Calibration without Ground Truth

Yuqing Kong[*,†]
Center on Frontiers of Computing Studies
School of Computer Science
Peking University
Beijing, China
yuqing.kong@pku.edu.cn

Mingyu Song
Center on Frontiers of Computing Studies
School of Computer Science
Peking University
Beijing, China
mingyu.song@pku.edu.cn

Yizhou Wang
Center on Frontiers of Computing Studies
School of Computer Science
State Key Laboratory of
General Artificial Intelligence
Peking University
Beijing, China
yizhou.wang@pku.edu.cn

Yifan Wu
Economics and Computation Group
Microsoft Research, New England
Cambridge, Massachusetts, United States
yifan.wu2357@gmail.com

Abstract

We propose a label-free post-processing framework that improves a strong but miscalibrated primary model using a weaker yet better-calibrated reference. Our key insight is that strict worst-case improvement is possible if and only if the two models are not *mutually calibrated*, meaning there does not exist a joint distribution over their reports and outcomes such that both predictions are simultaneously calibrated. We formalize this condition and connect it to no-arbitrage results from economics. Under this condition, we develop an efficient post-processing algorithm for the strong model's outputs based on Bregman projection, with a strict worst-case performance improvement guarantee. Experiments on three 8B LLM families across three multiple-choice benchmarks show that our method reduces the primary model's ECE by over 40% on common benchmarks.

1 Introduction

Modern neural networks and LLMs achieve high accuracy but are often systematically overconfident [Guo et al., 2017, Ovadia et al., 2019]. In principle, models should be calibrated, which requires reported confidence to align with the conditional likelihood of correctness [Dawid, 1982].

Several methods have been developed to mitigate LLM overconfidence, including supervised fine-tuning [Wang et al., 2025, Xiao et al., 2025] and temperature scaling with human annotations [Guo et al., 2017]. However, these approaches rely heavily on human-labeled data, which is challenging in highly specialized domains in practice. Things are even worse given the fact that temperature scaling performs poorly with limited labels [Mozafari et al., 2018]. Villalobos et al. [2024] predict that LLMs will exhaust the stock of public human text between 2026 and 2032, making calibrating LLMs *without ground truth* more practical and urgent.

In this regime, one promising direction is weak-to-strong generalization, where an existing weak model is used to improve a strong one [Burns et al., 2024]. This weak-to-strong generalization

[*]Authors are listed in alphabetical order.
[†]Corresponding author.

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

suggests that the weak reference can be valuable even when the model is imperfect. A natural instance of this pattern arises in calibration: *fine-tuned* LLMs after post-training or RLHF are highly capable but overconfident, while their *base* pre-trained counterparts are less accurate yet better calibrated [Achiam et al., 2023, Luo et al., 2025, Tan et al., 2026]. This pattern is common in many scenarios, and a well-calibrated weak model is relatively easy to obtain: a general-purpose text model may achieve higher accuracy than a structured baseline while producing less calibrated probabilities [Cruz et al., 2024], and in real-time social-media annotation where human labels are hard to obtain, a calibrated rule-based system such as VADER [Hutto and Gilbert, 2014] can serve as a reference.

This raises a natural question: can a calibrated weaker model be used to improve a stronger one without accessing ground-truth labels, while preserving the high capability of the stronger model? Recent work [Luo et al., 2025] pursues similar goals but does not establish a worst-case proper-loss improvement guarantee, while Tan et al. [2026] requires additional queries to weaker models or extra access to the hidden states of LLMs.

We name the stronger, miscalibrated model the **primary** predictor, and the weaker, better-calibrated model the **reference**. Our goal is to design a **post-processing algorithm** that maps the primary model's outputs to a strictly improved predictor with a provable worst-case improvement guarantee, while requiring no ground-truth labels. At deployment, the learned transformation uses only the primary model's output and does not require querying the reference predictor.

