uparrow}\}}\big] + \mathbb{E}\big[(1 - 2f^{\star}(X))\,\mathbb{I}_{\{X \in \mathcal{C}^{\downarrow}\}}\big] \;-\; 2\alpha\,(\tau^{\uparrow} + \tau^{\downarrow}). \tag{37}$$

On $\mathcal{C}^{\uparrow}$ we have $f^{\star}(X) \geq \frac{1}{2}$, so $2f^{\star}(X) - 1 = |2f^{\star}(X) - 1| \geq 0$; on $\mathcal{C}^{\downarrow}$ we have $f^{\star}(X) < \frac{1}{2}$, so $1 - 2f^{\star}(X) = |2f^{\star}(X) - 1| > 0$. Moreover, since the two sets are disjoint, $\mathbb{I}_{\{X \in \mathcal{C}^{\uparrow}\}} + \mathbb{I}_{\{\mathcal{C}^{\downarrow}\}} = \mathbb{I}_{\{X \in \mathcal{C}^{\uparrow} \cup \mathcal{C}^{\downarrow}\}}$. Therefore

$$\mathbb{E}\Big[(2f^{\star}(X) - 1)\,\mathbb{I}_{\{X \in \mathcal{C}^{\uparrow}\}}\Big] + \mathbb{E}\Big[(1 - 2f^{\star}(X))\,\mathbb{I}_{\{X \in \mathcal{C}^{\downarrow}\}}\Big] \;=\; \mathbb{E}\Big[\Big|2f^{\star}(X) - 1\Big|\,\mathbb{I}_{\{X \in \mathcal{C}^{\uparrow} \cup \mathcal{C}^{\downarrow}\}}\Big] \;=\; \Gamma \tag{38}$$

with $\Gamma \geq 0$ as the expectation of a nonnegative integrand. If we then combine these with Eq. (37), we obtain:

$$\mathrm{Acc}(f^{\star}(X)) - \mathrm{Acc}(f_0) \;\geq\; \Gamma \;-\; 2\alpha\,(\tau^{\uparrow} + \tau^{\downarrow}). \tag{39}$$

Because $\Gamma \geq 0$, the right-hand side is positive whenever $\Gamma > 2\alpha(\tau^{\uparrow} + \tau^{\downarrow})$, giving $\mathrm{Acc}(f^{\star}(X)) > \mathrm{Acc}(f_0)$. In case of perfect multicalibration $\alpha \to 0$, the bound reduces to $\mathrm{Acc}(f^{\star}(X)) - \mathrm{Acc}(f_0) \geq \Gamma$, strict in presence of $\Gamma > 0$. □

B Implementation Details

In principle, `CaLLM` is agnostic to the choice of multicalibration strategy. In practice, however, we rely on `MCGrad` (Tax et al., 2026) as our default multicalibrator because it can automatically search the supplied feature space and thereby uncover undocumented, latent sources of bias. We note that the guarantee in Theorem 2 additionally requires the multicalibrator to control violations defined jointly by confidence intervals and the base-decision groups g_- and g_+. This requirement is naturally satisfied by `MCGrad`, which searches for calibration violations using both the supplied features and the base confidence score. To exploit `MCGrad`'s ability to discover groups automatically, we encode both candidate responses using the open source `Muse-Glimmer-30B` (Meta, 2026) and denote their embeddings with $e_1, e_2 \in \mathbb{R}^d$. Following previous work (Zhang et al., 2026), we represent each comparison through the Principal Component Analysis (PCA) of the difference embedding, *i.e.*, $E(X) := \mathrm{PCA}(e_1 - e_2)$. Such contrastive representation suppresses components shared by the two responses and emphasizes their relative attributes, which are more directly aligned with the structure

of pairwise biases. When PCA is then applied to the difference embedding, the retained directions summarize variation in the relation between the two candidates rather than variation in their absolute representation. This choice is also consistent with Theorem 1: the multicalibrator can only control biases whose error pattern is captured by the supplied representation. A contrastive representation is therefore desirable as it better captures relative differences between the candidate responses.

Hyperparameter tuning. We perform five-fold cross validation. For each held-out fold, the remaining four folds form the calibration set on which we tune and fit `CaLLM` and the baselines. We assign all comparisons that share a question identifier or an identical prompt to the same fold, so that no prompt is seen both during fitting and at test time. Hyperparameters are tuned as follows for `CaLLM` and all baselines. We split the calibration set at random into 80% for training and 20% for validation, again keeping all comparisons of a question on the same side. For each method, dataset and held-out fold, we run 40 trials of Bayesian optimisation starting from the method's default configuration, score each trial by its validation log loss, and refit the best configuration on the whole calibration set. Specifically, for `CaLLM` we tune jointly the number $k \in \{4, 8, 16, 32\}$ of principal components retained in $E(X)$ with the hyperparameters of `MCGrad` (Tax et al., 2026), using its own tuning routine over the default search space. For the tunable baselines we use Optuna (Akiba et al., 2019) over the following hyperparamters (all searched on a log scale, with defaults in parentheses): for `CalibraEval`, we tune the anti-collapse weight $\lambda \in [0.01, 0.9]$ (0.1), the learning rate in $[10^{-3}, 0.3]$ (0.05) and the number of iterations in $\{50, \ldots, 2000\}$ (500); for `LenControl`, the inverse L_2 strength $C \in [10^{-3}, 10^{3}]$ (1) and the minimum model frequency in $\{1, \ldots, 60\}$ (10) ; for `IGLB`, the number of bins $M \in \{3, \ldots, 100\}$ (20), the maximum number of iterations in $\{100, \ldots, 2000\}$ (500) and the tolerance $\epsilon \in [10^{-4}, 0.2]$ (0.01). The other baselines do not have any hyperparameter to tune. The selected configurations are fold-specific and available in the supplementary material.

C Data

We use five pairwise evaluation benchmarks from the literature. `RewardBench` (`RWB`) (Lambert et al., 2025) consists of 3.0k prompt-chosen-rejected trios belonging to four different macro-categories: *chat* (strong vs. weak answers), *chat-hard* (tricky and adversarial answers), *safety* (refuse dangerous content and avoid over-refusals) and *reasoning* (coding and math problems). Answers are either LLM-generated (coming from 44 different models) or crafted manually, and the ground truth label is a manually verified correctness judgment rather than a human vote. `MTBench` (`MTB`) (Zheng et al., 2023) consists of 3.3k human votes on pairwise comparisons over 80 questions from eight categories (writing, roleplay, reasoning, math, coding, extraction, stem, humanities), with answers generated by six different LLMs (GPT-4, GPT-3.5, Claude-v1, Vicuna-13B, Alpaca-13B, and LLaMA-13B) and judged by 65 human annotators. We aggregate the individual votes into one comparison per (question, model pair) by majority vote. We further discard ties, i.e., cases where the annotators split evenly or could not identify a preferable option, and second-turn data, i.e., each question has two conversational turns, but we keep only the first one since each model's second-turn answer is conditioned on its own first turn answer. This leaves 709 pairwise comparisons. `ArenaExpl` (`AEX`) (Chiang et al., 2024) is a corpus of 106k pairwise comparisons extracted from conversational data where users write the prompt, receive two LLM-generated answers (coming from 55 different models), and act as the judge by expressing a preference. Prompts are written in different languages. We again discard ties (*i.e.* cases where the user does not prefer any of the two options, or where both responses are bad) and keep only single-turn conversations. This leaves 53.8k pairwise comparisons. Similarly, `PPE` (Frick et al., 2025) contains 16k prompts in different languages of users who cast the preference vote for one out of two options generated by two LLMs from a pool of 20 models. After discarding ties, the corpus amounts to 10.2k pairwise comparisons. Finally, `PKU` (Ji et al., 2025) consists of 82.1k pairwise comparisons over safety-oriented prompts generated combining human demonstrations with LLMs. Each pair carries two human preference labels: which response is safer and which is more helpful; we use the latter as ground truth. No ties are allowed. Both responses are generated by an LLM from the Alpaca-family (3 models in total). We use the *train* split of the original dataset which consists of 73.9k pairwise comparisons.

