Truthful Calibration Errors for Multi-Class Prediction

Yuxuan Lu[*]
School of Computer Science
Peking University
Beijing, China
yx_lu@pku.edu.cn

Yifan Wu[*]
Microsoft Research, New England
Cambridge, US
yifan.wu2357@gmail.com

Jason Hartline
McCormick School of Engineering
Northwestern University
Evanston, US
hartline@northwestern.edu

Lunjia Hu
Khoury College of Computer Sciences
Northeastern University
Boston, US
l.hu@northeastern.edu

Abstract

Calibrated predictions are useful because their numerical values can be interpreted as probabilities. Calibration errors are therefore widely used to evaluate and compare probabilistic predictors. Recently, Haghtalab et al. [2024] introduced truthfulness as an additional requirement for such measures. A calibration measure is truthful if a predictor minimizes its expected measured error by reporting the true conditional label distribution. Many standard empirical calibration errors are non-truthful: a predictor may appear better calibrated by distorting its probabilities. We study the implications of truthful and non-truthful calibration errors for practice. First, we introduce perfectly truthful calibration errors for full multiclass calibration and classwise calibration, two standard multiclass notions. More generally, our construction applies to any linear property of the label distribution, generalizing the truthful calibration error for binary predictions in Hartline et al. [2025]. We also identify a truthful correction for confidence calibration. Second, we characterize the decision-theoretic implications of these truthful errors. For fully calibrated probability predictors, truthful calibration errors preserve Blackwell dominance: a more informative predictor receives no larger expected error. Third, we show that this decision-theoretic interpretation explains and mitigates the well-observed ranking robustness problem of binned calibration errors. Empirically, non-truthful confidence-based errors can reverse model rankings when the number of bins changes, while our truthful classwise error gives more stable rankings across binning choices.

1 Introduction

Probabilistic predictions are useful only when their numerical values can be interpreted as probabilities. A weather forecast that reports a 40% chance of rain should mean that, among days receiving this forecast, rain occurs about 40% of the time. This requirement is calibration [Dawid, 1982]: for a binary label $y \in \{0,1\}$ and prediction $p \in [0,1]$, calibration asks that $\Pr[y=1 \mid p] = p$.

Calibration errors quantify deviations from perfect calibration. The canonical error is Expected Calibration Error (ECE), which measures the expected absolute prediction bias: $\mathbf{E}\left[\left|p - \Pr[y=1 \mid p]\right|\right]$.

A calibration error creates incentives. If a predictor can reduce its expected measured error by reporting probabilities different from the true conditional probabilities, then the measure is not only evaluating calibration; it is also rewarding strategic distortion. Haghtalab et al. [2024] initiate the study of approximate truthfulness in the online prediction problem. Hartline et al. [2025] propose a

[*]Co-first Authors

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

perfectly truthful error for binary classification in the batch setting. A calibration measure is truthful if, for every sequence of samples with true conditional label distributions, the predictor minimizes its expected measured calibration error by reporting these true conditional distributions. In other words, a truthful calibration error incentivizes true probability predictions. Proper scoring rules also incentivize truthful probability reports [Gneiting and Raftery, 2007], but their expected loss need not vanish for perfectly calibrated predictors. Calibration measures must satisfy completeness and soundness: their expected error should vanish asymptotically for calibrated predictors and remain bounded away from zero for miscalibrated predictors.

The following example explains the non-truthfulness of the empirical ECE.

Example 1.1 (Non-truthfulness of empirical ECE)**.** *Consider n samples whose true label probabilities are $p_i^* = i/n$, for $i = 1,\ldots,n$. The truthful predictor reports $p_i = p_i^*$. If empirical ECE is computed by conditioning on exact prediction values, then each reported value appears only once, so the empirical estimate of $\Pr[y=1 \mid p_i]$ has constant error. The truthful predictor therefore incurs empirical calibration error of order $O(1)$. By contrast, a predictor that always reports $1/2$ pools all samples together and achieves empirical error of order $O(1/\sqrt{n})$.*

Binning is the standard practical response to this problem, but remains untruthful. Instead of conditioning on exact prediction values, a binned calibration error partitions the prediction space into m bins and conditions on the bin. In Example 1.1, the truthful predictor has about n/m samples per bin and incurs sampling error of order $O(\sqrt{m/n})$, while the constant predictor still pools all samples into one bin and obtains error $O(1/\sqrt{n})$. Thus a non-truthful metric can prefer a less informative predictor simply because a bin with more samples induces lower average error.

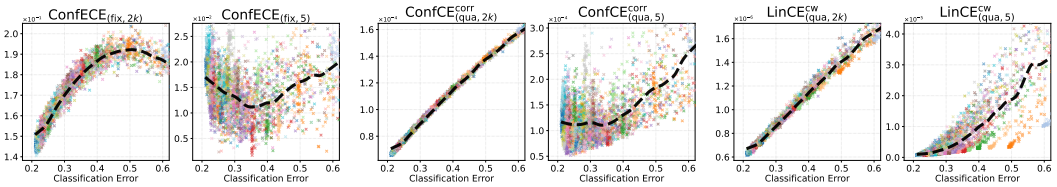
In this paper, we study the implications of truthful and non-truthful calibration errors for practice, especially for multiclass prediction. Our central message is that truthfulness is not merely an abstract incentive criterion. For the calibration errors we construct, truthfulness leads to calibration measures with a decision-theoretic interpretation, and this interpretation helps explain and mitigate the ranking robustness problems of standard binned calibration errors.

Our first contribution is to construct truthful calibration errors for standard multiclass notions, including full multiclass calibration and classwise calibration, through minimal changes to canonical binned ECE. We present these notions through a common language of *linear properties*: a property is a target summary of the label distribution, and it is linear when it can be written as $\Gamma_\phi(P) = \mathbb{E}_{y\sim P}[\phi(y)]$ for a fixed function of the label ϕ. Full multiclass calibration corresponds to $\phi(y) = e_y$ where e_y is the standard basis vector with a 1 in the coordinate of the realized label y and 0 elsewhere. Classwise calibration corresponds to the one-vs-all coordinates $\phi_r(y) = \mathbb{I}[y = r]$.

The construction replaces the ℓ_1 norm of each bin's summed residuals by its squared ℓ_2 norm, using the common, report-independent prefactor m/n^2. Quantile binning is used for the scalar completeness and soundness guarantees. For linear properties, we prove that this squared binned calibration error is truthful for every outcome-independent binning rule. This linear-property construction directly extends the binary construction of Hartline et al. [2025] using the same bias–variance decomposition: the expected error separates into a report-dependent squared bias term and a sampling-variance term, and reporting the true conditional property makes the bias term vanish. Our main new ingredients are the confidence correction and the decision-theoretic order-preservation results developed below.

We also examine confidence calibration, a common relaxation of full calibration in the multiclass setting. Given a multiclass prediction $p \in \Delta(\mathcal{Y})$, the confidence is the largest reported probability, $\max_{r\in\mathcal{Y}} p_r$. Confidence calibration asks that, among samples with the same reported confidence c, the predicted top label is correct with probability c. We show that simply applying a truthful binary calibration error to the reported confidence is not truthful. We then give a corrected confidence calibration error that restores finite-sample truthfulness while preserving the same asymptotic completeness and soundness requirements of a calibration error.

Our second contribution is to characterize the decision-theoretic implications of these truthful calibration errors. A downstream user observes a prediction and then chooses an action. If prediction $f(x)$ Blackwell dominates prediction $g(x)$, then every downstream decision-maker can achieve at least as much expected utility after observing $f(x)$ as after observing $g(x)$. We show that, in general, truthful calibration errors preserve this order among fully calibrated probability predictors. Under classwise or confidence calibration alone, we give separate guarantees for dominance over the corresponding families of proper losses. For our constructed errors, we further prove an approximate



(a) ConfECE is **non-truthful**. (b) Truthful $\mathrm{ConfCE}^{\mathrm{corr}}/m$. (c) Truthful $\mathrm{LinCE}^{\mathrm{cw}}/m$.

Figure 1: **Comparison of calibration errors across bin sizes.** Each dot is a classifier checkpoint after temperature scaling on the validation set[2]. Panel (a) shows the canonical confidence-based calibration error used in the literature, namely the same confidence error as in Example 1.1 but estimated after binning predictions into fixed bins. Panel (b) shows our constructed truthful calibration error for confidence, and panel (c) shows our constructed truthful calibration error for classwise calibration. Vertical axes in (b,c) show the constructed errors divided by m, preserving predictor rankings. In each panel, the left and right plots use different binning sizes, with $m = 2000$ bins and $m = 5$ bins, respectively. The qualitative trend is markedly more robust for the truthful errors in (b) and (c): their calibration error remains aligned with predictive quality across bin sizes, whereas the canonical non-truthful error in (a) exhibits a flip in trend when the number of bins changes.

version for nearly calibrated predictors. Informally, if two predictors differ enough in the associated downstream scoring loss, and the remaining calibration bias of the better predictor is smaller than this loss gap, then the truthful calibration error preserves the same ordering. Thus, a small miscalibration cannot reverse a ranking that is sufficiently separated in downstream decision quality.

Our third contribution is empirical: we show that the above decision-theoretic interpretation helps mitigate the well-observed ranking robustness problem of binned calibration errors. Prior empirical work has observed that conclusions about which predictor is better calibrated can change when the same binned calibration error is computed with a different binning granularity [Nixon et al., 2019, Minderer et al., 2021]. From our perspective, this instability is not only a finite-sample artifact, but is a symptom of a non-truthful measure that can reward non-truthful probabilities. For predictors that are already close to calibrated, our theory shows that truthful errors should not overturn a sufficiently clear downstream-loss ranking. We examine this ranking behavior empirically on neural-network predictors. Non-truthful confidence-based errors can reverse the ranking between predictors when the number of bins changes, whereas our truthful calibration errors preserve the same qualitative ordering across binning choices.

Together, these results show that truthfulness provides a useful design principle for calibration measurement and is achievable via simple modifications to existing calibration errors. It aligns the finite-sample evaluation objective with truthful probability reporting, connects calibration errors to downstream decision quality, and yields more robust empirical model rankings in multiclass prediction. Thus, we recommend that practitioners adopt our truthful calibration measures.

1.1 Related work

Truthful calibration measure. Recent work introduced the notion of *truthfulness* for a calibration measure. Haghtalab et al. [2024] first formalize this concept and design an approximately truthful calibration measure in the online setting. Qiao and Zhao [2025] study calibration errors that are approximately truthful and quantify decision-making payoff. Both papers focus on approximate truthfulness of a calibration measure in the online setting, while our paper focuses on perfect truthfulness in the batch setting. Hartline et al. [2025] design a perfectly truthful calibration measure for binary classification in the batch setting. Our work generalizes Hartline et al. [2025] beyond binary outcomes to multiclass classification tasks. We also provide empirical evaluations of the truthful error metric.

Calibration for multiclass prediction. Calibration in multiclass settings has been studied from several perspectives. Guo et al. [2017] introduced confidence calibration methods such as temperature scaling. Minderer et al. [2021] empirically observe that the ranking by calibration error is not robust to hyperparameter (binning size) selection, which motivated our paper. Kull et al. [2019] discuss confidence and classwise calibration. Gupta and Ramdas [2022] distinguish confidence calibration from top-label calibration and develop a framework for multiclass-to-binary reductions. Widmann et al. [2019] develop kernel-based tests for multiclass calibration, while Błasiok et al. [2023], Blasiok

[2]All points are evaluated on the CIFAR-100 test set and come from the same architecture, MobileNetV3-Small-1.0, trained on different subsets of the training set. Points with the same color correspond to checkpoints from the same training run.

and Nakkiran [2024] study distances from calibration and kernel-smoothed calibration measures. Gopalan et al. [2024] study projected smooth calibration. Zhao et al. [2021], Tang et al. [2025], Hegazy et al. [2026] study calibration relative to downstream decisions or utilities. Our focus is whether a calibration measure incentivizes truthful reporting when evaluated on a finite sample.

Non-robustness of calibration errors to binning granularity. It is a widely observed issue that existing binning-based calibration errors are sensitive to the choice of number of bins. As we mentioned, both works by Nixon et al. [2019] and Minderer et al. [2021] find that different choices of binning granularity can lead to completely opposite conclusions about which models are more calibrated. The work of Kumar et al. [2019] is motivated by recalibration methods that are believed to produce well-calibrated predictions, which turn out to be less calibrated when evaluated using more bins. We show that the truthful calibration errors we introduce can yield more consistent rankings across binning choices when calibration bias is small relative to downstream-loss gaps after recalibration.