We characterize the key condition when strict improvement is possible as **mutual calibration**: two predictors are mutually calibratable if there exists a joint distribution over both predictions and outcomes under which both predictors are simultaneously calibrated. For full-distribution reports, our central result (Theorem 3.3) shows that a strictly improving transformation exists *if and only if* the two predictors are *not* mutually calibrated. This condition also connects to economics: mutual calibration parallels the common-prior assumption, and its failure implies guaranteed arbitrage opportunities [Milgrom and Stokey, 1982, Arieli et al., 2021, Morris, 2020]. For a chosen differentiable strictly proper loss, our algorithm exploits exactly this contradiction to guarantee a worst-case loss reduction. We provide the formal connection in Appendix E.

The following example illustrates how contradictory predictions create an arbitrage opportunity:

Example 1.1 (Arbitrage in Contradictory Predictions). Consider two predictors forecasting the probability that an investment will double in value.

- The primary predictor outputs 10% for half of the samples (Group A of timestamps) and 70% for the other half (Group B of timestamps).
- The reference predictor always outputs a base rate of 50% and is known to be calibrated.

The primary predictor implies an average base rate of $0.5 \times 70\% + 0.5 \times 10\% = 40\%$, contradicting the reference's 50%, so the two predictors are not mutually calibratable.

Consider an arbitrageur offering the following contracts ($0 < \epsilon < 0.1$).

- **To the reference predictor:** the arbitrageur pays $(0.5 + \epsilon)$ if the investment doubles, and receives 0.5 if not.
- **To the primary predictor:** When predicting 0.7, the arbitrageur receives 0.3 if the investment doubles, and pays $(0.7 + \epsilon)$ otherwise; when predicting 0.1, receives 0.9 if the investment doubles, and pays $(0.1 + \epsilon)$ otherwise.

Both predictors accept these contracts because their subjective expected profits are strictly positive. The arbitrageur's payoff is $0.4 - \epsilon$ in Group A and $-0.2 - \epsilon$ in Group B, independently of the realized outcome. Since the groups have equal probability, the expected payoff is $0.1 - \epsilon > 0$ under every possible outcome, indicating an arbitrage opportunity through the contradictory predictions.

Our algorithm has two steps: (1) define the *reference-compatible set*, which is a compact convex set including all transformations of the primary predictor that can be induced by ground truth extensions under which the reference is calibrated; and (2) project the primary predictor onto this set under the Bregman divergence induced by the chosen proper loss, choosing the transformation that minimally adjusts the primary model while restoring mutual calibration (Figure 1). When two predictors are not mutually calibrated, the projection guarantees a proper-loss reduction by reducing inconsistency.

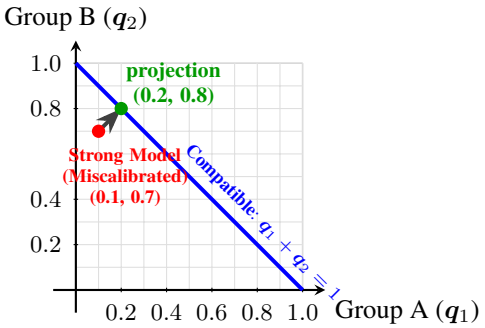

Figure 1: Projection onto the reference-compatible set: transformations induced by ground truth extensions under which the reference is calibrated. The red point is the primary predictor, and the blue line is the constraint imposed by the calibrated reference. Under the chosen Bregman geometry (here we use Brier loss so the projection is Euclidean), the green point is the compatible map closest to the primary predictor. The specific numbers are derived from Example 1.1.

1.1 Our Contributions

Theory We formalize the problem of improving a (approximately) miscalibrated primary model using a calibrated reference and develop an efficient Bregman-projection-based algorithm.

- **Condition for Strict Improvement.** In the full-distribution setting, we characterize when strict improvement is possible: precisely when the primary and reference models are *not* mutually calibrated. We link mutual calibration to the No-Trade Theorem in economics, and establish a correspondence between maximal extractable arbitrage profit and calibration error metrics (Appendix E).
- **Worst-Case Improvement Guarantee.** Our algorithm outputs a transformation that reduces the strong model's expected proper loss by at least its Bregman divergence to the *reference-compatible set* (the "distance" between the red and green points in Figure 1) for every ground-truth extension under which the reference is calibrated (Theorem 3.3). Furthermore, we relax the assumption of a perfectly calibrated reference, quantifying how the guarantee degrades with reference calibration error (Theorem C.1).
- **Generality of the Framework.** Our framework applies broadly to:
  - *Diverse Metrics:* Optimization under proper losses satisfying the stated regularity conditions, including Brier Score, Log Loss, and Confidence Loss.
  - *General Elicitable Properties:* Extensions beyond full predictive distributions to elicitable reports such as means, medians, and top-one classification with confidence.