Note that family bias can only be measured on comparisons where exactly one of the two answers comes from the judge's own model family. For `Muse-Glimmer-30B` this is the case in all

datasets except PKU, where both answers of a pair are generated by the same model. We consider answers from Llama models as the same family as Muse, since both are developed by Meta. For Qwen2.5-7B-Instruct, only PPE and AEX contain answers generated by the same model family, so family is not measurable on MTB, RWB and PKU.

D METRICS

ECCE. Expected Cumulative Calibration Error (ECCE) (Arrieta-Ibarra et al., 2022) was first introduced as a reliable parameter free alternative to binning-based calibration metrics. ECCE computes the cumulative calibration residuals:

$$C_k(f) := \frac{1}{n} \sum_{i=1}^{k} \left(y_{(i)} - f(\mathbf{x}_{(i)}) \right), \qquad k = 1, \ldots, n \tag{40}$$

where $f(\mathbf{x}_{(1)}) \leq f(\mathbf{x}_{(2)}) \leq \cdots \leq f(\mathbf{x}_{(n)})$ denote the ordered values of f over a dataset.

Definition 4. ECCE-R *is the range of the empirical cumulative calibration error defined as:*

$$\text{ECCE-R} := \max_{0 \leq k \leq n} C_k(f) - \min_{0 \leq k \leq n} C_k(f). \tag{41}$$

Thus, ECCE-R measures the full range of the cumulative residual curve and therefore captures the largest cumulative calibration error over any contiguous score interval (Arrieta-Ibarra et al., 2022).

MCE. The *Multicalibration* property of a model is evaluated via the multicalibration Error (MCE). This relies on the Expected Cumulative Calibration Error (ECCE) (Arrieta-Ibarra et al., 2022). The Multicalibration Error (MCE) is then the largest standardized ECCE-R across all considered groups and confidence bands (Guy et al., 2025; Tax et al., 2026).

Definition 3 (Multicalibration Error)**.** *Consider a collection $\mathcal{G}$ of measurable group indicators $g : \mathcal{X} \to \{0, 1\}$. For each $g \in \mathcal{G}$, let $\mathcal{D}_g := \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_g}$ denote the sample restricted to instances satisfying $g(\mathbf{x}_i) = 1$, and let ECCE-R_g denote the ECCE-R computed restricting the sample to $\mathcal{D}_g$. Under the null hypothesis of perfect calibration, define the corresponding noise scale as $\tau_g(f) := \frac{1}{n_g} \sqrt{\sum_{i=1}^{n_g} f(x_i)\big(1 - f(x_i)\big)}$. The empirical Multicalibration Error of f over $\mathcal{G}$ is*

$$\text{MCE} := \max_{g \in \mathcal{G}} \frac{\text{ECCE-R}_g(f)}{\tau_g(f)} \tag{42}$$

The normalization accounts for the different sampling variability of groups of different sizes, while the outer maximum measures the strongest statistical evidence of group-wise miscalibration.

Log loss. We additionally report the binary log loss as a measure of probabilistic predictive performance:

$$\text{LogLoss}(f) := -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \log f(\mathbf{x}_i) + (1 - y_i) \log\big(1 - f(\mathbf{x}_i)\big) \right]. \tag{43}$$

Log loss is a strictly proper scoring rule (Gneiting and Raftery, 2007), meaning that its expected value is uniquely minimized when $f(\mathbf{x}) = \mathbb{P}(Y = 1 \mid X = \mathbf{x})$. It therefore rewards probabilistic predictions that are both well calibrated and informative, while strongly penalizing confident incorrect predictions. Accordingly, we use log loss as a measure of the overall quality of the judge's probabilistic predictions, with lower values indicating better predictive performance. A value of zero is attained only when every observed outcome is assigned probability one for the correct class.

E ADDITIONAL EXPERIMENTS

E.1 RESULTS ON SPECIFIED AND UNSPECIFIED GROUPS

We complement the per-bias analysis of Q1 with two additional analyses. First, we consider the *specified* biases jointly (*i.e.* position, verbosity and family): the MCE is computed over every position,

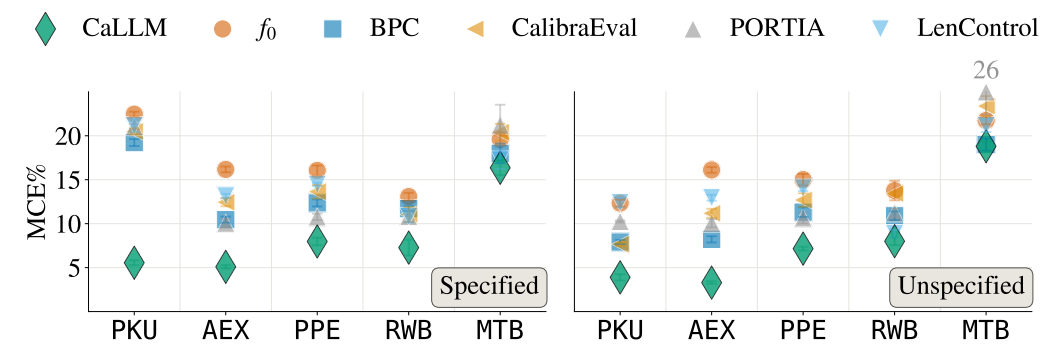


Figure 3: Average per-axis $\text{MCE}_{\%}$ ($\pm$ s.e.) on the *specified* and *unspecified* groups for each method across datasets.

verbosity and family group together with all their pairwise and three-way intersections. Second, we consider *unspecified* groups. Following Tax et al. (2026), we pass the concatenated embeddings $[e_Q, e_1, e_2]$ of the prompt and the two responses to the MCE as numerical segment features to create 1000 groups. These groups depend on neither labels nor predictions and are identical for all methods.

Accounting for interactions between biases increases `CaLLM`'s advantage. Figure 3 (left) reports the joint $\text{MCE}_{\%}$ over the *specified* groups and their intersections. `CaLLM` attains the lowest $\text{MCE}_{\%}$ on all five datasets, whereas it is the best method on 11 of 14 (dataset-bias) pairs when each bias is scored alone. A baseline that targets one specific bias does not mitigate other biases as well; *e.g.* on `RWB`, `BPC` has the lowest position error (6.10 ± 0.26) but a joint error of 11.71 ± 0.27, whereas `CaLLM` reaches 7.27 ± 0.89.

`CaLLM` also mitigates the bias on unspecified groups. Figure 3 (right) reports the $\text{MCE}_{\%}$ on the *unspecified* groups. `CaLLM` attains the lowest $\text{MCE}_{\%}$ on four of five datasets and lowers that of f_0 by a median of $\approx$52%, more than twice the best baseline (`BPC`, $\approx$25%). Against the best baseline on each dataset, it lowers the error by 60% on `AEX`, 49% on `PKU`, 33% on `PPE`, and 10% on `RWB`. On `MTB`, its gain over `BPC` (1%) is within one standard error. Thus, `CaLLM` is also effective at mitigating the bias for unspecified groups.

E.2 Judge ablation: results with a different judge

We repeat the experiments from Section 4 using `Qwen 2.5-7B-Instruct` as the judge, taking its verbalized confidence as f_0.