2 Preliminaries

Each sample $(x,y)\sim D$ consists of a feature $x\in X$ and an outcome $y\in\mathcal{Y}$. For multiclass prediction, $\mathcal{Y}=\{1,...,k\}$. We write $[n]=\{1,...,n\}$ and let $\mathbb{I}[E]$ denote the indicator of event E. Throughout, every argmax uses a fixed deterministic tie-breaking rule.

Linear properties. A property is a target summary of a label distribution. A *linear property* is specified by a map $\phi:\mathcal{Y}\to\mathbb{R}^d$ and maps a distribution $P\in\Delta(\mathcal{Y})$ to

$$\Gamma_\phi(P):=\mathbf{E}_{y\sim P}[\phi(y)]\in\mathcal{R}_\phi,\qquad \mathcal{R}_\phi:=\mathrm{conv}\{\phi(y):y\in\mathcal{Y}\}.$$

A predictor for Γ_ϕ is a map $g:X\to\mathcal{R}_\phi$. It is calibrated for Γ_ϕ if

$$\mathbf{E}[\phi(y)\,|\,g(x)=u]=u\qquad\text{for every report } u\in\mathcal{R}_\phi.$$

Binary calibration is the special case $\phi(y)=y$ for $y\in\{0,1\}$. Full multiclass calibration is the special case $\phi(y)=e_y$, where e_y is the y-th standard basis vector. Classwise calibration corresponds to the k scalar properties $\phi_r(y)=\mathbb{I}[y=r]$, evaluated on the coordinate reports p_r.

Truthfulness. A batch calibration measure for Γ_ϕ maps n reports and outcomes to a real number:

$$\mathrm{CAL}_\phi:\mathcal{R}_\phi^n\times\mathcal{Y}^n\to\mathbb{R}.$$

It is *truthful* if reporting the true property values minimizes expected error. Formally, for any distributions $P_1,...,P_n\in\Delta(\mathcal{Y})$, let

$$\mu_i^*:=\Gamma_\phi(P_i)=\mathbf{E}_{y_i\sim P_i}[\phi(y_i)].$$

Then CAL_ϕ is truthful if, for every alternative report $\mu_1,...,\mu_n\in\mathcal{R}_\phi$,

$$\mathbf{E}\big[\mathrm{CAL}_\phi(\mu_{1:n}^*;y_{1:n})\big]\le\mathbf{E}\big[\mathrm{CAL}_\phi(\mu_{1:n};y_{1:n})\big],$$

where $y_i\sim P_i$ independently.

Completeness and soundness. Completeness and soundness are different from truthfulness: they are asymptotic properties that detect miscalibration, not finite-sample reporting incentives. Let $\{\mathrm{CAL}_n\}_{n\ge1}$ be a sequence of sample calibration measures for a calibration notion C; for example, C may be calibration for a linear property Γ_ϕ, classwise calibration, or confidence calibration. We say that the sequence $\{\mathrm{CAL}_n\}_{n\ge1}$ is *complete in the limit* for C if, for every C-calibrated predictor h,

$$\lim_{n\to\infty}\mathbf{E}[\mathrm{CAL}_n(h(x_1),...,h(x_n);y_1,...,y_n)]=0.$$

It is *sound in the limit* for C if, for every predictor h that is not C-calibrated,

$$\liminf_{n\to\infty}\mathbf{E}[\mathrm{CAL}_n(h(x_1),...,h(x_n);y_1,...,y_n)]>0.$$

For binned measures, the limit uses the specified binning rules. Deterministic positive rescaling preserves truthfulness and rankings at fixed n,m, but need not preserve completeness or soundness. The squared errors below use m/n^2 to recover their scalar population targets under quantile binning.

Binning. A binning rule B maps reports $u_{1:n}$ to m bins partitioning $[n]$, without depending on the realized outcomes. The prescribed count $m \geq 1$ is report-independent and includes empty bins, which contribute zero. For scalar reports, two standard choices are fixed bins and quantile bins. Fixed bins first choose intervals $I_1,...,I_m$ that partition the report space and then set $B_j = \{i \in [n] : u_i \in I_j\}$. In the common case of reports in $[0,1]$, one may take equal-width bins $I_j = ((j-1)/m, j/m]$ for $j = 1,...,m$, with $0 \in I_1$. Quantile binning with $1 \leq m \leq n$ sorts the reports, breaking ties by sample index independently of the outcomes, and sets (after relabeling in sorted order)

$$B_j = \left\{ i \in [n] : \frac{(j-1)n}{m} < i \leq \frac{jn}{m} \right\}, \qquad j = 1,...,m.$$

For a sequence m_n, we call the associated sequence of quantile binning rules $\{\mathsf{B}^{Q}_{m_n}\}_{n \geq 1}$ a *consistent regime* if $m_n \to \infty$ and $n/m_n \to \infty$, i.e., the number of bins grows while the number of samples per bin also diverges. The truthfulness results below hold for any outcome-independent binning rule.

Definition 2.1 (Binned calibration error for a linear property)**.** *For a map $\phi : \mathcal{Y} \to \mathbb{R}^d$ and a binning rule B, define the ℓ_1 and ℓ_2 versions of calibration error:*

$$\mathrm{LinECE}_{\phi,\mathsf{B}}(u_{1:n}; y_{1:n}) := \frac{1}{n} \sum_{B \in \mathsf{B}(u_{1:n})} \left\| \sum_{i \in B} \big(u_i - \phi(y_i)\big) \right\|_1,$$

$$\mathrm{LinCE}_{\phi,\mathsf{B}}(u_{1:n}; y_{1:n}) := \frac{m}{n^2} \sum_{B \in \mathsf{B}(u_{1:n})} \left\| \sum_{i \in B} \big(u_i - \phi(y_i)\big) \right\|_2^2.$$

When $d = 1$, $\phi(y) = y$, and B is quantile binning, $\mathrm{LinCE}_{\phi,\mathsf{B}}$ is a normalized version of the squared quantile-binned calibration error ℓ_2-QECE of Hartline et al. [2025], while $\mathrm{LinECE}_{\phi,\mathsf{B}}$ is its ℓ_1 analogue. Our truthfulness results below use the squared ℓ_2 form $\mathrm{LinCE}_{\phi,\mathsf{B}}$.

3 Truthful calibration measures

The key observation is that squared binned residuals admit a bias-variance decomposition for every linear property. This gives a single truthfulness theorem that covers binary calibration, full multiclass calibration, classwise calibration, and other moment properties of the outcome.

3.1 Truthful calibration error for linear properties

Theorem 3.1 (Truthfulness for linear properties)**.** *For every signal $\phi : \mathcal{Y} \to \mathbb{R}^d$ and every outcome-independent binning rule B, the measure $\mathrm{LinCE}_{\phi,\mathsf{B}}$ is truthful for the linear property $\Gamma_\phi(P) = \mathbf{E}_P[\phi(y)]$.*

More precisely, let $P_1,...,P_n \in \Delta(\mathcal{Y})$ be arbitrary, let $\mu_i^* = \Gamma_\phi(P_i)$, and draw $y_i \sim P_i$ independently. Then for every report $u_{1:n} \in \mathcal{R}_\phi^n$,

$$\mathbf{E}\big[\mathrm{LinCE}_{\phi,\mathsf{B}}(u_{1:n}; y_{1:n})\big] = \frac{m}{n^2} \sum_{B \in \mathsf{B}(u_{1:n})} \left\| \sum_{i \in B} (u_i - \mu_i^*) \right\|_2^2 + \frac{m}{n^2} \sum_{i=1}^{n} \mathbf{E}_{y_i \sim P_i}\big[\|\phi(y_i) - \mu_i^*\|_2^2\big].$$

Thus the truthful report $u_i = \mu_i^*$ makes the bias term vanish and minimizes the expected error. Moreover, in the scalar case $d = 1$, if the binning rule is quantile binning with $m_n \to \infty$ and $m_n/n \to 0$, then, for any fixed predictor $h : X \to \mathcal{R}_\phi$ and i.i.d. samples $(x_i, y_i) \sim D$,

$$\lim_{n \to \infty} \mathbf{E}[\mathrm{LinCE}_{\phi,\mathsf{B}_n}(h(x_{1:n}); y_{1:n})] = \mathbf{E}\Big[\big(h(x) - \mathbf{E}[\phi(y) \mid h(x)]\big)^2\Big].$$

Consequently, the sequence is complete and sound in the limit for calibration of Γ_ϕ.

The proof is deferred to Appendix C. Since the truthful value is only the sampling-variance term, it is $O(m/n)$ for bounded ϕ. For scalar reports, quantile binning with $m \to \infty$ and $m/n \to 0$ makes the remaining bias converge to the population squared calibration error for Γ_ϕ.

Classwise calibration as a linear property. For multiclass prediction, take $\phi_r(y) = \mathbb{I}[y = r]$ for each class r. Given probabilistic reports $p_i \in \Delta(\mathcal{Y})$, define the classwise squared binned error by running the scalar rule separately on each coordinate with a scalar outcome-independent binning rule B:

$$\mathrm{LinCE}^{\mathrm{cw}}_{\mathsf{B}}(p_{1:n}; y_{1:n}) := \frac{1}{k}\sum_{r=1}^{k} \mathrm{LinCE}_{\phi_r,\mathsf{B}}\big(p^{(r)}_{1:n}; y_{1:n}\big),$$

where $p_i^{(r)}$ is the r-th coordinate of p_i and B is applied to the scalar reports $p_1^{(r)},\ldots,p_n^{(r)}$ for each class r. By Theorem 3.1, this measure is truthful for classwise calibration. Equivalently, truthfulness is preserved by averaging the truthful scalar errors for the k one-vs-all linear properties, using the same prescribed bin count m for each class. The same vector construction also covers full multiclass calibration through $\phi(y) = e_y$; in the experiments below, we focus on the classwise specialization.

3.2 Confidence calibration

Confidence calibration differs from the linear properties above because the evaluated coordinate depends on the report itself. For a multiclass report $p_i \in \Delta(\mathcal{Y})$, let

$$r_i := \operatorname*{argmax}_{r \in \mathcal{Y}} p_i^{(r)}, \qquad c_i := p_i^{(r_i)}, \qquad z_i := \mathbb{I}[y_i = r_i].$$

The usual confidence aggregation applies a binary calibration error to the pairs (c_i, z_i), which is not generally truthful: a predictor can change which class becomes the argmax and thereby change the binary problem being evaluated. A concrete counterexample is given in Example C.1.

The following corrected confidence measure adds the classification-error term $\frac{m}{n}\big[1 - \frac{1}{n}\sum_{i=1}^{n} z_i\big]$. The common factor m preserves its weight relative to the squared term. The correction is at most m/n, so it vanishes in the consistent regime while restoring finite-sample truthfulness.

Definition 3.2 (Corrected squared confidence error)**.** *Let* B *be a binning rule applied to the scalar confidences* $c_{1:n}$*. Define*

$$\mathrm{ConfCE}^{\mathrm{corr}}_{\mathsf{B}}(p_{1:n}; y_{1:n}) := \frac{m}{n^2}\sum_{B \in \mathsf{B}(c_{1:n})}\left(\sum_{i \in B}(c_i - z_i)\right)^2 + \frac{m}{n}\left[1 - \frac{1}{n}\sum_{i=1}^{n} z_i\right].$$

Theorem 3.3 (Truthful confidence calibration)**.** *For every outcome-independent binning rule* B*,* $\mathrm{ConfCE}^{\mathrm{corr}}_{\mathsf{B}}$ *is truthful for confidence calibration. If the binning rule is quantile binning with* $m_n \to \infty$ *and* $m_n/n \to 0$*, then for any fixed predictor* h *and i.i.d. samples* $(x_i, y_i) \sim D$*,*

$$\lim_{n\to\infty} \mathbf{E}\big[\mathrm{ConfCE}^{\mathrm{corr}}_{\mathsf{B}_n}(h(x_{1:n}); y_{1:n})\big] = \mathbf{E}\Big[\big(c(x) - \mathbf{E}[z \mid c(x)]\big)^2\Big],$$

where $c(x) = \max_r h^{(r)}(x)$ *and* $z = \mathbb{I}[y = r_h(x)]$*. Thus the sequence is complete and sound in the limit for confidence calibration.*

The proof is deferred to Appendix C.