Experiment We evaluate our method on three mainstream open-source 8B LLM families across three multiple-choice benchmarks. We set the better-calibrated Base models as *reference* predictors to improve their *primary* Instruct counterparts. Experimental results show significant reductions in both the Brier Score and Expected Calibration Error (ECE) relative to the Instruct models. Despite being label-free, our method substantially improves the calibration of the primary model. Unlike standard post-hoc calibration methods that modify only confidence scores, our approach re-estimates the full probability distribution; empirically, it preserves the high prediction accuracy of the Instruct models and sometimes yields marginal accuracy gains (Section 5).

1.2 Related Work

We discuss the most closely related work here; more related work is deferred to Appendix A.

Mutual Calibration and the No-Trade Theorem The notion of *mutual calibration* closely parallels the Common Prior Assumption and rationality in the economic literature. Seminal work by Aumann [1976] shows that Bayesian agents who share a common prior cannot "agree to disagree". This insight was extended to market settings by the No-Trade Theorem [Milgrom and Stokey, 1982],

which states that risk-averse agents with common priors cannot profit from trade in an equilibrium. More recently, Arieli et al. [2021] and Morris [2020] characterize the set of *feasible* joint distributions of posterior beliefs as those that admit no arbitrage, formalizing feasibility through arbitrage-free constraints. This notion of feasibility aligns precisely with our definition of mutual calibration. We leverage this connection to show that failures of mutual calibration correspond to exploitable arbitrage opportunities, which in turn can be used in a learning context to systematically optimize model predictions and improve performance.

Calibration Baselines Classical post-hoc calibration uses a labeled calibration set, including temperature scaling [Guo et al., 2017], histogram binning and isotonic regression [Zadrozny and Elkan, 2001, 2002], and richer probabilistic methods such as GP calibration and PAC-Bayes recalibration [Wenger et al., 2020, Fujisawa and Futami, 2025]. These methods serve as supervised baselines, while our target setting is calibration transfer when labels are absent. Closest to our setting, Luo et al. [2025] and Tan et al. [2026] develop label-free calibration methods that exploit the empirical observation that base LLMs are often better calibrated than their post-trained counterparts: Luo et al. [2025] use agreement with the base model to guide unsupervised temperature scaling, while BaseCal [Tan et al., 2026] either scores post-trained model responses with the base model or learns a hidden-state projection and uses the base model's output layer to estimate confidence. Unlike our approach, these empirically effective methods lack worst-case loss-improvement guarantees.

Weak-to-Strong Generalization and Self-Improvement. A closely related line of work studies weak-to-strong generalization, using supervision from a weak model to improve a stronger model [Burns et al., 2024]. Recent theoretical work analyzes this phenomenon using techniques closely related to ours: Charikar et al. [2024] characterize the gain in weak-to-strong generalization under squared loss, Mulgund and Pabbaraju [2025] extend the analysis beyond squared loss, and Yao et al. [2025a] compare reverse-KL and forward-KL losses. Yao et al. [2025b] relate student calibration to teacher calibration and disagreement, while Xu et al. [2026] characterize weak-to-strong risk gains through teacher–student Bregman misfit. These works share our high-level motivation of using a weaker source of information to improve a stronger model and also employ Bregman geometry. The main difference is that standard weak-to-strong generalization typically uses weak labels or pseudo-labels to train a strong student, whereas both predictors remain frozen in our setting: the weak model serves as a calibrated probabilistic reference, and we post-process the strong model using only the unlabeled joint distribution of their reports. On the self-improvement side of language models, Mohri et al. [2025] uses Bregman projection to improve coherence across answers to self-generated paraphrases, rather than mutual calibration between two predictors.