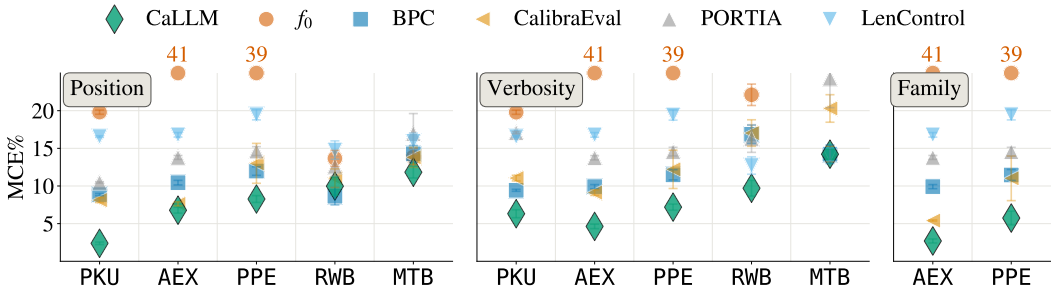


Figure 4: Average per-group $\text{MCE}_{\%}$ ($\pm$ s.e.) for `CaLLM` and the mitigation baselines across datasets with `Qwen 2.5-7B-Instruct` used as the judge. The self-enhancement axis contains `AEX` and `PPE` only: `PKU`, `RWB` and `MTB` do not contain `qwen`-family responses.

Table 3: $MCE_{\%}$ per bias group (± 1 s.e.) for CaLLM and CaLLM $+g$ using the Qwen 2.5-7B-Instruct judge. The *family* axis contains results for AEX and PPE only: PKU, RWB and MTB contain no qwen-family response.

	PKU		AEX		PPE		RWB		MTB	
	CaLLM	CaLLM+g	CaLLM	CaLLM+g	CaLLM	CaLLM+g	CaLLM	CaLLM+g	CaLLM	CaLLM+g
Position	**2.4 ± 0.2**	**2.4 ± 0.2**	6.8 ± 0.4	**2.4 ± 0.3**	8.3 ± 0.5	**3.7 ± 0.5**	10.0 ± 1.2	**7.9 ± 1.1**	**11.8 ± 0.8**	12.0 ± 0.8
Verbosity	6.3 ± 0.5	**2.8 ± 0.2**	4.6 ± 0.3	**2.3 ± 0.1**	7.2 ± 0.4	**4.0 ± 0.3**	9.7 ± 1.0	**7.4 ± 0.8**	14.4 ± 0.7	**13.5 ± 1.1**
Family	–	–	2.7 ± 0.3	**2.1 ± 0.2**	5.7 ± 0.9	**4.7 ± 0.9**	–	–	–	–

Table 4: Average accuracy and log-loss (± 1 s.e.) across datasets using Qwen 2.5-7B-Instruct as judge.

	PKU		AEX		PPE		RWB		MTB	
	Acc	$\mathcal{L}_{\log}$	Acc	$\mathcal{L}_{\log}$	Acc	$\mathcal{L}_{\log}$	Acc	$\mathcal{L}_{\log}$	Acc	$\mathcal{L}_{\log}$
CaLLM	**.68 ± .01**	**.60 ± .01**	**.61 ± .01**	**.65 ± .01**	**.61 ± .01**	**.66 ± .01**	**.84 ± .01**	**.37 ± .01**	**.80 ± .01**	**.45 ± .02**
f_0	.66 ± .01	.72 ± .01	.58 ± .01	.88 ± .01	.59 ± .01	.87 ± .01	.80 ± .01	.50 ± .02	**.80 ± .01**	.51 ± .02
BPC	.66 ± .01	.65 ± .01	.59 ± .01	.71 ± .01	.60 ± .01	.72 ± .01	.80 ± .01	.42 ± .01	**.80 ± .01**	.46 ± .02
CalibraEval	.66 ± .01	.63 ± .01	.59 ± .01	.68 ± .01	.60 ± .01	.69 ± .01	.80 ± .01	.43 ± .01	**.80 ± .01**	.47 ± .02
PORTIA	.59 ± .01	.71 ± .01	.59 ± .01	.75 ± .01	.60 ± .01	.74 ± .01	.76 ± .01	.50 ± .01	.70 ± .03	.59 ± .06
LenControl	.66 ± .01	.64 ± .01	.59 ± .01	.69 ± .01	.60 ± .01	.69 ± .01	.80 ± .01	.45 ± .01	**.80 ± .01**	.46 ± .02

Q1. The bias mitigation transfers to a different judge. Figure 4 reports the per-group $MCE_{\%}$ ($\pm$ s.e.) for CaLLM and the Fig. 2 baselines when Qwen 2.5-7B-Instruct is used as the judge. CaLLM attains the lowest $MCE_{\%}$ on 10 of the 12 (dataset, bias group) pairs: four out of five on *Position*, four out of five on *Verbosity*, and two out of two on *Family*. Even when the verbalized confidences come from a different judge, CaLLM still mitigates the bias effectively.

Q2. Declaring the bias groups helps regardless of the judge. Table 3 repeats the CaLLM vs. CaLLM $+g$ comparison using the Qwen 2.5-7B-Instruct judge to obtain the verbalized confidences. As in Q2, giving CaLLM the explicit bias axes further reduces the bias on most (dataset, bias group) pairs.

Q3. CaLLM improves the predictive performance also when Qwen 2.5-7B-Instruct is used as judge. Table 4 reports the average accuracy and log-loss for CaLLM and the baselines on each dataset. CaLLM improves on every baseline in both metrics and on every dataset, except on MTB where it ties with other baselines. Multicalibrating Qwen 2.5-7B-Instruct therefore improves predictive performance, and the main empirical findings therefore also hold for this judge.

E.3 Embedder ablation: results using a separate open-source embedder

We refit CaLLM unchanged with the embeddings obtained from an off-the-shelf embedder, *i.e.* Qwen3-Embedding-0.6B, to show that the method also applies to judges whose hidden states are not accessible, such as API-only models.

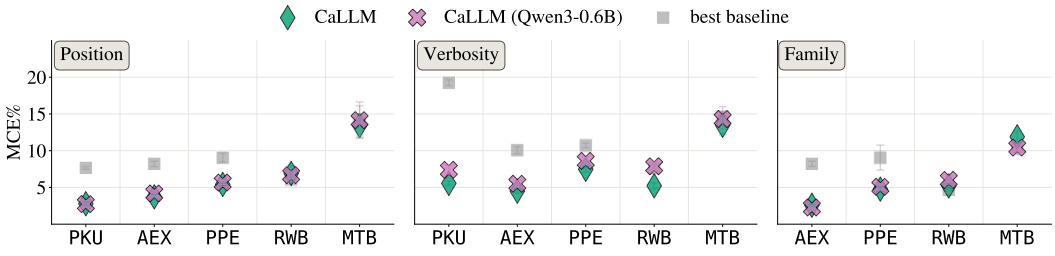


Figure 5: Average per-group $MCE_{\%}$ (± 1 standard error) for CaLLM (embeddings from the judge's last hidden state), CaLLM(Qwen3-0.6B) (embeddings from Qwen3-Embedding-0.6B) and, as a reference, the *best* of the Fig. 2 baselines across datasets. Family is undefined on PKU, since it contains no response from the judge's own family.

Table 5: $\text{MCE}_{\%}$ per bias group (± 1 standard error) for CaLLM and CaLLM+g across datasets using Qwen3-Embedding-0.6B embeddings. Self-enhancement is undefined on PKU because the dataset does not contain muse-family responses.