Why the correction is necessary. Let q_i denote the true probability of the reported top label r_i. Without the correction, the expected confidence error contains the variance term $\sum_i q_i(1-q_i)$ rather than $\sum_i (1-q_i^2)$. The former is not monotone in the selected true probability q_i, so changing the reported argmax can reduce the measured error even when the confidence report is not truthful. The correction changes the objective so that selecting the true most-likely class and reporting its probability is optimal.

4 Blackwell dominance of the constructed truthful errors

The calibration assumption must match the information revealed by a report. For a linear-property predictor, we assume calibration conditional on its entire reported property. For a multiclass probability predictor h, *full calibration* means $\Pr[y = r \mid h(x)] = h^{(r)}(x)$ almost surely for every r. Classwise calibration conditions only on each coordinate separately, and confidence calibration only on the maximum coordinate; neither implies full calibration. The multiclass Blackwell conclusions below require full calibration. We also give separate dominance guarantees under the weaker classwise and confidence notions, using restricted families of proper losses.

Blackwell dominance between predictions. A downstream user acts after seeing a prediction. For a predictor $h: X \to \mathcal{Z}_h$, define

$$\pi_h(z) := \Pr[y \mid h(x) = z]$$

for the posterior distribution over outcomes after observing the prediction z. For fully calibrated probability predictors, $\pi_h(h(x)) = h(x)$ almost surely. We say that prediction f *Blackwell dominates* prediction g, written $f \succeq_B g$, if a Markov kernel K can generate $\widetilde{G} \sim K(\cdot \mid f(x))$, conditionally independently of y given $f(x)$, such that $(\widetilde{G}, y)$ has the same law as $(g(x), y)$ [Blackwell, 1951]. Equivalently, for every bounded downstream decision problem $u: A \times \mathcal{Y} \to [0,1]$,

$$\mathbf{E}\left[\max_{a \in A} \mathbf{E}[u(a,y) \mid f(x)]\right] \geq \mathbf{E}\left[\max_{a \in A} \mathbf{E}[u(a,y) \mid g(x)]\right].$$

In words, after best-responding to the observed prediction, a decision-maker who sees $f(x)$ can achieve at least as much expected utility as one who sees $g(x)$, in every downstream task.

4.1 Consequence of truthfulness: preserving orders between calibrated predictors

The following result extends the Blackwell ordering of calibrated single reports under proper losses to truthful measures of independent batches.

Theorem 4.1 (Truthful batch measures preserve Blackwell dominance)**.** *Let* $\mathrm{CAL}_\phi : \mathcal{R}_\phi^n \times \mathcal{Y}^n \to \mathbb{R}$ *be truthful for the linear property* Γ_ϕ*. Let* $f, g: X \to \mathcal{R}_\phi$ *be calibrated for* Γ_ϕ*. If* $f \succeq_B g$*, then*

$$\mathbf{E}[\mathrm{CAL}_\phi(f(x_{1:n}); y_{1:n})] \leq \mathbf{E}[\mathrm{CAL}_\phi(g(x_{1:n}); y_{1:n})],$$

where $(x_i, y_i)_{i=1}^n$ *are drawn i.i.d. from* D*.*

Taking $\phi(y) = e_y$ in Theorem 4.1 gives the multiclass conclusion for any truthful measure of probability reports, including $\mathrm{LinCE}_{\mathsf{B}}^{\mathrm{cw}}$ and $\mathrm{ConfCE}_{\mathsf{B}}^{\mathrm{corr}}$, when f, g are *fully* calibrated. Classwise or confidence calibration alone does not justify this conclusion; counterexamples appear in Appendix D.4. Truthfulness also rewards canonical recalibration:

Theorem 4.2 (Canonical recalibration reduces expected truthful error)**.** *Let* $\mathrm{CAL}_\phi : \mathcal{R}_\phi^n \times \mathcal{Y}^n \to \mathbb{R}$ *be truthful for the linear property* Γ_ϕ*, and let* $h : X \to \mathcal{R}_\phi$ *be any predictor. Define its canonical recalibration by*

$$\bar{h}(x) := \mathbf{E}[\phi(y) \mid h(x)].$$

Then

$$\mathbf{E}\big[\mathrm{CAL}_\phi(\bar{h}(x_{1:n}); y_{1:n})\big] \leq \mathbf{E}[\mathrm{CAL}_\phi(h(x_{1:n}); y_{1:n})].$$

Proofs of Theorems 4.1 and 4.2 are deferred to Appendix D.

4.2 The order between approximately calibrated predictors

For our constructed errors, exact proper-loss identities yield ranking guarantees under the corresponding calibration notions. For miscalibrated predictors, a separate sufficient condition compares the recalibrated proper-loss gap with a nonnegative residual term.

Let CAL denote one of the constructed errors: $\mathrm{LinCE}_{\phi,\mathsf{B}}$, $\mathrm{LinCE}_{\mathsf{B}}^{\mathrm{cw}}$, or $\mathrm{ConfCE}_{\mathsf{B}}^{\mathrm{corr}}$. Define the associated losses by

$$\ell_\phi(u,y) := \|u - \phi(y)\|_2^2, \quad \ell_{\mathrm{Brier}}(p,y) := \sum_{r=1}^{k} \big(p^{(r)} - \mathbb{I}[y = r]\big)^2,$$

$$\ell_{\mathrm{conf}}(p,y) := 1 - \mathbb{I}[y = r_p] + \big(p^{(r_p)} - \mathbb{I}[y = r_p]\big)^2, \qquad r_p \in \operatorname*{argmax}_{r \in \mathcal{Y}} p^{(r)}.$$

Under linear-property, classwise, or confidence calibration, respectively, the expected error is m/n times the associated loss risk in the table (Lemma D.1). These identities need only the stated weaker calibration notions; interpreting the reports as posterior distributions in a multiclass Blackwell comparison requires full calibration.

constructed error CAL	associated loss ℓ_{CAL}
$\mathrm{LinCE}_{\phi,\mathsf{B}}$	$\ell_\phi(u,y) = \|u - \phi(y)\|_2^2$
$\mathrm{LinCE}_{\mathsf{B}}^{\mathrm{cw}}$	$\ell_{\mathrm{Brier}}(p,y)/k$
$\mathrm{ConfCE}_{\mathsf{B}}^{\mathrm{corr}}$	$\ell_{\mathrm{conf}}(p,y)$

Dominance for the weaker calibration notions. Let $\mathcal{L}_{\mathrm{cw}}$ be the family of all bounded losses $\ell(p,y)=k^{-1}\sum_r s_r(p_r,\mathbb{I}[y=r])$, where each s_r is a proper binary loss. Let $\mathcal{L}_{\mathrm{conf}}$ be the family of all bounded losses $\ell(p,y)=s(c_p,\mathbb{I}[y=r_p])$ that are proper as *multiclass* losses, where $c_p=\max_r p_r$ and $r_p=\mathrm{argmax}_r p_r$. Here properness means $\mathbf{E}_{y\sim q}[\ell(q,y)]\le\mathbf{E}_{y\sim q}[\ell(p,y)]$ for every $p,q\in\Delta(\mathcal{Y})$. For $t\in\{\mathrm{cw},\mathrm{conf}\}$, write $f\succeq_t g$ if $\mathbf{E}[\ell(f(x),y)]\le\mathbf{E}[\ell(g(x),y)]$ for every $\ell\in\mathcal{L}_t$. These orders compare reported-loss risks, not all downstream decision problems.

Theorem 4.3 (Dominance under classwise or confidence calibration)**.** *For i.i.d. samples and the same prescribed n,m, if f,g are classwise calibrated and $f\succeq_{\mathrm{cw}} g$, then $\mathbf{E}[\mathrm{LinCE}^{\mathrm{cw}}_{\mathsf{B}}(f)]\le\mathbf{E}[\mathrm{LinCE}^{\mathrm{cw}}_{\mathsf{B}}(g)]$. If f,g are confidence calibrated and $f\succeq_{\mathrm{conf}} g$, then $\mathbf{E}[\mathrm{ConfCE}^{\mathrm{corr}}_{\mathsf{B}}(f)]\le\mathbf{E}[\mathrm{ConfCE}^{\mathrm{corr}}_{\mathsf{B}}(g)]$. Both conclusions hold for every outcome-independent scalar binning rule.*

The proof and the relation to full-calibration Blackwell dominance are in Appendix D.4.

Any of these losses may be rescaled by a deterministic positive constant, if desired, without affecting the ordering statements below. For a possibly miscalibrated predictor h, let $\bar{h}$ be its canonical recalibration: it keeps the same prediction $h(x)$ but replaces the report by the true conditional property given that prediction. Thus, for a linear property,

$$\bar{h}_\phi(x):=\mathbf{E}[\phi(y)\,|\,h(x)],$$

and for full multiclass prediction, $\bar{h}(x):=\pi_h(h(x))$.

For each constructed error, define the calibration-bias term $\rho_{\mathrm{CAL}}(h)\ge 0$ by the decomposition

$$\frac{n}{m}\mathbf{E}[\mathrm{CAL}(h)]=\mathbf{E}\big[\ell_{\mathrm{CAL}}(\bar{h}(x),y)\big]+\rho_{\mathrm{CAL}}(h).$$

For $\mathrm{LinCE}_{\phi,\mathsf{B}}$, this bias is the squared binned residual between the reported property $h(x)$ and the canonical property $\bar{h}_\phi(x)$. For corrected confidence calibration, it includes both confidence miscalibration and the regret from selecting a label that is not most likely under the posterior. For both multiclass errors, $\bar{h}$ uses the full posterior conditional on the entire probability vector. Consequently, $\rho_{\mathrm{CAL}}(h)$ need not vanish under classwise or confidence calibration alone; the approximate guarantee requires its stated bound separately.

Theorem 4.4 (Approximate calibrated orders for the constructed errors)**.** *Fix one of the constructed errors $\mathrm{CAL}\in\{\mathrm{LinCE}_{\phi,\mathsf{B}},\mathrm{LinCE}^{\mathrm{cw}}_{\mathsf{B}},\mathrm{ConfCE}^{\mathrm{corr}}_{\mathsf{B}}\}$, with the same prescribed bin count m for both predictors, and let ℓ_{CAL} be its associated proper loss. Suppose predictors f and g satisfy*

$$\rho_{\mathrm{CAL}}(f)\le\epsilon.$$

If their canonically recalibrated proper-loss risks are separated by

$$\mathbf{E}\big[\ell_{\mathrm{CAL}}(\bar{f}(x),y)\big]+\epsilon\le\mathbf{E}[\ell_{\mathrm{CAL}}(\bar{g}(x),y)],$$

then the constructed calibration error preserves the ranking:

$$\mathbf{E}[\mathrm{CAL}(f)]\le\mathbf{E}[\mathrm{CAL}(g)].$$

Equivalently, in payoff form, if the bounded scoring payoff associated with ℓ_{CAL} favors $\bar{f}$ over $\bar{g}$ by at least ϵ, then the constructed truthful error preserves the ranking.

The proof is deferred to Appendix D. Equivalently, the associated bounded scoring task favors $\bar{f}$ over $\bar{g}$ by the same margin.

5 Empirical evaluations

Our experiments follow the standard post-hoc calibration pipeline used in empirical studies of neural-network calibration, such as Minderer et al. [2021]. We first train neural-network classifiers on standard multiclass prediction tasks, then apply simple post-hoc calibration methods to their predicted probabilities.

Experimental goal. The purpose of the experiments is to examine how the calibration errors analyzed in our theory behave in a realistic neural-network setting and to clarify the source of the inconsistent rankings observed in prior work. Different calibration errors, and even the same error with different hyperparameters, such as the number of bins, naturally quantify calibration bias in

different ways. We therefore should not expect all errors to produce identical numerical values. However, they should still respect certain decision-theoretic principles. Full calibration supports the Blackwell guarantee, whereas classwise and confidence calibration support the respective proper-loss orders in Theorem 4.3. For miscalibrated predictors, Theorem 4.4 gives a sufficient condition involving the residual term and the recalibrated loss gap.

The experiments examine how the constructed errors track their associated losses after post-hoc calibration and how these rankings change with binning. Closeness to the classwise or confidence calibrated-loss identity is an empirical diagnostic. It does not establish full calibration, bound the full-posterior residual ρ_{CAL}, or verify the recalibrated loss gap in Theorem 4.4. We therefore interpret these comparisons as empirical ranking behavior, rather than verification of that theorem.

5.1 Experimental setup

Dataset. We use the `CIFAR-100` dataset, split into 50000 training images, 5000 validation images, and 5000 test images [Krizhevsky, 2009]. The validation split is used only for temperature scaling, and all reported empirical results are evaluated on the test split.

Models. We evaluate nine pretrained image classifiers fine-tuned on `CIFAR-100`: MobileNetV3-Small-0.5/0.75/1.0 and ResNet10t/18/34/50/101/152. For each training run, we keep model checkpoints throughout fine-tuning and treat them as a sequence of probabilistic predictors along training. Following Minderer et al. [2021], we apply temperature scaling to every checkpoint, with the temperature parameter chosen to minimize the cross-entropy on the validation set.

Evaluation pool. Our experiments use two complementary collections of predictor traces. For the within-model experiments, we focus on MobileNetV3-Small-1.0 and collect 40 training traces obtained from random training subsets of different sizes. This provides a dense view of how calibration errors behave across checkpoints within a single architecture. For the cross-model experiments, we consider all nine architectures and use 5 training traces for each model, again obtained from random training subsets of different sizes. This yields a diverse pool of predictors for comparing calibration rankings across architectures.

Binning. Empirically, we find that fixed-bin and quantile-binned evaluations give very similar trends in our experiments. Their main difference is therefore theoretical rather than empirical: truthfulness holds for any outcome-independent binning rule, while quantile binning is used for the limiting completeness/soundness guarantee and for the main empirical presentation. For this reason, we use quantile-binned calibration errors in the main text and defer most fixed-bin results to the appendix. When we need to specify the binning rule and the number of bins explicitly for any calibration error CAL, we write $\mathrm{CAL}_{(\mathrm{qua},m)}$ for the version with m quantile bins and $\mathrm{CAL}_{(\mathrm{fix},m)}$ for the version with m fixed bins.

Evaluation metrics. We evaluate predictors using several losses, including classification error, cross-entropy loss, Brier loss, spherical loss, and confidence loss. Across our model checkpoints, these losses exhibit nearly linear trends with one another, so for simplicity we use classification error as the primary reference loss. We compare our truthful $\mathrm{LinCE}_{\mathsf{B}}^{\mathrm{cw}}$ (classwise calibration) and $\mathrm{ConfCE}_{\mathsf{B}}^{\mathrm{corr}}$ (confidence calibration) against the widely used $\mathrm{ConfCE}_{\mathsf{B}}$ and $\mathrm{ConfECE}_{\mathsf{B}}$.

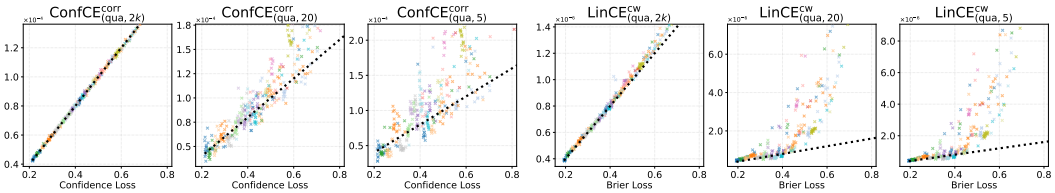
Plotting scale. Squared-error plots use the $1/n^2$ prefactor: our constructed errors are displayed divided by m, including the entire confidence correction, as specified in the captions. The uncorrected baseline and ConfECE are unchanged. Multiplying by m recovers the theoretical scale and preserves all rankings and Spearman correlations, including those between bin counts. Plot titles identify metric families.

5.2 Results

Assumption validation. After temperature scaling, the evaluated checkpoints lie close to the calibrated-predictor loss lines for the confidence and classwise measures. These diagnostics concern the corresponding weaker calibration notions; they do not establish full calibration of the reported probability vectors. We defer the plots to Appendix E.1.

Dominance preservation. The full-calibration Blackwell guarantee motivates comparing reported proper-loss risks, but evaluating finitely many losses does not establish Blackwell dominance. Predictor f empirically dominates g if it has no larger risk on every evaluated loss and strictly lower risk on at least one. The plots show predictors from a maximum-cardinality chain in this order, which

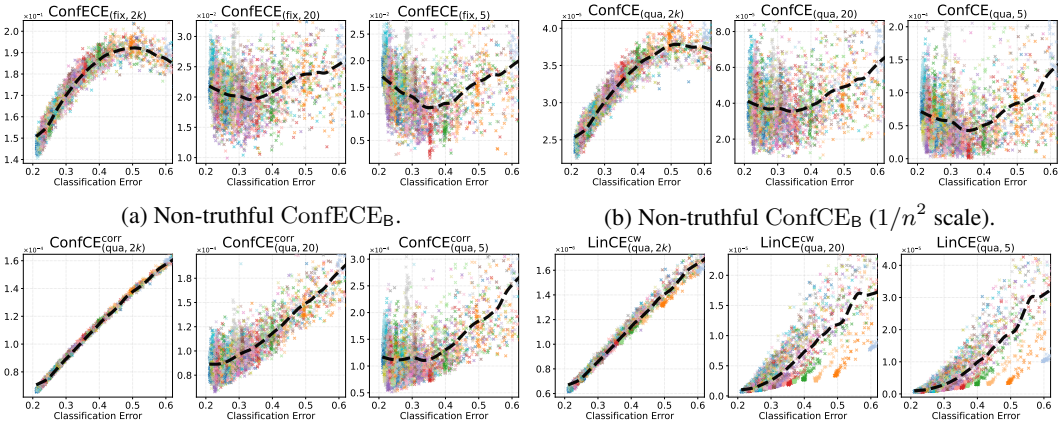
we call a maximum dominance set. Figure 2 shows that both truthful calibration errors track this ordering closely. The ordering is stable across binning choices, and the small number of remaining inversions is consistent with sampling noise and slight residual miscalibration. The analogous within-model plots are deferred to Appendix E.2.



(a) $\mathrm{ConfCE}_{\mathrm{B}}^{\mathrm{corr}}/m$ on the maximum dominance set. (b) $\mathrm{LinCE}_{\mathrm{B}}^{\mathrm{cw}}/m$ on the maximum dominance set.

Figure 2: Empirical validation of dominance preservation on the cross-model maximum dominance set. Each panel visualizes only the predictors belonging to the extracted dominance total order; predictors outside this order are omitted. The panels plot corrected confidence error against confidence loss and classwise calibration error against Brier loss. Our truthful calibration errors achieve high rank correlation, and the few remaining inversions can be attributed to sampling error and the fact that the predictors are only nearly calibrated. The analogous within-model plots are deferred to the appendix.

Binning robustness. Prior empirical work reported that the relationship between calibration and predictive performance can flip when only the number of bins changes. We now move beyond ranking preservation on the selected dominance subset and quantitatively examine all predictors in the within-model traces. Figure 3 shows that this instability is concentrated in the non-truthful confidence-based errors: their curves can rise again late in training, making better predictors appear less calibrated. By contrast, for the truthful calibration errors $\mathrm{ConfCE}_{\mathrm{B}}^{\mathrm{corr}}$ and $\mathrm{LinCE}_{\mathrm{B}}^{\mathrm{cw}}$, the trend is stable across binning choices, with lower classification error continuing to align with lower calibration error. Thus, truthful errors are quantitatively more robust to the choice of binning size. This supports our interpretation that the observed flip is a measurement artifact of non-truthful calibration errors rather than a property of the underlying predictors. We provide the analogous cross-model plots in the appendix, and Table 1 reports the corresponding rank correlations across binning choices.



(a) Non-truthful $\mathrm{ConfECE}_{\mathrm{B}}$. (b) Non-truthful $\mathrm{ConfCE}_{\mathrm{B}}$ ($1/n^2$ scale).

(c) Truthful $\mathrm{ConfCE}_{\mathrm{B}}^{\mathrm{corr}}$, divided by m. (d) Truthful $\mathrm{LinCE}_{\mathrm{B}}^{\mathrm{cw}}$, divided by m.

Figure 3: Binning robustness on the within-model MobileNetV3-Small-1.0 traces. Each panel plots calibration error against classification error for temperature-scaled checkpoints, and the curves are kernel-regression trend lines for different bin sizes. The truthful calibration errors in (c) and (d) remain monotone: lower-loss predictors also have lower calibration error across binning choices. By contrast, the non-truthful calibration errors in (a) and (b) can flip their relationship with predictive loss as the binning size changes.

We discuss the limitations in Appendix A.

Acknowledgments and Disclosure of Funding

The work was done while Yuxuan Lu was a visiting student researcher at Northwestern University, supported by the IDEAL Institute.

References

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

D. Blackwell. Comparison of experiments. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, volume 2, pages 93–103. University of California Press, 1951.

J. Blasiok and P. Nakkiran. Smooth ece: Principled reliability diagrams via kernel smoothing. In *The Twelfth International Conference on Learning Representations*, 2024.

J. Błasiok, P. Gopalan, L. Hu, and P. Nakkiran. A unifying theory of distance from calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1727–1740, 2023.

A. P. Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982. doi: 10.1080/01621459.1982.10477856. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1982.10477856.

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

P. Gopalan, L. Hu, and G. N. Rothblum. On computationally efficient multi-class calibration. In S. Agrawal and A. Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 1983–2026. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/gopalan24a.html.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

C. Gupta and A. Ramdas. Top-label calibration and multiclass-to-binary reductions. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=WqoBaaPHS-.

N. Haghtalab, M. Qiao, K. Yang, and E. Zhao. Truthfulness of calibration measures. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

J. Hartline, L. Hu, and Y. Wu. A perfectly truthful calibration measure. *arXiv preprint arXiv:2508.13100*, 2025.

M. Hegazy, M. I. Jordan, and A. Dieuleveut. Scalable utility-aware multiclass calibration. In E. Khan, Y. Li, A. Solin, and A. Ramdas, editors, *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pages 397–405. PMLR, 2026. URL https://proceedings.mlr.press/v300/hegazy26a.html.

A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

M. Kull, M. Perello Nieto, M. Kängsepp, T. Silva Filho, H. Song, and P. Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/8ca01ea920679a0fe3728441494041b9-Paper.pdf.

A. Kumar, P. S. Liang, and T. Ma. Verified uncertainty calibration. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/f8c0c968632845cd133308b1a494967f-Paper.pdf.

M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran, and M. Lucic. Revisiting the calibration of modern neural networks. *Advances in neural information processing systems*, 34:15682–15694, 2021.

J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran. Measuring calibration in deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 38–41, 2019.

M. Qiao and E. Zhao. Truthfulness of decision-theoretic calibration measures. *arXiv preprint arXiv:2503.02384*, 2025.

J. Tang, J. Wu, Z. S. Wu, and J. Zhang. Dimension-free decision calibration for nonlinear loss functions. *arXiv preprint arXiv:2504.15615*, 2025.

D. Widmann, F. Lindsten, and D. Zachariah. Calibration tests in multi-class classification: A unifying framework. In *Advances in Neural Information Processing Systems*, volume 32, pages 12257–12267, 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/hash/1c336b8080f82bcc2cd2499b4c57261d-Abstract.html`.

S. Zhao, M. Kim, R. Sahoo, T. Ma, and S. Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 22313–22324. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/file/bbc92a647199b832ec90d7cf57074e9e-Paper.pdf`.

A Limitations

The multiclass Blackwell guarantees require full calibration of the reported probability vectors. Classwise and confidence calibration alone support the respective proper-loss identities and restricted dominance guarantees, but not the full Blackwell conclusion. For miscalibrated predictors, our approximate ranking result assumes that the remaining calibration bias is small relative to the gap in the associated downstream loss. Thus, the theory does not claim that truthful calibration errors recover decision-theoretic rankings for arbitrary predictors; it characterizes the regime in which calibration errors can be interpreted as stable proxies for downstream decision quality.

Our experiments examine the ranking behavior of temperature-scaled predictors under the standard post-hoc calibration pipeline used in prior work. The results therefore should not be interpreted as applying unchanged to settings where the predictor is substantially miscalibrated. This limitation is especially relevant for post-trained language models, whose reported probabilities or confidence scores can remain poorly calibrated after instruction tuning, reinforcement learning, or other post-training procedures, e.g., [Achiam et al., 2023]. In such regimes, the calibration-bias term may dominate the downstream loss gap, and the ranking guarantees in this paper need not hold.