	PKU		AEX		PPE		RWB		MTB	
	CaLLM	CaLLM+g	CaLLM	CaLLM+g	CaLLM	CaLLM+g	CaLLM	CaLLM+g	CaLLM	CaLLM+g
Position	2.7 ± 0.2	$\mathbf{2.2 \pm 0.1}$	4.2 ± 0.3	$\mathbf{2.1 \pm 0.2}$	5.7 ± 0.3	$\mathbf{4.1 \pm 0.2}$	6.7 ± 1.0	$\mathbf{5.6 \pm 0.5}$	14.2 ± 2.5	$\mathbf{11.9 \pm 1.5}$
Verbosity	7.4 ± 0.6	$\mathbf{2.4 \pm 0.1}$	5.5 ± 0.3	$\mathbf{2.3 \pm 0.2}$	8.6 ± 0.5	$\mathbf{4.3 \pm 0.1}$	7.9 ± 0.7	$\mathbf{5.2 \pm 0.5}$	14.3 ± 1.7	$\mathbf{11.7 \pm 1.4}$
Family	–	–	2.3 ± 0.1	$\mathbf{2.2 \pm 0.1}$	$\mathbf{5.1 \pm 0.3}$	$\mathbf{5.1 \pm 0.3}$	$\mathbf{6.0 \pm 0.6}$	6.6 ± 0.7	$\mathbf{10.4 \pm 0.8}$	11.1 ± 1.7

Table 6: Average accuracy and log-loss ($\pm$ s.e.) for CaLLM and its ablated variant CaLLM(Qwen3-0.6B) across datasets. *best baseline* is the strongest of the Fig. 2 baselines on each (dataset, metric) pair.

	PKU		AEX		PPE		RWB		MTB	
	ACC	$\mathcal{L}_{\log}$	ACC	$\mathcal{L}_{\log}$	ACC	$\mathcal{L}_{\log}$	ACC	$\mathcal{L}_{\log}$	ACC	$\mathcal{L}_{\log}$
CaLLM	**.69 ± .01**	**.57 ± .01**	**.68 ± .01**	**.60 ± .01**	**.67 ± .01**	**.61 ± .01**	**.93 ± .01**	**.21 ± .01**	**.81 ± .01**	**.41 ± .02**
CaLLM (Qwen3-0.6B)	**.69 ± .01**	.58 ± .01	.66 ± .01	.61 ± .01	.66 ± .01	.62 ± .01	.91 ± .01	.27 ± .01	**.81 ± .01**	.42 ± .02
best baseline	.64 ± .01	.65 ± .01	.65 ± .01	.69 ± .01	.66 ± .01	.66 ± .01	.88 ± .01	.29 ± .01	.81 ± .01	.43 ± .02

CaLLM **does not need the judge's hidden states to mitigate the bias.** We compare CaLLM (*i.e.* the embeddings are the judge's last hidden state) with its ablated variant CaLLM(Qwen3-0.6B), which takes them from the off-the-shelf sentence embedder (Qwen3-Embedding-0.6B). As a reference we additionally report the *best* of the Fig. 2 baselines on that (dataset, bias group) pair.

Figure 5 reports the per-group $\text{MCE}_{\%}$ ($\pm$ standard error) for CaLLM, its ablated variant and the *best* baseline. On *Position* and *Family* the two variants perform very similarly on nearly every dataset, and where they differ most (*i.e. Family* on MTB) the ablated variant is the better one (10.4 ± 0.8 vs. 11.9 ± 0.8). *Verbosity* degrades on all five datasets, plausibly because answer length is a property that the judge's own states encode directly. Even so, CaLLM(Qwen3-0.6B) still beats the *best baseline* on 10 of the 14 (dataset, bias group) pairs, including four of the five on verbosity. As for CaLLM, its losses concentrate on the two smallest datasets: RWB (all three groups, by at most 1.3 points) and *Position* on MTB, where it is behind the best baseline by only 0.03 points. These results show that CaLLM's bias reduction does not require access to the judge's internal representations.

Declaring the bias groups still helps under an off-the-shelf embedder. Table 5 repeats the CaLLM vs. CaLLM+g comparison of Q2 with the embeddings taken from Qwen3-Embedding-0.6B. Handing the reference the three group indicators lowers the $\text{MCE}_{\%}$ on every dataset for *Position* and *Verbosity*, and the gain is largest on verbosity: this falls from 7.4 ± 0.6 to 2.4 ± 0.1 on PKU and from 8.6 ± 0.5 to 4.3 ± 0.1 on PPE. For *Family* the change is always within one standard error, in either direction. The value of declaring a bias axis is therefore unchanged by the embedder.

Using an off-the-shelf embedder still improves the predictive performance against the baselines. Table 6 reports accuracy and log-loss for CaLLM, CaLLM(Qwen3-0.6B) and the *best* Q1 baseline for each (dataset, metric) pair. The two variants are close on four of the five datasets, while on RWB the judge's own states buy a real margin (.927 vs. .911 accuracy, .203 vs. .264 log-loss). Crucially, both variants beat the strongest baseline on all ten (dataset, metric) pairs, so CaLLM still outperforms every baseline even when using an off-the-shelf embedder.

E.4 Ablation study: multicalibration method

In this experiment, we evaluate alternative multicalibration strategies, all fitted to the same base scores f_0, validating our choice of MCGrad as multicalibrator (still denoted as CaLLM in both tables and figures for coherence with the main text). We emphasize that CaLLM is agnostic to the choice of multicalibrator and can be used with any multicalibration procedure. We consider two *multicalibration* baselines, which correct the scores within groups: LOGR, a group-conditional unbiased logistic regression (Gopalan et al., 2022), and IGLB, iterative grouped linear binning (Detommaso et al., 2024). Like CaLLM, these baselines have no access to the bias groups, so we build their groups with the clustering strategy of Detommaso et al. (2024). That is, we embed each comparison (the prompt followed by both responses) with the same encoder as CaLLM, reduce the embedding to 20 dimensions with UMAP, and fit a Gaussian mixture on the reduced embedding together with f_0, choosing the number of components in $\{2, \ldots, 20\}$ by BIC. Each component with at least 30

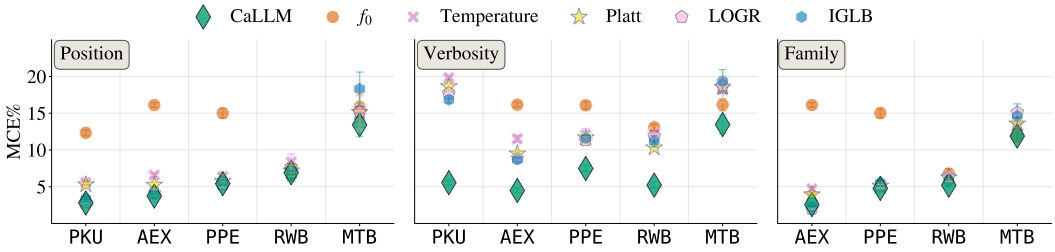


Figure 6: Average per-group $\text{MCE}_{\%}$ (± 1 s.e.) across datasets for `CaLLM` and the calibration and multicalibration baselines. Family is undefined on `PKU`, since it contains no response from the judge's own family.

Table 7: Average ACC and log-loss $\mathcal{L}_{\log}$ ($\pm$ s.e.) across datasets. `CaLLM` always matches or improves both metrics over all calibration and multicalibration strategies.

	PKU		AEX		PPE		RWB		MTB	
	ACC	$\mathcal{L}_{\log}$	ACC	$\mathcal{L}_{\log}$	ACC	$\mathcal{L}_{\log}$	ACC	$\mathcal{L}_{\log}$	ACC	$\mathcal{L}_{\log}$
`CaLLM`	**.69 ± .01**	**.57 ± .01**	**.68 ± .01**	**.60 ± .01**	**.67 ± .01**	**.61 ± .01**	**.93 ± .01**	**.21 ± .01**	**.81 ± .01**	**.41 ± .02**
f_0	.63 ± .01	.73 ± .01	.63 ± .01	.76 ± .01	.64 ± .01	.74 ± .01	.86 ± .01	.36 ± .02	.79 ± .01	.49 ± .02
`Temperature`	.64 ± .01	.63 ± .01	.65 ± .01	.64 ± .01	.65 ± .01	.63 ± .01	.88 ± .01	.30 ± .01	.80 ± .01	.43 ± .01
`Platt`	.64 ± .01	.63 ± .01	.65 ± .01	.64 ± .01	.65 ± .01	.63 ± .01	.88 ± .01	.30 ± .01	.80 ± .01	.43 ± .01
`LOGR`	.64 ± .01	.62 ± .01	.65 ± .01	.63 ± .01	.65 ± .01	.63 ± .01	.88 ± .01	.30 ± .01	.81 ± .01	.43 ± .01
`IGLB`	.64 ± .01	.62 ± .01	.64 ± .01	.63 ± .01	.65 ± .01	.63 ± .01	.88 ± .01	.30 ± .01	.81 ± .01	.44 ± .02

calibration comparisons defines a group. We also include f_0 for comparison with the uncalibrated judge. Finally, we compare with two *global* calibration strategies: temperature scaling (Guo et al., 2017) and Platt scaling (Platt, 1999).