B Additional preliminaries

B.1 Calibration measure: completeness and soundness

The two requirements for a calibration measure are *completeness* and *soundness*. Completeness and soundness ensure that the calibration error vanishes only for calibrated predictors and does not vanish for miscalibrated predictors. Note that proper losses are truthful, but properness alone does not ensure completeness. For example, if $y \sim \mathrm{Bernoulli}(1/2)$, the constant predictor $h(x) = 1/2$ is perfectly calibrated, while its expected average squared loss is $1/4$ for every sample size. Thus, squared loss is sound but not complete as a calibration measure.

Definition B.1 (Completeness)**.** *A sequence of calibration measures* $\{\mathrm{CAL}_n\}_{n\geq 1}$ *is* complete in the limit *if, for any calibrated predictor* f*,*

$$\lim_{n\to\infty} \mathbf{E}[\mathrm{CAL}_n((f(x_i))_{i=1}^n, (y_i)_{i=1}^n)] = 0.$$

Definition B.2 (Soundness)**.** *A sequence of calibration measures* $\{\mathrm{CAL}_n\}_{n\geq 1}$ *is* sound in the limit *if, for any miscalibrated predictor* f*,*

$$\liminf_{n\to\infty} \mathbf{E}[\mathrm{CAL}_n((f(x_i))_{i=1}^n, (y_i)_{i=1}^n)] > 0.$$

In this paper, we focus on several weaker definitions of high-dimensional calibration. The classwise aggregation focuses on classwise calibration [Kull et al., 2019]. Instead of conditioning on the high-dimensional prediction, classwise calibration requires unbiasedness conditioned on a prediction of a single dimension.

Definition B.3 (Classwise Calibration [Kull et al., 2019])**.** *A predictor $f : X \to \Delta(\mathcal{Y})$ is (perfectly) classwise* calibrated *if for every $p \in [0,1]$ and every $r \in \mathcal{Y} = \{1,...,k\}$,*

$$\Pr_{(x,y)\sim D}[y = r | f_r(x) = p] = p.$$

The confidence aggregation is induced from confidence calibration. Confidence calibration requires unbiasedness conditioned on the highest prediction.

Definition B.4 (Confidence Calibration)**.** *A predictor $f : X \to \Delta(\mathcal{Y})$ is (perfectly) confidence* calibrated *if for every $p \in [0,1]$,*

$$\Pr_{(x,y)\sim D}[y = \operatorname{argmax}_r f_r(x) | \max_r f_r(x) = p] = p,$$

where argmax *uses the fixed deterministic tie-breaking rule from the main text.*

B.2 Commonly used proper losses

We define the Brier loss and the classification error here.

Definition B.5 (Brier Loss)**.** *The multiclass Brier loss is*

$$l_{\text{Brier}}(p,y) := \sum_{r=1}^{k} \left(p^{(r)} - \mathbb{I}[y = r]\right)^2.$$

Definition B.6 (Classification Error)**.** *The multiclass classification error is*

$$l_{\text{Cla}}(p,y) := \mathbb{I}\left[y \neq \operatorname*{argmax}_{r \in \{1,...,k\}} p_r\right].$$

C Proofs for Section 3

Proof of Theorem 3.1. Fix arbitrary reports $u_{1:n} \in \mathcal{R}_\phi^n$, and let

$$\mu_i^* := \Gamma_\phi(P_i) = \mathbf{E}_{y_i \sim P_i}[\phi(y_i)]$$

for $i = 1,...,n$. Because the binning rule B is outcome-independent, conditioning on $u_{1:n}$ fixes the partition $\mathsf{B}(u_{1:n})$. For each i, define the centered noise

$$\xi_i := \phi(y_i) - \mu_i^*.$$

Then $\mathbf{E}[\xi_i] = 0$, and the variables $\xi_1,...,\xi_n$ are independent.

For any bin $B \in \mathsf{B}(u_{1:n})$,

$$\sum_{i\in B}(u_i - \phi(y_i)) = \sum_{i\in B}(u_i - \mu_i^*) - \sum_{i\in B}\xi_i.$$

Taking squared Euclidean norms and expectations, the cross term vanishes because the second sum has mean zero. Independence also removes the cross-covariances among the ξ_i, so

$$\begin{aligned}\mathbf{E}\left[\left\|\sum_{i\in B}(u_i - \phi(y_i))\right\|_2^2\right] &= \left\|\sum_{i\in B}(u_i - \mu_i^*)\right\|_2^2 + \sum_{i\in B}\mathbf{E}\left[\|\xi_i\|_2^2\right] \\ &= \left\|\sum_{i\in B}(u_i - \mu_i^*)\right\|_2^2 + \sum_{i\in B}\mathbf{E}_{y_i\sim P_i}\left[\|\phi(y_i) - \mu_i^*\|_2^2\right].\end{aligned}$$

Summing over bins and multiplying by the report-independent factor m/n^2 yields

$$\mathbf{E}\big[\mathrm{LinCE}_{\phi,\mathsf{B}}(u_{1:n};y_{1:n})\big]=\frac{m}{n^2}\sum_{B\in\mathsf{B}(u_{1:n})}\Big\|\sum_{i\in B}(u_i-\mu_i^*)\Big\|_2^2 + \frac{m}{n^2}\sum_{i=1}^{n}\mathbf{E}_{y_i\sim P_i}\big[\|\phi(y_i)-\mu_i^*\|_2^2\big].$$

The second term is independent of the report, while the first term is nonnegative and vanishes at the truthful report $u_i=\mu_i^*$ for all i. This proves truthfulness.

For the scalar limiting claim, let $U=h(x)$ and $V=\phi(y)$, so $U,V\in[a,b]=\mathcal{R}_\phi$. If $a=b$, the error and the population calibration error are both zero. Otherwise, write $r(u)=u-\mathbf{E}[V\mid U=u]$ and $M=b-a$, so $|r(u)|\le M$ almost surely. Let $(U_i,V_i)_{i=1}^n$ be i.i.d. copies of (U,V) and set $m=m_n$. For each quantile bin B, let $s_B=|B|$ and $\bar r_B=s_B^{-1}\sum_{i\in B}r(U_i)$. Conditional independence gives

$$\mathbf{E}[\mathrm{LinCE}_{\phi,\mathsf{B}_n}\mid U_{1:n}]=Q_n+\frac{m}{n^2}\sum_{i=1}^{n}\mathrm{Var}(V_i\mid U_i),\qquad Q_n:=\frac{m}{n^2}\sum_B s_B^2\bar r_B^2.$$

The variance term has expectation $(m/n)\mathbf{E}[\mathrm{Var}(V\mid U)]=O(m/n)\to 0$.

To identify the bias limit, define

$$A_n:=\frac{1}{n}\sum_B s_B\bar r_B^2,\qquad R_n:=\frac{1}{n}\sum_B\sum_{i\in B}(r(U_i)-\bar r_B)^2.$$

The within-bin variance identity gives $A_n=n^{-1}\sum_i r(U_i)^2-R_n$. Since every bin size is either $\lfloor n/m\rfloor$ or $\lceil n/m\rceil$, we have $|ms_B/n-1|\le m/n$ and hence

$$|Q_n-A_n|\le\frac{m}{n}A_n\le M^2\frac{m}{n}\longrightarrow 0.$$

It remains to show $\mathbf{E}[R_n]\to 0$. For any $\eta>0$, choose a step function q, constant on a partition of $[a,b]$ into K intervals, with $|q|\le M$ and $\mathbf{E}[(r(U)-q(U))^2]<\eta$. Such functions are dense in $L^2(P_U)$, including when P_U has atoms. Let $\bar q_B=s_B^{-1}\sum_{i\in B}q(U_i)$. Since the bin average minimizes the sum of squared deviations within its bin,

$$R_n\le\frac{2}{n}\sum_i(r(U_i)-q(U_i))^2+\frac{2}{n}\sum_B\sum_{i\in B}(q(U_i)-\bar q_B)^2.$$

The reports are sorted, so at most $K-1$ bins straddle a boundary between intervals on which q is constant. Each such bin has at most $\lceil n/m\rceil$ samples; all other bins contribute zero to the second sum. Thus

$$\mathbf{E}[R_n]\le 2\eta+8M^2(K-1)\left(\frac{1}{m}+\frac{1}{n}\right).$$

Taking $n\to\infty$ and then $\eta\downarrow 0$ proves $\mathbf{E}[R_n]\to 0$. Combining the preceding bounds yields

$$\lim_{n\to\infty}\mathbf{E}[\mathrm{LinCE}_{\phi,\mathsf{B}_n}]=\mathbf{E}[r(U)^2]=\mathbf{E}\big[(h(x)-\mathbf{E}[\phi(y)\mid h(x)])^2\big].$$

This limit is zero exactly when h is calibrated for Γ_ϕ, which proves completeness and soundness. The argument allows tied reports and does not require n to be divisible by m. For example, when $U\equiv 0$ and $V\equiv 1$, the statistic is $(m/n^2)\sum_B s_B^2\to 1$ (exactly 1 for equal-sized bins), rather than vanishing as $1/m$. □

For fixed k, averaging the k scalar limits proves that $\mathrm{LinCE}^{\mathrm{cw}}_{\mathsf{B}_n}$ converges in expectation to $k^{-1}\sum_{r=1}^k\mathbf{E}[(h^{(r)}(x)-\Pr(y=r\mid h^{(r)}(x)))^2]$. This is zero exactly for classwise-calibrated predictors, proving classwise completeness and soundness in the consistent quantile regime.

Example C.1 (Confidence aggregation can be non-truthful)**.** *Let $k\ge 3$, and let j be the label selected by the fixed deterministic tie-breaking rule when the reported distribution is uniform. Choose two distinct labels $s,t\ne j$, and suppose the true posterior is the same on every sample:*

$$p_i^*=p^*:=\frac{1}{2}e_s+\frac{1}{k}e_j+\left(\frac{1}{2}-\frac{1}{k}\right)e_t.$$

Thus class s is uniquely most likely, while class j has true probability $1/k$. Consider applying the truthful binary calibration error $\mathrm{LinCE}_{\mathrm{id},\mathsf{B}}$ to the confidence pairs (c_i,z_i) produced by confidence aggregation.

Under truthful reporting, each selected confidence equals $c_i=1/2$, and the selected class is correct with probability $1/2$, so $z_i\sim\mathrm{Bernoulli}(1/2)$. The residuals have mean zero, so independence gives, for any binning rule with prescribed bin count m,

$$\mathbf{E}[\mathrm{LinCE}_{\mathrm{id},\mathsf{B}}(c_{1:n};z_{1:n})]=\frac{m}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n}z_i\right]=\frac{m}{4n}.$$

If instead the predictor always reports the uniform distribution $(1/k,\dots,1/k)$, deterministic tie-breaking selects class j. Then $c_i=1/k$ and $z_i\sim\mathrm{Bernoulli}(1/k)$. With the same prescribed bin count m, the residuals again have mean zero, so

$$\mathbf{E}[\mathrm{LinCE}_{\mathrm{id},\mathsf{B}}(c_{1:n};z_{1:n})]=\frac{m}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n}z_i\right]=\frac{m}{n}\frac{1}{k}\left(1-\frac{1}{k}\right)<\frac{m}{4n}.$$

Thus confidence aggregation can assign smaller expected error to a non-truthful report, even when the underlying binary calibration error is truthful.

Proof of Theorem 3.3. Let $p_1^*,\dots,p_n^*\in\Delta(\mathcal{Y})$ be the true class distributions and draw $y_i\sim p_i^*$ independently. Fix arbitrary reported distributions $p_1,\dots,p_n$, and write

$$r_i:=\operatorname*{argmax}_{r\in\mathcal{Y}}p_i^{(r)},\qquad c_i:=p_i^{(r_i)},\qquad q_i:=(p_i^*)^{(r_i)},\qquad z_i:=\mathbb{I}[y_i=r_i].$$

Conditional on the reports, the selected classes r_i, confidences c_i, and bins $\mathsf{B}(c_{1:n})$ are fixed, while the labels $z_i\sim\mathrm{Bernoulli}(q_i)$ remain independent.