Fig. 6 shows the per-group $\text{MCE}_{\%}$ of `CaLLM` and the other calibration strategies across datasets. Correcting the scores within groups decreases the biases: `MCGrad` beats both global calibrators on all 14 (dataset–bias) pairs, while `LOGR` and `IGLB` fail to beat them mostly on the two smallest datasets, `RWB` and `MTB`. `MCGrad` is the best on 11 of 14 pairs, with the largest gap on *Verbosity*, where it lowers the error of the best calibrator by at least 27%. Indeed, verbosity depends on how the two responses compare, which the clusters of the other multicalibration approaches are unlikely to isolate; `MCGrad` instead searches the contrastive $\phi(X)$ for the groups where the judge is miscalibrated.

We report in Table 7 accuracy and log-loss on all five datasets. `MCGrad` is the best on both metrics for all datasets with the sole exception of accuracy in `MTB`: `MCGrad` ties with the two multicalibration strategies *i.e.*, `LOGR` and `IGLB`.

E.5 Validating Recoverability

In this section, we assess whether `CaLLM` can recover the input-space groups $\mathcal{G}$ from the supplied representation $[\phi(X), f_0(X)]$. This condition directly reduces the terms φ and ψ in Theorem 1 as illustrated in Corollary 1. Specifically, we take $\mathcal{H}$ to be the tree family used by `MCGrad` and measure how accurately it predicts membership in the *position*, *verbosity*, and *family* bias groups. We carry this analysis using `Muse-Glimmer-30B` as both the judge and the encoder as in Section 4.

Table 8: Recoverability results across datasets.

	PKU	AEX	PPE	RWB	MTB
Position	0.50 ± 0.01	0.76 ± 0.01	0.76 ± 0.01	0.93 ± 0.01	0.66 ± 0.05
Verbosity	0.92 ± 0.01	0.90 ± 0.01	0.89 ± 0.01	0.95 ± 0.01	0.94 ± 0.02
Family	-	0.98 ± 0.01	0.98 ± 0.01	0.99 ± 0.01	0.98 ± 0.02

Table 9: Average per-group $\text{MCE}_{\%}$ and MCE_{σ} ($\pm$ standard error) for CaLLM and the mitigation baselines across datasets. The *family* axis does not show PKU since it does not contain muse-family responses.

		$\text{MCE}_{\%}$			MCE_{σ}		
Data	Method	Position	Verbosity	Family	Position	Verbosity	Family
PKU	CaLLM	**2.78 ± 0.23**	**5.55 ± 0.28**	–	**3.85 ± 0.31**	**7.68 ± 0.38**	–
	f_0	12.34 ± 0.40	22.47 ± 0.28	–	18.08 ± 0.61	32.92 ± 0.47	–
	BPC	7.91 ± 0.25	19.25 ± 0.39	–	11.22 ± 0.36	27.28 ± 0.56	–
	CalibraEval	7.66 ± 0.19	20.51 ± 0.35	–	10.18 ± 0.26	27.26 ± 0.47	–
	PORTIA	10.26 ± 0.10	21.06 ± 0.31	–	14.73 ± 0.19	30.22 ± 0.45	–
	LenControl	12.40 ± 0.54	21.16 ± 0.70	–	18.16 ± 0.75	31.00 ± 1.02	–
AEX	CaLLM	**3.72 ± 0.35**	**4.49 ± 0.27**	**2.58 ± 0.16**	**4.27 ± 0.41**	**5.14 ± 0.29**	**2.96 ± 0.19**
	f_0	16.11 ± 0.33	16.16 ± 0.33	16.11 ± 0.33	20.82 ± 0.39	20.88 ± 0.40	20.82 ± 0.39
	BPC	8.21 ± 0.36	10.46 ± 0.35	8.21 ± 0.36	10.09 ± 0.42	12.84 ± 0.41	10.09 ± 0.42
	CalibraEval	11.18 ± 0.51	12.44 ± 0.46	11.18 ± 0.51	14.06 ± 0.66	15.63 ± 0.58	14.06 ± 0.66
	PORTIA	10.05 ± 0.47	10.06 ± 0.47	10.05 ± 0.47	12.79 ± 0.58	12.80 ± 0.58	12.79 ± 0.58
	LenControl	12.96 ± 0.34	13.15 ± 0.24	12.96 ± 0.34	16.86 ± 0.42	17.11 ± 0.30	16.86 ± 0.42
PPE	CaLLM	**5.41 ± 0.18**	**7.48 ± 0.41**	**4.77 ± 0.34**	**2.63 ± 0.10**	**3.62 ± 0.19**	**2.32 ± 0.17**
	f_0	15.01 ± 0.65	16.08 ± 0.52	15.01 ± 0.65	8.35 ± 0.38	8.93 ± 0.30	8.35 ± 0.38
	BPC	9.05 ± 0.60	12.36 ± 0.40	9.24 ± 0.51	4.80 ± 0.33	6.55 ± 0.21	4.91 ± 0.28
	CalibraEval	9.25 ± 1.77	13.67 ± 0.71	9.06 ± 1.71	4.91 ± 1.07	7.12 ± 0.59	4.81 ± 1.05
	PORTIA	10.15 ± 0.42	10.71 ± 0.32	10.15 ± 0.42	5.55 ± 0.23	5.86 ± 0.18	5.55 ± 0.23
	LenControl	13.82 ± 0.58	14.47 ± 0.18	13.82 ± 0.58	7.75 ± 0.34	8.10 ± 0.12	7.75 ± 0.34
RWB	CaLLM	6.90 ± 1.11	**5.22 ± 0.41**	5.21 ± 0.27	3.70 ± 0.57	**2.82 ± 0.25**	2.81 ± 0.18
	f_0	7.35 ± 0.91	13.08 ± 0.42	6.78 ± 0.67	3.35 ± 0.41	5.98 ± 0.15	3.09 ± 0.28
	BPC	**6.10 ± 0.26**	10.92 ± 0.53	5.69 ± 0.26	**2.48 ± 0.09**	4.45 ± 0.22	2.32 ± 0.11
	CalibraEval	7.70 ± 0.43	10.71 ± 0.92	7.29 ± 0.22	3.14 ± 0.18	4.42 ± 0.55	2.99 ± 0.20
	PORTIA	6.24 ± 0.73	10.21 ± 0.60	4.83 ± 0.38	2.80 ± 0.31	4.59 ± 0.23	**2.17 ± 0.15**
	LenControl	9.55 ± 0.73	7.62 ± 0.50	**4.68 ± 0.40**	4.94 ± 0.35	3.95 ± 0.28	2.42 ± 0.21
MTB	CaLLM	**13.42 ± 1.59**	**13.46 ± 1.28**	11.89 ± 0.79	**2.23 ± 0.26**	**2.20 ± 0.20**	1.96 ± 0.12
	f_0	15.86 ± 2.03	16.15 ± 0.82	12.00 ± 0.92	2.86 ± 0.40	2.85 ± 0.09	2.15 ± 0.22
	BPC	14.30 ± 2.05	15.76 ± 1.11	**10.51 ± 1.04**	2.43 ± 0.37	2.63 ± 0.16	**1.79 ± 0.19**
	CalibraEval	16.67 ± 2.23	17.92 ± 1.12	13.47 ± 1.88	2.91 ± 0.42	3.09 ± 0.20	2.35 ± 0.36
	PORTIA	14.50 ± 1.27	21.07 ± 2.76	15.12 ± 2.94	2.42 ± 0.24	3.44 ± 0.39	2.56 ± 0.55
	LenControl	14.12 ± 1.99	14.63 ± 0.65	13.54 ± 1.74	2.63 ± 0.38	2.69 ± 0.10	2.53 ± 0.34

Results are reported in Table 8. Across the 14 available dataset-group pairs, the bias groups are generally well recoverable from the representation, with median accuracies of ≈ 0.76 for *Position*, ≈ 0.92 for *Verbosity*, and ≈ 0.98 for *Family*. The main exception is *Position* on PKU, where recoverability is at chance level (0.50). Interestingly, CaLLM still substantially reduces the *Position* bias on PKU, and explicitly providing the position groups yields little further improvement. This is consistent with our analysis: poor group recoverability need not prevent bias mitigation when the corresponding judge-error components, λ and χ, are sufficiently small, allowing larger group-side terms φ and ψ without inducing large residual bias. Thus, CaLLM can mitigate position bias even when position membership itself is poorly recoverable.

E.6 Detailed results

We provide all our results in what follows.

Table 10: Average per-group $\text{MCE}_{\%}$ and MCE_{σ} ($\pm$ standard error) for `CaLLM` and `CaLLM`+g. The *family* axis does not show `PKU` since it does not contain `muse`-family responses.

		$\text{MCE}_{\%}$			MCE_{σ}		
Dataset	Method	Position	Verbosity	Family	Position	Verbosity	Family
PKU	`CaLLM`	2.78 ± 0.23	5.55 ± 0.28	–	3.85 ± 0.31	7.68 ± 0.38	–
	`CaLLM`+g	$\mathbf{2.56 \pm 0.10}$	$\mathbf{2.81 \pm 0.18}$	–	$\mathbf{3.58 \pm 0.14}$	$\mathbf{3.92 \pm 0.25}$	–
AEX	`CaLLM`	3.72 ± 0.35	4.49 ± 0.27	2.58 ± 0.16	4.27 ± 0.41	5.14 ± 0.29	2.96 ± 0.19
	`CaLLM`+g	$\mathbf{2.52 \pm 0.30}$	$\mathbf{2.55 \pm 0.24}$	$\mathbf{2.50 \pm 0.19}$	$\mathbf{2.90 \pm 0.34}$	$\mathbf{2.93 \pm 0.29}$	$\mathbf{2.87 \pm 0.22}$
PPE	`CaLLM`	5.41 ± 0.18	7.48 ± 0.41	4.77 ± 0.34	2.63 ± 0.10	3.62 ± 0.19	2.32 ± 0.17
	`CaLLM`+g	$\mathbf{3.67 \pm 0.30}$	$\mathbf{3.99 \pm 0.18}$	$\mathbf{4.56 \pm 0.41}$	$\mathbf{1.79 \pm 0.15}$	$\mathbf{1.95 \pm 0.09}$	$\mathbf{2.23 \pm 0.21}$
RWB	`CaLLM`	6.90 ± 1.11	5.22 ± 0.41	5.21 ± 0.27	3.70 ± 0.57	$\mathbf{2.82 \pm 0.25}$	$\mathbf{2.81 \pm 0.18}$
	`CaLLM`+g	$\mathbf{4.66 \pm 0.45}$	$\mathbf{4.70 \pm 0.40}$	$\mathbf{4.64 \pm 0.15}$	$\mathbf{2.83 \pm 0.21}$	2.87 ± 0.22	2.83 ± 0.10
MTB	`CaLLM`	13.42 ± 1.59	13.46 ± 1.28	11.89 ± 0.79	2.23 ± 0.26	2.20 ± 0.20	1.96 ± 0.12
	`CaLLM`+g	$\mathbf{12.53 \pm 1.54}$	$\mathbf{12.33 \pm 0.86}$	$\mathbf{11.63 \pm 0.23}$	$\mathbf{2.10 \pm 0.24}$	$\mathbf{2.05 \pm 0.14}$	$\mathbf{1.91 \pm 0.04}$

Table 11: Average per-group $\text{MCE}_{\%}$ and MCE_{σ} ($\pm$ standard error) for `CaLLM` and the mitigation baselines across datasets over the three biases documented in the literature (*i.e.* position, verbosity, and family) and the *unspecified* groups.

		$\text{MCE}_{\%}$		MCE_{σ}	
Dataset	Method	Specified	Unspecified	Specified	Unspecified
PKU	`CaLLM`	$\mathbf{5.55 \pm 0.28}$	$\mathbf{3.89 \pm 0.36}$	$\mathbf{7.68 \pm 0.38}$	$\mathbf{5.38 \pm 0.50}$
	f_0	22.46 ± 0.28	12.34 ± 0.40	32.91 ± 0.47	18.08 ± 0.61
	`BPC`	19.23 ± 0.40	7.91 ± 0.25	27.26 ± 0.56	11.22 ± 0.36
	`CalibraEval`	20.49 ± 0.35	7.66 ± 0.19	27.24 ± 0.47	10.18 ± 0.26
	`PORTIA`	21.04 ± 0.31	10.26 ± 0.10	30.20 ± 0.45	14.73 ± 0.19
	`LenControl`	21.15 ± 0.71	12.40 ± 0.54	30.99 ± 1.02	18.16 ± 0.75
AEX	`CaLLM`	$\mathbf{5.09 \pm 0.19}$	$\mathbf{3.29 \pm 0.14}$	$\mathbf{5.83 \pm 0.21}$	$\mathbf{3.77 \pm 0.17}$
	f_0	16.17 ± 0.33	16.11 ± 0.33	20.89 ± 0.40	20.82 ± 0.39
	`BPC`	10.46 ± 0.35	8.21 ± 0.36	12.84 ± 0.41	10.09 ± 0.42
	`CalibraEval`	12.44 ± 0.46	11.18 ± 0.51	15.63 ± 0.58	14.06 ± 0.66
	`PORTIA`	10.06 ± 0.47	10.05 ± 0.47	12.80 ± 0.58	12.79 ± 0.58
	`LenControl`	13.15 ± 0.24	12.96 ± 0.34	17.11 ± 0.30	16.86 ± 0.42
PPE	`CaLLM`	$\mathbf{7.97 \pm 0.43}$	$\mathbf{7.16 \pm 0.21}$	$\mathbf{3.87 \pm 0.20}$	$\mathbf{3.47 \pm 0.10}$
	f_0	16.05 ± 0.56	15.01 ± 0.65	8.93 ± 0.32	8.35 ± 0.38
	`BPC`	12.36 ± 0.41	11.31 ± 0.55	6.56 ± 0.21	6.00 ± 0.29
	`CalibraEval`	13.64 ± 0.78	12.69 ± 0.76	7.12 ± 0.63	6.61 ± 0.56
	`PORTIA`	10.77 ± 0.37	10.66 ± 0.62	5.90 ± 0.21	5.83 ± 0.33
	`LenControl`	14.47 ± 0.18	14.06 ± 0.36	8.11 ± 0.12	7.89 ± 0.22
RWB	`CaLLM`	$\mathbf{7.27 \pm 0.89}$	$\mathbf{8.00 \pm 0.43}$	$\mathbf{3.90 \pm 0.45}$	$\mathbf{4.29 \pm 0.16}$
	f_0	13.08 ± 0.42	13.77 ± 1.12	5.98 ± 0.15	6.28 ± 0.46
	`BPC`	11.71 ± 0.27	10.90 ± 0.49	4.77 ± 0.08	4.44 ± 0.18
	`CalibraEval`	11.05 ± 0.85	13.42 ± 0.22	4.54 ± 0.50	5.52 ± 0.41
	`PORTIA`	10.83 ± 0.67	11.23 ± 0.44	4.87 ± 0.27	5.06 ± 0.25
	`LenControl`	10.82 ± 0.67	8.87 ± 0.46	5.60 ± 0.35	4.58 ± 0.17
MTB	`CaLLM`	$\mathbf{16.36 \pm 0.84}$	$\mathbf{18.80 \pm 0.57}$	$\mathbf{2.70 \pm 0.11}$	$\mathbf{3.12 \pm 0.13}$
	f_0	19.59 ± 0.56	21.74 ± 0.45	3.50 ± 0.12	3.88 ± 0.05
	`BPC`	18.00 ± 1.11	19.00 ± 0.69	3.04 ± 0.19	3.21 ± 0.12
	`CalibraEval`	20.40 ± 0.96	23.38 ± 1.13	3.55 ± 0.20	4.07 ± 0.22
	`PORTIA`	21.21 ± 2.32	26.31 ± 3.89	3.51 ± 0.35	4.37 ± 0.63
	`LenControl`	17.26 ± 1.05	21.16 ± 0.72	3.20 ± 0.19	3.93 ± 0.14