For any bin $B\in\mathsf{B}(c_{1:n})$,

$$\sum_{i\in B}(c_i-z_i)=\sum_{i\in B}(c_i-q_i)-\sum_{i\in B}(z_i-q_i).$$

The second sum has conditional mean zero, so as in the previous proof,

$$\mathbf{E}\left[\left(\sum_{i\in B}(c_i-z_i)\right)^2\middle|p_{1:n}\right]=\left(\sum_{i\in B}(c_i-q_i)\right)^2+\sum_{i\in B}q_i(1-q_i).$$

Summing over bins gives

$$\begin{aligned}&\mathbf{E}\left[\frac{m}{n^2}\sum_{B\in\mathsf{B}(c_{1:n})}\left(\sum_{i\in B}(c_i-z_i)\right)^2\middle|p_{1:n}\right]\\&\quad=\frac{m}{n^2}\sum_{B\in\mathsf{B}(c_{1:n})}\left(\sum_{i\in B}(c_i-q_i)\right)^2+\frac{m}{n^2}\sum_{i=1}^{n}q_i(1-q_i).\end{aligned}$$

The correction term contributes

$$\mathbf{E}\left[\frac{m}{n}\left(1-\frac{1}{n}\sum_{i=1}^{n}z_i\right)\middle|p_{1:n}\right]=\frac{m}{n^2}\sum_{i=1}^{n}(1-q_i).$$

Therefore

$$\frac{n^2}{m}\mathbf{E}\left[\mathrm{ConfCE}_{\mathsf{B}}^{\mathrm{corr}}(p_{1:n};y_{1:n})\,|\,p_{1:n}\right]=\sum_{B\in\mathsf{B}(c_{1:n})}\left(\sum_{i\in B}(c_i-q_i)\right)^2+\sum_{i=1}^{n}(1-q_i^2).$$

The first term is nonnegative and, for fixed selected classes $r_1,\dots,r_n$, is minimized by setting $c_i=q_i$ for every i. The second term depends only on the selected classes and is minimized by maximizing

each q_i, that is, by choosing $r_i \in \operatorname{argmax}_{r\in\mathcal{Y}}(p_i^*)^{(r)}$. The truthful report $p_i = p_i^*$ attains both conditions simultaneously (with deterministic tie-breaking selecting one maximizer when several classes tie). Hence no alternative report has smaller expected error, proving truthfulness.

For limiting completeness and soundness, fix a predictor $h: X \to \Delta(\mathcal{Y})$ and let

$$c(x) := \max_{r\in\mathcal{Y}} h^{(r)}(x), \qquad z := \mathbb{I}[y = r_h(x)],$$

where $r_h(x)$ is the top label selected by the fixed deterministic tie-breaking rule. Along a consistent quantile regime $\mathsf{B}_n = \mathsf{B}^Q_{m_n}$, the corrected confidence error decomposes as

$$\mathrm{ConfCE}^{\mathrm{corr}}_{\mathsf{B}_n}(h(x_{1:n}); y_{1:n}) = \mathrm{LinCE}_{\mathrm{id},\mathsf{B}_n}(c(x_{1:n}); z_{1:n}) + \frac{m_n}{n}\left[1 - \frac{1}{n}\sum_{i=1}^{n} z_i\right].$$

The correction term is always between 0 and m_n/n, so it vanishes in expectation because $m_n/n \to 0$. Applying the scalar limit just proved with $U = c(x)$ and $V = z$ gives

$$\lim_{n\to\infty} \mathbf{E}\big[\mathrm{ConfCE}^{\mathrm{corr}}_{\mathsf{B}_n}(h(x_{1:n}); y_{1:n})\big] = \mathbf{E}\big[(c(x) - \mathbf{E}[z \mid c(x)])^2\big].$$

This limit is zero exactly for confidence-calibrated h and strictly positive otherwise, proving completeness and soundness. □

D Proofs for Section 4

D.1 Proof of Theorem 4.1

Proof of Theorem 4.1. Let $F_i := f(x_i)$ for $i = 1,\ldots,n$. By Blackwell dominance, there is a Markov kernel K for which a simulated report $\widetilde{G}_i \sim K(\cdot \mid F_i)$ is conditionally independent of y_i given F_i, and $(\widetilde{G}_i, y_i)$ has the same law as $(g(x_i), y_i)$. Use independent kernel draws for the i.i.d. samples. Conditional on $F_{1:n}, \widetilde{G}_{1:n}$, the outcomes $y_1,\ldots,y_n$ remain independent. This uses a garbling coupling, not an assumption that the original same-input pair $(f(x), g(x))$ is itself Markov with y.

Since f is calibrated for Γ_ϕ,

$$\mathbf{E}[\phi(y_i) \mid F_i] = F_i.$$

Because $\widetilde{G}_i$ is sampled from F_i without access to y_i, conditioning on it does not change this conditional property value:

$$\mathbf{E}[\phi(y_i) \mid F_{1:n}, \widetilde{G}_{1:n}] = \mathbf{E}[\phi(y_i) \mid F_i] = F_i.$$

Thus, after conditioning on $F_{1:n}, \widetilde{G}_{1:n}$, the n conditional outcome distributions have truthful reports $F_1,\ldots,F_n$, while $\widetilde{G}_1,\ldots,\widetilde{G}_n$ are alternative reports in $\mathcal{R}_\phi$. By the truthfulness of CAL_ϕ,

$$\mathbf{E}\Big[\mathrm{CAL}_\phi(F_{1:n}; y_{1:n}) \Big| F_{1:n}, \widetilde{G}_{1:n}\Big] \le \mathbf{E}\Big[\mathrm{CAL}_\phi(\widetilde{G}_{1:n}; y_{1:n}) \Big| F_{1:n}, \widetilde{G}_{1:n}\Big].$$

Taking expectations and using the equality in law of $(\widetilde{G}_{1:n}, y_{1:n})$ and $(g(x_{1:n}), y_{1:n})$ proves the claim. □

D.2 Proof of Theorem 4.2

Proof of Theorem 4.2. Let $U_i := h(x_i)$ and $\bar{U}_i := \bar{h}(x_i)$ for $i = 1,\ldots,n$, where

$$\bar{U}_i = \mathbf{E}[\phi(y_i) \mid U_i].$$

Because the samples are i.i.d., conditional on $U_{1:n}$ the outcomes $y_1,\ldots,y_n$ are independent, and the truthful report for the i-th conditional distribution is exactly $\bar{U}_i$. Therefore, by truthfulness of CAL_ϕ,

$$\mathbf{E}\big[\mathrm{CAL}_\phi(\bar{U}_{1:n}; y_{1:n}) \big| U_{1:n}\big] \le \mathbf{E}[\mathrm{CAL}_\phi(U_{1:n}; y_{1:n}) | U_{1:n}].$$

Taking expectation over $U_{1:n}$ yields

$$\mathbf{E}\big[\mathrm{CAL}_\phi(\bar{h}(x_{1:n}); y_{1:n})\big] \le \mathbf{E}[\mathrm{CAL}_\phi(h(x_{1:n}); y_{1:n})].$$ □

D.3 Proof of associated losses

Lemma D.1 (Proper-loss identities for the constructed errors)**.** *Let $(x_i,y_i)_{i=1}^n$ be drawn i.i.d. from D, and let m be the prescribed, report-independent bin count (the same for every class in the classwise construction).*

1. If $h: X \to \mathcal{R}_\phi$ is calibrated for Γ_ϕ, then for every outcome-independent binning rule B,

$$\mathbf{E}[\mathrm{LinCE}_{\phi,\mathsf{B}}(h(x_{1:n});y_{1:n})] = \frac{m}{n}\mathbf{E}_{(x,y)\sim D}\big[\ell_\phi(h(x),y)\big].$$

2. If $h: X \to \Delta(\mathcal{Y})$ is classwise calibrated, then for every outcome-independent binning rule B,

$$\mathbf{E}[\mathrm{LinCE}^{\mathrm{cw}}_{\mathsf{B}}(h(x_{1:n});y_{1:n})] = \frac{m}{n}\mathbf{E}_{(x,y)\sim D}\big[\ell_{\mathrm{Brier}}(h(x),y)/k\big].$$

3. If $h: X \to \Delta(\mathcal{Y})$ is confidence calibrated, then for every outcome-independent binning rule B,

$$\mathbf{E}[\mathrm{ConfCE}^{\mathrm{corr}}_{\mathsf{B}}(h(x_{1:n});y_{1:n})] = \frac{m}{n}\mathbf{E}_{(x,y)\sim D}\big[\ell_{\mathrm{conf}}(h(x),y)\big].$$

Proof. We prove the three identities in order.

For the linear-property identity, let $U_i := h(x_i)$. Since h is calibrated for Γ_ϕ,

$$\mathbf{E}[\phi(y_i)\,|\,U_i] = U_i \qquad \text{for each } i.$$

Conditioning on $U_{1:n}$ fixes the bins of B, because the binning rule is outcome-independent. Therefore the centered residuals $U_i - \phi(y_i)$ have conditional mean zero and are independent across i, so the cross terms vanish:

$$\mathbf{E}[\mathrm{LinCE}_{\phi,\mathsf{B}}(h(x_{1:n});y_{1:n})\,|\,U_{1:n}] = \frac{m}{n^2}\sum_{i=1}^n \mathbf{E}\big[\|U_i-\phi(y_i)\|_2^2\,|\,U_i\big].$$

Taking expectation and using that the samples are i.i.d. gives

$$\begin{aligned}\mathbf{E}[\mathrm{LinCE}_{\phi,\mathsf{B}}(h(x_{1:n});y_{1:n})] &= \frac{m}{n}\mathbf{E}_{(x,y)\sim D}\big[\|h(x)-\phi(y)\|_2^2\big] \\ &= \frac{m}{n}\mathbf{E}_{(x,y)\sim D}\big[\ell_\phi(h(x),y)\big].\end{aligned}$$

For the classwise identity, fix a class $r \in \{1,\dots,k\}$ and define $U_i^{(r)} := h^{(r)}(x_i)$ and $Z_i^{(r)} := \mathbb{I}[y_i = r]$. Classwise calibration gives

$$\mathbf{E}[Z_i^{(r)}\,|\,U_i^{(r)}] = U_i^{(r)}.$$

Applying the linear-property identity above to the scalar property $\phi_r(y) = \mathbb{I}[y=r]$ and the binning rule B yields

$$\mathbf{E}\Big[\mathrm{LinCE}_{\phi_r,\mathsf{B}}\big(h^{(r)}(x_{1:n});y_{1:n}\big)\Big] = \frac{m}{n}\mathbf{E}_{(x,y)\sim D}\big[(h^{(r)}(x)-\mathbb{I}[y=r])^2\big].$$

Averaging over $r = 1,\dots,k$ and using the definition of $\mathrm{LinCE}^{\mathrm{cw}}_{\mathsf{B}}$ gives

$$\begin{aligned}\mathbf{E}[\mathrm{LinCE}^{\mathrm{cw}}_{\mathsf{B}}(h(x_{1:n});y_{1:n})] &= \frac{m}{kn}\sum_{r=1}^k \mathbf{E}_{(x,y)\sim D}\big[(h^{(r)}(x)-\mathbb{I}[y=r])^2\big] \\ &= \frac{m}{n}\mathbf{E}_{(x,y)\sim D}\big[\ell_{\mathrm{Brier}}(h(x),y)/k\big].\end{aligned}$$

For the confidence identity, let

$$r_i := \operatorname*{argmax}_{r\in\mathcal{Y}} h^{(r)}(x_i), \qquad c_i := h^{(r_i)}(x_i), \qquad z_i := \mathbb{I}[y_i = r_i].$$

Confidence calibration implies

$$\mathbf{E}[z_i\,|\,c_i] = c_i.$$

Conditioning on $c_{1:n}$ fixes the bins of $\mathsf{B}(c_{1:n})$, and the variables $z_1,...,z_n$ remain independent. Hence

$$\mathbf{E}\left[\frac{m}{n^2}\sum_{B\in\mathsf{B}(c_{1:n})}\left(\sum_{i\in B}(c_i-z_i)\right)^2\middle| c_{1:n}\right] = \frac{m}{n^2}\sum_{i=1}^{n} c_i(1-c_i).$$

The correction term contributes

$$\mathbf{E}\left[\frac{m}{n}\left(1-\frac{1}{n}\sum_{i=1}^{n} z_i\right)\middle| c_{1:n}\right] = \frac{m}{n^2}\sum_{i=1}^{n}(1-c_i).$$

Adding the two pieces gives

$$\mathbf{E}[\mathrm{ConfCE}^{\mathrm{corr}}_{\mathsf{B}}(h(x_{1:n});y_{1:n})|c_{1:n}] = \frac{m}{n^2}\sum_{i=1}^{n}(1-c_i^2).$$

On the other hand,

$$\ell_{\mathrm{conf}}(h(x_i),y_i) = 1-z_i+(c_i-z_i)^2,$$

so by confidence calibration,

$$\mathbf{E}\big[\ell_{\mathrm{conf}}(h(x_i),y_i)\,|\,c_i\big] = 1-c_i^2.$$

Taking expectations and using i.i.d. sampling yields

$$\mathbf{E}[\mathrm{ConfCE}^{\mathrm{corr}}_{\mathsf{B}}(h(x_{1:n});y_{1:n})] = \frac{m}{n}\mathbf{E}_{(x,y)\sim D}\big[\ell_{\mathrm{conf}}(h(x),y)\big]. \qquad \square$$

D.4 Proof of the notion-specific dominance guarantees

Proof of Theorem 4.3. For classwise dominance, take the binary proper loss $s_r(c,z)=(c-z)^2$ for every class r. Their average is the classwise loss ℓ_{Brier}/k, so the assumed dominance gives

$$\mathbf{E}[\ell_{\mathrm{Brier}}(f(x),y)/k] \le \mathbf{E}[\ell_{\mathrm{Brier}}(g(x),y)/k].$$

Since both predictors are classwise calibrated, Lemma D.1 identifies these risks with $(n/m)\mathbf{E}[\mathrm{LinCE}^{\mathrm{cw}}_{\mathsf{B}}(f)]$ and $(n/m)\mathbf{E}[\mathrm{LinCE}^{\mathrm{cw}}_{\mathsf{B}}(g)]$, respectively. Multiplication by m/n proves the classwise claim.