Table 12: Average per-group $MCE_{\%}$ and MCE_{σ}($\pm$ standard error) for CaLLM and the mitigation baselines across datasets with `Qwen 2.5-7B-Instruct` used as the judge. The *family* axis contains AEX and PPE only: PKU, RWB and MTB do not contain qwen-family responses.

Dataset	Method	$MCE_{\%}$			MCE_{σ}		
		Position	Verbosity	Family	Position	Verbosity	Family
PKU	CaLLM	**2.39 ± 0.17**	**6.32 ± 0.53**	–	**3.17 ± 0.23**	**8.39 ± 0.69**	–
	f_0	19.81 ± 0.28	19.80 ± 0.28	–	31.48 ± 0.45	31.45 ± 0.45	–
	BPC	8.86 ± 0.15	9.43 ± 0.13	–	12.91 ± 0.21	13.73 ± 0.20	–
	CalibraEval	8.13 ± 0.33	11.05 ± 0.38	–	10.95 ± 0.48	14.87 ± 0.49	–
	PORTIA	10.47 ± 0.10	17.09 ± 0.14	–	14.50 ± 0.12	23.66 ± 0.21	–
	LenControl	16.56 ± 0.11	16.55 ± 0.11	–	26.13 ± 0.18	26.10 ± 0.18	–
AEX	CaLLM	**6.78 ± 0.39**	**4.63 ± 0.29**	**2.70 ± 0.27**	**7.37 ± 0.44**	**5.03 ± 0.31**	**2.93 ± 0.29**
	f_0	40.95 ± 0.42	40.95 ± 0.42	40.95 ± 0.42	58.37 ± 0.55	58.35 ± 0.55	58.37 ± 0.55
	BPC	10.44 ± 0.30	9.92 ± 0.27	9.92 ± 0.27	12.39 ± 0.36	11.77 ± 0.31	11.77 ± 0.30
	CalibraEval	7.61 ± 0.28	9.12 ± 0.38	5.40 ± 0.10	8.12 ± 0.29	9.72 ± 0.39	5.76 ± 0.09
	PORTIA	13.79 ± 0.26	13.75 ± 0.29	13.79 ± 0.26	17.11 ± 0.32	17.06 ± 0.35	17.11 ± 0.32
	LenControl	16.79 ± 0.30	16.79 ± 0.30	16.79 ± 0.30	21.45 ± 0.37	21.45 ± 0.37	21.45 ± 0.37
PPE	CaLLM	**8.27 ± 0.45**	**7.19 ± 0.41**	**5.74 ± 0.91**	**3.81 ± 0.20**	**3.31 ± 0.19**	**2.64 ± 0.42**
	f_0	39.11 ± 0.72	39.18 ± 0.71	39.11 ± 0.72	24.17 ± 0.41	24.17 ± 0.39	24.17 ± 0.41
	BPC	11.98 ± 0.57	11.49 ± 0.68	11.46 ± 0.72	6.21 ± 0.31	5.94 ± 0.37	5.94 ± 0.39
	CalibraEval	13.01 ± 2.65	12.21 ± 2.54	11.02 ± 2.96	6.89 ± 1.97	6.46 ± 1.87	5.91 ± 2.07
	PORTIA	14.63 ± 0.59	14.57 ± 0.57	14.55 ± 0.58	7.84 ± 0.29	7.79 ± 0.28	7.80 ± 0.29
	LenControl	19.47 ± 0.69	19.43 ± 0.68	19.47 ± 0.69	10.97 ± 0.38	10.93 ± 0.38	10.97 ± 0.38
RWB	CaLLM	9.98 ± 1.21	**9.70 ± 1.04**	–	3.51 ± 0.44	**3.42 ± 0.38**	–
	f_0	13.66 ± 1.07	22.11 ± 1.42	–	5.24 ± 0.41	8.47 ± 0.54	–
	BPC	**8.64 ± 1.15**	16.87 ± 1.21	–	**2.96 ± 0.42**	5.74 ± 0.40	–
	CalibraEval	11.04 ± 1.24	17.03 ± 1.76	–	3.83 ± 0.60	5.89 ± 0.85	–
	PORTIA	12.61 ± 1.31	16.30 ± 1.84	–	4.00 ± 0.42	5.17 ± 0.57	–
	LenControl	14.78 ± 1.22	12.72 ± 1.16	–	5.99 ± 0.47	5.15 ± 0.44	–
MTB	CaLLM	**11.83 ± 0.77**	14.22 ± 0.69	–	**1.85 ± 0.16**	**2.19 ± 0.13**	–
	f_0	13.88 ± 1.07	14.25 ± 1.03	–	2.39 ± 0.23	2.44 ± 0.23	–
	BPC	14.28 ± 1.10	14.22 ± 1.23	–	2.23 ± 0.16	2.20 ± 0.15	–
	CalibraEval	13.80 ± 1.33	20.30 ± 1.83	–	2.16 ± 0.29	3.06 ± 0.24	–
	PORTIA	16.95 ± 2.67	24.30 ± 3.90	–	2.48 ± 0.39	3.52 ± 0.54	–
	LenControl	15.93 ± 1.08	**14.18 ± 0.86**	–	2.94 ± 0.25	2.58 ± 0.16	–

Table 13: Average per-group $MCE_{\%}$ and MCE_{σ} ($\pm$1 s.e.) for CaLLM (embeddings from the judge's last hidden state), `CaLLM(Qwen3-0.6B)` (embeddings from `Qwen3-Embedding-0.6B`) and, as a reference, the *best* of the Fig. 2 baselines across datasets. Family is undefined on PKU, since it contains no response from the judge's own family.