For confidence dominance, we first verify that the associated loss ℓ_{conf} is a bounded proper multiclass loss of the required confidence-based form. Fix a true label distribution $q\in\Delta(\mathcal{Y})$ and a reported distribution p. Write $r=r_p$, $c=p^{(r)}$, and $q_*:=\max_j q_j$. Then

$$\begin{aligned}\mathbf{E}_{y\sim q}[\ell_{\mathrm{conf}}(p,y)] &= 1-q_r^2+(c-q_r)^2\\ &\ge 1-q_*^2 = \mathbf{E}_{y\sim q}[\ell_{\mathrm{conf}}(q,y)].\end{aligned}$$

The last equality also holds when the maximum is tied, because the fixed tie-breaking rule selects a maximizer. Moreover, $\ell_{\mathrm{conf}}(p,y)=s(c,\mathbb{I}[y=r_p])$ with $s(c,z)=1-z+(c-z)^2$. Thus the assumed confidence dominance implies

$$\mathbf{E}[\ell_{\mathrm{conf}}(f(x),y)] \le \mathbf{E}[\ell_{\mathrm{conf}}(g(x),y)].$$

Confidence calibration and Lemma D.1 again turn this inequality into the claimed ordering of expected corrected confidence errors. Both arguments apply to every outcome-independent scalar binning rule with the same prescribed n,m in the two evaluations. $\square$

Under full calibration, Blackwell dominance implies both of these notion-specific dominance relations: each loss in either family is a bounded proper multiclass loss, and a fully calibrated report is a Bayes act for its conditional label distribution. A more informative experiment therefore has no larger expected loss. Weak calibration alone does not justify this implication, as the following example shows.

Example D.2 (Weak calibration does not suffice for full-vector Blackwell ordering)**.** *Let $k=3$, let $r+1$ denote the cyclic successor of $r\in\{1,2,3\}$, and let $x=(r,b)$ have probabilities $\Pr[x=(r,0)]=2/9$ and $\Pr[x=(r,1)]=1/9$ for every r. Set $y=r$ deterministically and define*

$$f(r,b)=\begin{cases}\frac{2}{3}e_r+\frac{1}{3}e_{r+1}, & b=0,\\ \frac{1}{3}e_r+\frac{2}{3}e_{r+1}, & b=1,\end{cases}\qquad g(r,b)=e_r.$$

For each coordinate, the report $2/3$ is correct with conditional probability $2/3$, the report $1/3$ with conditional probability $1/3$, and the report 0 is never correct. Hence f is classwise calibrated. Its confidence is always $2/3$, and its top label is correct exactly when $b=0$, an event of probability $2/3$, so it is also confidence calibrated. The predictor g is fully calibrated.

All six reports of f are distinct and identify y, so g is a deterministic function of f and $f\succeq_B g$. Nevertheless, the proper-loss identities give

$$\mathbf{E}[\mathrm{LinCE}^{\mathrm{cw}}_{\mathsf{B}}(f)]=\frac{4m}{27n}>0=\mathbf{E}[\mathrm{LinCE}^{\mathrm{cw}}_{\mathsf{B}}(g)],$$
$$\mathbf{E}[\mathrm{ConfCE}^{\mathrm{corr}}_{\mathsf{B}}(f)]=\frac{5m}{9n}>0=\mathbf{E}[\mathrm{ConfCE}^{\mathrm{corr}}_{\mathsf{B}}(g)].$$

Indeed, $\|f(x)\|_2^2=5/9$ and $c_f(x)^2=4/9$ almost surely. Thus even simultaneous classwise and confidence calibration does not allow full-vector Blackwell dominance to replace the corresponding notion-specific dominance assumption.

D.5 Proof of Theorem 4.4

The benchmark $\bar{h}$ for the classwise and confidence errors is the full posterior $\Pr[y\,|\,h(x)]$, not separate scalar recalibrations. The tower property makes $\bar{h}$ fully calibrated. Applying Theorem 4.2 with $\phi(y)=e_y$ and then Lemma D.1 proves $\rho_{\mathrm{CAL}}(h)\geq 0$. The same argument uses the given ϕ for a linear-property error. This residual is not solely a measure of weak miscalibration: in Example D.2, $\bar{f}=g$ has zero associated loss, but $\rho_{\mathrm{cw}}(f)=4/27$ and $\rho_{\mathrm{conf}}(f)=5/9$ despite both weak calibration conditions.

Proof. Using the decomposition

$$\frac{n}{m}\mathbf{E}[\mathrm{CAL}(h)]=\mathbf{E}\big[\ell_{\mathrm{CAL}}(\bar{h}(x),y)\big]+\rho_{\mathrm{CAL}}(h),$$

we obtain

$$\begin{aligned}\frac{n}{m}\mathbf{E}[\mathrm{CAL}(f)]-\frac{n}{m}\mathbf{E}[\mathrm{CAL}(g)]&=\mathbf{E}[\ell_{\mathrm{CAL}}(\bar{f}(x),y)]-\mathbf{E}[\ell_{\mathrm{CAL}}(\bar{g}(x),y)]+\rho_{\mathrm{CAL}}(f)-\rho_{\mathrm{CAL}}(g)\\&\leq-\epsilon+\epsilon\leq 0,\end{aligned}$$

where the first inequality uses the assumed proper-loss separation together with $\rho_{\mathrm{CAL}}(f)\leq\epsilon$ and $\rho_{\mathrm{CAL}}(g)\geq 0$. Multiplying by m/n proves the theorem. □

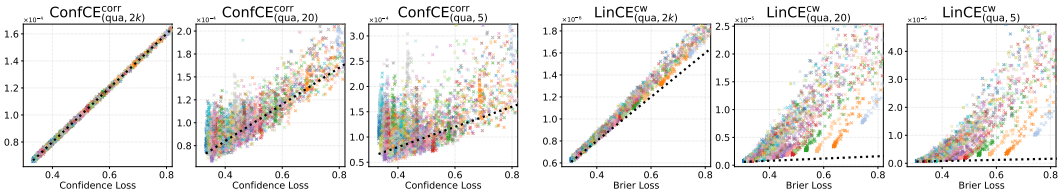
E Additional empirical results

E.1 Assumption validation: predictors are approximately calibrated

This section collects supplementary empirical plots deferred from the main text. In particular, Figure 4 provides the assumption-validation plots showing that, after temperature scaling, the evaluated predictors lie close to the calibrated-predictor loss lines for both confidence and classwise calibration. These diagnostics do not establish full calibration of the reported probability vectors.

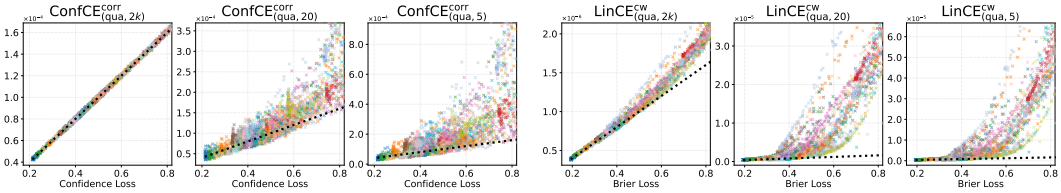
E.2 Dominance preservation and binning robustness: omitted figures

We obtain the maximum dominance set by finding a longest path in the directed acyclic graph whose edges encode the empirical dominance relation in Section 5.



(a) Within-model, confidence calibration.

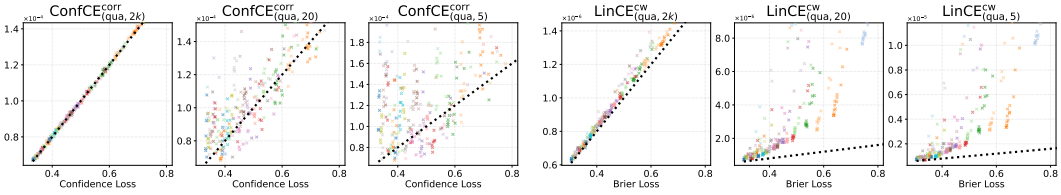
(b) Within-model, classwise calibration.



(c) Cross-model, confidence calibration.

(d) Cross-model, classwise calibration.

Figure 4: Assumption validation plots deferred from the main text. Vertical axes show the constructed errors divided by m, following the plotting convention in Section 5. In each panel, the dotted line is the calibrated-predictor identity $\mathbf{E}[\textsc{Cal}/m] = \mathbf{E}[\ell_{\textsc{Cal}}]/n$; the normalized error CAL instead has expectation $(m/n)\mathbf{E}[\ell_{\textsc{Cal}}]$. After temperature scaling on the validation set, the within-model and cross-model checkpoints stay close to this line, providing an empirical diagnostic for confidence and classwise calibration.



(a) $\mathrm{ConfCE}^{\mathrm{corr}}_{\mathrm{B}}/m$ on the maximum dominance set.

(b) $\mathrm{LinCE}^{\mathrm{cw}}_{\mathrm{B}}/m$ on the maximum dominance set.

Figure 5: Empirical validation of dominance preservation on the within-model maximum dominance set with the same setting as Figure 2. Each vertical axis shows the corresponding constructed error divided by its prescribed bin count m.

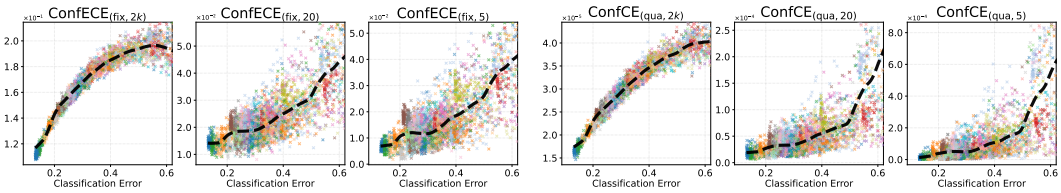
F Experimental details and reproducibility

All paper experiments use `CIFAR-100` distributed through `torchvision`. The code keeps the official 50,000-image training split intact and deterministically partitions the official 10,000-image test split into 5,000 validation and 5,000 test examples using random seed 42. The validation split is used only for post-hoc scaling; all reported figures and tables are evaluated on the test split.

In the experiments, we fine-tune nine ImageNet-pretrained `timm` models: `mobilenetv3_small_050`, `mobilenetv3_small_075`, `mobilenetv3_small_100`, `resnet10t`, `resnet18`, `resnet34`, `resnet50`, `resnet101`, and `resnet152`. For `CIFAR-100`, all runs use resized 224×224 inputs, ImageNet normalization, SGD with momentum 0.9, learning rate 10^{-3}, weight decay 5×10^{-4}, cosine learning rate schedule, batch size 256, 100 epochs, and mixed-precision training. The training transform uses resize, random crop with padding 4, and random horizontal flip; the evaluation transform uses deterministic resize.