Dataset	Method	$MCE_{\%}$			MCE_{σ}		
		Position	Verbosity	Family	Position	Verbosity	Family
PKU	CaLLM	2.78 ± 0.23	**5.55 ± 0.28**	–	3.85 ± 0.31	**7.68 ± 0.38**	–
	CaLLM (Qwen3-0.6B)	**2.73 ± 0.22**	7.39 ± 0.62	–	**3.74 ± 0.31**	10.12 ± 0.84	–
	best baseline	7.66 ± 0.19	19.25 ± 0.39	–	10.18 ± 0.26	27.26 ± 0.47	–
AEX	CaLLM	**3.72 ± 0.35**	**4.49 ± 0.27**	2.58 ± 0.16	**4.27 ± 0.41**	**5.14 ± 0.29**	2.96 ± 0.19
	CaLLM (Qwen3-0.6B)	4.17 ± 0.31	5.49 ± 0.28	**2.28 ± 0.14**	4.70 ± 0.36	6.19 ± 0.31	**2.57 ± 0.16**
	best baseline	8.21 ± 0.36	10.06 ± 0.47	8.21 ± 0.36	10.09 ± 0.42	12.80 ± 0.58	10.09 ± 0.42
PPE	CaLLM	**5.41 ± 0.18**	**7.48 ± 0.41**	**4.77 ± 0.34**	**2.63 ± 0.10**	**3.62 ± 0.19**	**2.32 ± 0.17**
	CaLLM (Qwen3-0.6B)	5.65 ± 0.27	8.60 ± 0.55	5.08 ± 0.27	2.73 ± 0.14	4.14 ± 0.25	2.45 ± 0.14
	best baseline	9.05 ± 0.60	10.71 ± 0.32	9.06 ± 1.71	4.80 ± 0.33	5.86 ± 0.18	4.81 ± 1.05
RWB	CaLLM	6.90 ± 1.11	**5.22 ± 0.41**	5.21 ± 0.27	3.70 ± 0.57	**2.82 ± 0.25**	2.81 ± 0.18
	CaLLM (Qwen3-0.6B)	6.67 ± 0.95	7.88 ± 0.69	6.02 ± 0.63	2.94 ± 0.43	3.44 ± 0.22	2.64 ± 0.27
	best baseline	**6.10 ± 0.26**	7.62 ± 0.50	**4.68 ± 0.40**	**2.48 ± 0.09**	3.95 ± 0.28	**2.17 ± 0.15**
MTB	CaLLM	**13.42 ± 1.59**	**13.46 ± 1.28**	11.89 ± 0.79	**2.23 ± 0.26**	**2.20 ± 0.20**	1.96 ± 0.12
	CaLLM (Qwen3-0.6B)	14.15 ± 2.50	14.31 ± 1.68	**10.43 ± 0.76**	2.44 ± 0.47	2.39 ± 0.27	**1.74 ± 0.16**
	best baseline	14.12 ± 1.99	14.63 ± 0.65	10.51 ± 1.04	2.42 ± 0.24	2.63 ± 0.16	1.79 ± 0.19

Table 14: Average per-group $\text{MCE}_{\%}$ and MCE_{σ} (± 1 s.e.) for CaLLM and the calibration and multicalibration baselines across datasets. Family is undefined on PKU, since it contains no response from the judge's own family.

Dataset	Method	$\text{MCE}_{\%}$			MCE_{σ}		
		Position	Verbosity	Family	Position	Verbosity	Family
PKU	CaLLM	**2.78 ± 0.23**	**5.55 ± 0.28**	–	3.85 ± 0.31	**7.68 ± 0.38**	–
	f_0	12.34 ± 0.40	22.47 ± 0.28	–	18.08 ± 0.61	32.92 ± 0.47	–
	Temperature	5.56 ± 0.28	19.84 ± 0.20	–	7.10 ± 0.36	25.32 ± 0.29	–
	Platt	5.29 ± 0.20	18.64 ± 0.51	–	6.76 ± 0.26	23.80 ± 0.64	–
	LOGR	2.87 ± 0.15	17.71 ± 0.48	–	**3.68 ± 0.19**	22.73 ± 0.61	–
	IGLB	3.33 ± 0.36	16.86 ± 0.47	–	4.30 ± 0.46	21.78 ± 0.60	–
AEX	CaLLM	**3.72 ± 0.35**	**4.49 ± 0.27**	2.58 ± 0.16	**4.27 ± 0.41**	**5.14 ± 0.29**	2.96 ± 0.19
	f_0	16.11 ± 0.33	16.16 ± 0.33	16.11 ± 0.33	20.82 ± 0.39	20.88 ± 0.40	20.82 ± 0.39
	Temperature	6.58 ± 0.22	11.52 ± 0.43	4.70 ± 0.22	7.11 ± 0.25	12.46 ± 0.46	5.08 ± 0.23
	Platt	5.23 ± 0.51	9.50 ± 0.25	3.93 ± 0.22	5.67 ± 0.56	10.28 ± 0.26	4.26 ± 0.25
	LOGR	4.13 ± 0.44	8.86 ± 0.27	**1.99 ± 0.10**	4.50 ± 0.48	9.64 ± 0.28	**2.17 ± 0.10**
	IGLB	4.00 ± 0.35	8.73 ± 0.30	2.32 ± 0.07	4.38 ± 0.39	9.55 ± 0.31	2.53 ± 0.07
PPE	CaLLM	5.41 ± 0.18	**7.48 ± 0.41**	**4.77 ± 0.34**	2.63 ± 0.10	**3.62 ± 0.19**	**2.32 ± 0.17**
	f_0	15.01 ± 0.65	16.08 ± 0.52	15.01 ± 0.65	8.35 ± 0.38	8.93 ± 0.30	8.35 ± 0.38
	Temperature	6.37 ± 0.32	12.07 ± 0.87	5.12 ± 0.47	2.99 ± 0.15	5.65 ± 0.41	2.40 ± 0.22
	Platt	5.75 ± 0.28	11.69 ± 0.47	5.11 ± 0.23	2.70 ± 0.13	5.48 ± 0.21	2.40 ± 0.11
	LOGR	**5.28 ± 0.40**	11.33 ± 0.43	5.03 ± 0.24	**2.49 ± 0.19**	5.34 ± 0.19	2.37 ± 0.12
	IGLB	5.74 ± 0.26	11.59 ± 0.42	5.51 ± 0.35	2.73 ± 0.14	5.51 ± 0.18	2.62 ± 0.16
RWB	CaLLM	6.90 ± 1.11	**5.22 ± 0.41**	**5.21 ± 0.27**	3.70 ± 0.57	**2.82 ± 0.25**	2.81 ± 0.18
	f_0	7.35 ± 0.91	13.08 ± 0.42	6.78 ± 0.67	3.35 ± 0.41	5.98 ± 0.15	3.09 ± 0.28
	Temperature	8.38 ± 1.07	12.16 ± 0.38	6.21 ± 0.68	3.16 ± 0.41	4.58 ± 0.13	2.34 ± 0.25
	Platt	7.17 ± 0.76	10.33 ± 0.29	6.24 ± 0.48	2.71 ± 0.29	3.90 ± 0.09	2.36 ± 0.18
	LOGR	7.65 ± 0.74	11.88 ± 0.31	6.43 ± 0.73	2.92 ± 0.28	4.53 ± 0.10	2.45 ± 0.27
	IGLB	**6.50 ± 0.68**	11.20 ± 0.76	5.89 ± 0.71	**2.45 ± 0.23**	4.23 ± 0.27	**2.22 ± 0.23**
MTB	CaLLM	**13.42 ± 1.59**	**13.46 ± 1.28**	**11.89 ± 0.79**	**2.23 ± 0.26**	**2.20 ± 0.20**	**1.96 ± 0.12**
	f_0	15.86 ± 2.03	16.15 ± 0.82	12.00 ± 0.92	2.86 ± 0.40	2.85 ± 0.09	2.15 ± 0.22
	Temperature	15.37 ± 1.99	18.31 ± 1.64	13.26 ± 0.65	2.33 ± 0.33	2.72 ± 0.18	1.99 ± 0.13
	Platt	15.15 ± 2.32	18.53 ± 2.08	13.57 ± 0.52	2.31 ± 0.38	2.75 ± 0.25	2.04 ± 0.11
	LOGR	15.17 ± 2.66	18.77 ± 1.28	15.10 ± 0.33	2.35 ± 0.43	2.86 ± 0.17	2.32 ± 0.09
	IGLB	18.29 ± 2.31	19.33 ± 1.61	14.63 ± 1.64	3.10 ± 0.43	3.20 ± 0.21	2.42 ± 0.28