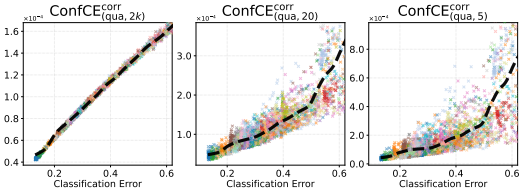
The within-model experiments use 40 traces from `mobilenetv3_small_100`: the full-data run plus subset runs from 2.5% to 97.5% in increments of 2.5%. The cross-model experiments use 45 traces: for each of the nine architectures above, we use the full-data run and subset runs at 20%, 40%, 60%, and 80%.

Every checkpoint is evaluated after temperature scaling. For each epoch, the temperature parameter is fit on validation logits by minimizing validation cross-entropy with L-BFGS and is then applied to test logits. The paper presets use all available epochs except warm-up epochs (filtered by classification error) when generating the figures and tables.

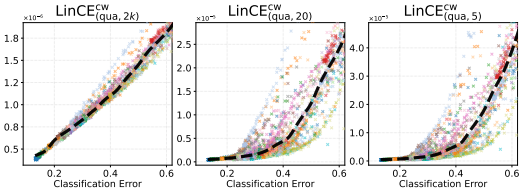


(a) Non-truthful $\mathrm{ConfECE}_{\mathrm{B}}$.

(b) Non-truthful $\mathrm{ConfCE}_{\mathrm{B}}$ ($1/n^2$ scale).



(c) Truthful $\mathrm{ConfCE}_{\mathrm{B}}^{\mathrm{corr}}/m$.



(d) Truthful $\mathrm{LinCE}_{\mathrm{B}}^{\mathrm{cw}}/m$.

Figure 6: Binning robustness on the cross-model traces with the same setting as Figure 3. The truthful squared errors in (c,d) are displayed divided by m, including the entire confidence correction; the baseline in (b) retains its $1/n^2$ scale. These display conventions preserve all reported ranks and rank correlations.

All training runs were executed on the same machine using a single GPU. The machine has dual Intel Xeon Platinum 8470Q CPUs (104 physical cores / 208 threads total), one NVIDIA RTX PRO 6000 Blackwell GPU with about 96 GB device memory, and 1.0 TiB system RAM. The software stack uses PyTorch, `torchvision`, and `timm`.

Table 1: Spearman rank correlations on the test split for the calibration-error comparisons in our binning-robustness analysis. For each calibration-error family, the table reports the correlation between classification error and each calibration error, as well as the pairwise correlations among calibration errors computed with different bin choices. The last two columns compare the within-model MobileNetV3-Small-1.0 setting and the cross-model setting. The reported $\pm$ values denote 1.96 standard deviations for confidence estimation. Truthful calibration errors remain stable across binning choices, whereas non-truthful calibration errors are substantially more sensitive to the choice of bin size. In particular, truthful errors display better rank correlation among errors computed with different binning sizes.

Calibration Error	Item 1	Item 2	Within-Model Spearman	Cross-Model Spearman
ConfECE	Classification Error	$\mathrm{ConfECE}_{(\mathrm{fix},2k)}$	0.972 $\pm$ 0.002	0.974 $\pm$ 0.002
	Classification Error	$\mathrm{ConfECE}_{(\mathrm{fix},20)}$	-0.009 $\pm$ 0.033	0.654 $\pm$ 0.018
	Classification Error	$\mathrm{ConfECE}_{(\mathrm{fix},5)}$	-0.277 $\pm$ 0.030	0.604 $\pm$ 0.020
	$\mathrm{ConfECE}_{(\mathrm{fix},2k)}$	$\mathrm{ConfECE}_{(\mathrm{fix},20)}$	-0.006 $\pm$ 0.033	0.672 $\pm$ 0.018
	$\mathrm{ConfECE}_{(\mathrm{fix},2k)}$	$\mathrm{ConfECE}_{(\mathrm{fix},5)}$	-0.285 $\pm$ 0.030	0.624 $\pm$ 0.020
	$\mathrm{ConfECE}_{(\mathrm{fix},20)}$	$\mathrm{ConfECE}_{(\mathrm{fix},5)}$	0.660 $\pm$ 0.019	0.841 $\pm$ 0.009
ConfCE	Classification Error	$\mathrm{ConfCE}_{(\mathrm{qua},2k)}$	0.966 $\pm$ 0.002	0.976 $\pm$ 0.002
	Classification Error	$\mathrm{ConfCE}_{(\mathrm{qua},20)}$	0.045 $\pm$ 0.033	0.639 $\pm$ 0.019
	Classification Error	$\mathrm{ConfCE}_{(\mathrm{qua},5)}$	-0.121 $\pm$ 0.032	0.582 $\pm$ 0.021
	$\mathrm{ConfCE}_{(\mathrm{qua},2k)}$	$\mathrm{ConfCE}_{(\mathrm{qua},20)}$	0.070 $\pm$ 0.033	0.662 $\pm$ 0.018
	$\mathrm{ConfCE}_{(\mathrm{qua},2k)}$	$\mathrm{ConfCE}_{(\mathrm{qua},5)}$	-0.115 $\pm$ 0.032	0.608 $\pm$ 0.020
	$\mathrm{ConfCE}_{(\mathrm{qua},20)}$	$\mathrm{ConfCE}_{(\mathrm{qua},5)}$	0.652 $\pm$ 0.019	0.828 $\pm$ 0.010
$\mathrm{ConfCE}^{\mathrm{corr}}$	Classification Error	$\mathrm{ConfCE}^{\mathrm{corr}}_{(\mathrm{qua},2k)}$	0.998 $\pm$ 0.000	0.999 $\pm$ 0.000
	Classification Error	$\mathrm{ConfCE}^{\mathrm{corr}}_{(\mathrm{qua},20)}$	0.691 $\pm$ 0.017	0.891 $\pm$ 0.007
	Classification Error	$\mathrm{ConfCE}^{\mathrm{corr}}_{(\mathrm{qua},5)}$	0.272 $\pm$ 0.031	0.753 $\pm$ 0.014
	$\mathrm{ConfCE}^{\mathrm{corr}}_{(\mathrm{qua},2k)}$	$\mathrm{ConfCE}^{\mathrm{corr}}_{(\mathrm{qua},20)}$	0.695 $\pm$ 0.017	0.895 $\pm$ 0.006
	$\mathrm{ConfCE}^{\mathrm{corr}}_{(\mathrm{qua},2k)}$	$\mathrm{ConfCE}^{\mathrm{corr}}_{(\mathrm{qua},5)}$	0.272 $\pm$ 0.031	0.760 $\pm$ 0.014
	$\mathrm{ConfCE}^{\mathrm{corr}}_{(\mathrm{qua},20)}$	$\mathrm{ConfCE}^{\mathrm{corr}}_{(\mathrm{qua},5)}$	0.642 $\pm$ 0.020	0.884 $\pm$ 0.007
$\mathrm{LinCE}^{\mathrm{cw}}$	Classification Error	$\mathrm{LinCE}^{\mathrm{cw}}_{(\mathrm{qua},2k)}$	0.994 $\pm$ 0.000	0.992 $\pm$ 0.000
	Classification Error	$\mathrm{LinCE}^{\mathrm{cw}}_{(\mathrm{qua},20)}$	0.887 $\pm$ 0.007	0.875 $\pm$ 0.008
	Classification Error	$\mathrm{LinCE}^{\mathrm{cw}}_{(\mathrm{qua},5)}$	0.884 $\pm$ 0.007	0.873 $\pm$ 0.008
	$\mathrm{LinCE}^{\mathrm{cw}}_{(\mathrm{qua},2k)}$	$\mathrm{LinCE}^{\mathrm{cw}}_{(\mathrm{qua},20)}$	0.899 $\pm$ 0.006	0.910 $\pm$ 0.006
	$\mathrm{LinCE}^{\mathrm{cw}}_{(\mathrm{qua},2k)}$	$\mathrm{LinCE}^{\mathrm{cw}}_{(\mathrm{qua},5)}$	0.895 $\pm$ 0.007	0.907 $\pm$ 0.006
	$\mathrm{LinCE}^{\mathrm{cw}}_{(\mathrm{qua},20)}$	$\mathrm{LinCE}^{\mathrm{cw}}_{(\mathrm{qua},5)}$	0.997 $\pm$ 0.000	0.997 $\pm$ 0.000

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Introduction state the paper's three main contributions: truthful multiclass calibration errors, Blackwell-dominance/order-preservation results, and empirical robustness to binning choices; these are developed in Sections 3 to 5 and scoped by Appendix A.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Table 2: Spearman rank correlations across binning choices on the selected dominance subsets.

Calibration error	Bin counts	Within-model	Cross-model
ConfECE (fixed)	2k vs. 20	0.151 ± 0.114	0.790 ± 0.041
	2k vs. 5	-0.096 ± 0.115	0.762 ± 0.045
	20 vs. 5	0.690 ± 0.065	0.855 ± 0.029
ConfCE (quantile)	2k vs. 20	0.224 ± 0.112	0.769 ± 0.044
	2k vs. 5	0.107 ± 0.115	0.713 ± 0.053
	20 vs. 5	0.688 ± 0.066	0.850 ± 0.030
$\mathrm{ConfCE}^{\mathrm{corr}}$ (quantile)	2k vs. 20	0.826 ± 0.040	0.953 ± 0.010
	2k vs. 5	0.502 ± 0.091	0.877 ± 0.025
	20 vs. 5	0.740 ± 0.057	0.932 ± 0.014
$\mathrm{LinCE}^{\mathrm{cw}}$ (quantile)	2k vs. 20	0.939 ± 0.015	0.937 ± 0.013
	2k vs. 5	0.938 ± 0.015	0.938 ± 0.013
	20 vs. 5	0.997 ± 0.001	0.996 ± 0.001

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The paper has a dedicated Limitations section, Appendix A, which states that the theory requires nearly calibrated predictors and that the guarantees need not hold in substantially miscalibrated regimes such as post-trained language models.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions appear in the statements of the main theorems in Sections 3 and 4, and complete proofs are provided in Appendices C, D, D.3 and D.5.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental pipeline is described in Section 5, and Appendix F specifies the data split, model families, training hyperparameters, temperature-scaling procedure, evaluation pools, and hardware/software stack used for the reported figures.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper uploads the source code with instructions in supplementary materials.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Section 5 specifies the dataset split, model families, temperature scaling, evaluation pools, binning conventions, and evaluation metrics; Appendix F adds the optimizer, learning-rate schedule, batch size, epochs, transforms, and checkpoint-selection details.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The empirical evidence is presented through plots and rank-correlation summaries in Figures 1 to 3 and appendix E. The tables in Appendix E presents confidence estimates.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix F reports the hardware and software environment (CPU, GPU, system RAM, and software stack).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper is a theoretical and empirical study of calibration measurement using public benchmark data and pretrained image classifiers, as described in Section 5 and appendix F; it does not involve human subjects, crowdsourcing, or release of high-risk generative models.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The manuscript focuses on theory and benchmark evaluation in Sections 3 to 5, which does not imply immediate societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk generative models or scraped datasets; the empirical study uses CIFAR-100 and pretrained image classifiers for evaluation only (Section 5 and appendix F).

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper uses CIFAR-100 via torchvision and ImageNet-pretrained timm classifiers. CIFAR-100 is cited to Krizhevsky [2009]; the official CIFAR-100 page does not appear to specify an explicit license, and the Hugging Face dataset card lists the license as unknown. ImageNet is governed by ImageNet terms of access, which permit non-commercial research and educational use. We do not redistribute either dataset.

    Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper's contributions are theoretical constructions and empirical evaluations; it does not introduce or release a new dataset, model, or benchmark asset in Sections 3 to 5 and appendix F.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The experiments use CIFAR-100 and pretrained image classifiers only, with no crowdsourcing or human-subject data collection; see Section 5 and appendix F.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human subjects; the empirical study is based on CIFAR-100 and pretrained image classifiers as described in Section 5 and appendix F.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method and experiments concern calibration theory and image-classifier evaluation, not LLM-based methodology; see Sections 3 to 5 and appendix A.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.