2 Problem Formulation

Setup. We consider a prediction setting with an outcome variable $Y \in \mathcal{Y}$ and two predictive models. We treat the output of the predictive models as random variables:

- $\boldsymbol{Q}_1$: The **primary predictor** (e.g., a strong, but miscalibrated model).
- $\boldsymbol{Q}_0$: The **reference predictor** (e.g., a weak, but calibrated model).

We first consider the standard case where predictors output full probability distributions over outcomes, $\boldsymbol{Q} \in \Delta(\mathcal{Y})$. We assume the existence of a calibrated reference predictor. In applications, prior model or domain knowledge may support this assumption. When a small labeled diagnostic set is available, it can provide empirical evidence about calibration performance comparisons (see Appendix G.1).

Assumption 2.1 (Calibrated Reference)**.** We assume the reference predictor $\boldsymbol{Q}_0$ is *calibrated*, meaning $\Pr(Y = y \mid \boldsymbol{Q}_0 = \boldsymbol{q}) = \boldsymbol{q}(y)$ for all $y \in \mathcal{Y}$.

Assumption relaxation. Appendix C relaxes perfect reference calibration and derives a lower bound on loss improvement with an explicit calibration-error penalty (Theorem C.1). Under the conditions of that theorem, the bound guarantees strict improvement when the geometric gain exceeds the miscalibration penalty. Empirically, we observe that even when the reference is only better calibrated than the primary, our method can exploit this relative advantage to reduce error.

We assume access to the joint distribution over $\boldsymbol{Q}_0$, $\boldsymbol{Q}_1$, denoted by $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$, estimated by the empirical joint distribution of the two predictors' outputs on an **unlabeled** dataset. We focus on

finite supports $\mathcal{Q}_0, \mathcal{Q}_1$. We say a distribution $\widetilde{\mathcal{D}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1,Y}$ *extends* $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$ when the latter is exactly its $(\boldsymbol{Q}_0, \boldsymbol{Q}_1)$-marginal.

Problem Statement. We aim to learn a transformation map $h : \mathcal{Q}_1 \mapsto \Delta(\mathcal{Y})$ solely from $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$ such that $h(\boldsymbol{Q}_1)$ strictly outperforms $\boldsymbol{Q}_1$ under a chosen proper loss, "transferring" calibration information from $\boldsymbol{Q}_0$ to $\boldsymbol{Q}_1$ without requiring $\boldsymbol{Q}_0$ during subsequent deployment.

Formally, we evaluate performance using a *(strictly) proper loss* $\ell : \Delta(\mathcal{Y}) \times \mathcal{Y} \to \mathbb{R}$, such that the expected loss is (uniquely) minimized when the predicted distribution matches the true posterior. Specifically, for any true distribution $\boldsymbol{p} \in \Delta(\mathcal{Y})$ of the outcome y, it holds that $\mathbb{E}_{y\sim\boldsymbol{p}}\left[\ell(\boldsymbol{p}, y)\right] \leq \mathbb{E}_{y\sim\boldsymbol{p}}\left[\ell(\boldsymbol{q}, y)\right]$ for any forecast $\boldsymbol{q}$. Common examples include the Log-Loss and the Brier Score.

Definition 2.2 (Strictly Improving Transformation)**.** Given $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$, a deterministic transformation $h : \mathcal{Q}_1 \to \Delta(\mathcal{Y})$ is *strictly improving* if:

1. For every distribution $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1,Y}$ extending $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$ under which $\boldsymbol{Q}_0$ is calibrated, applying h weakly improves the loss: $\mathbb{E}[\ell(h(\boldsymbol{Q}_1), Y)] \leq \mathbb{E}[\ell(\boldsymbol{Q}_1, Y)]$.
2. There exists such an extension $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1,Y}$ leading to strict inequality.

We aim to find a strictly improving transformation. This is not guaranteed in all cases. For instance, if both predictors $\boldsymbol{Q}_0$ and $\boldsymbol{Q}_1$ are already providing the ground truth, no transformation will improve the performance. Thus, we will quantify the lower bound of improvement as a function of $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$, and explore the conditions on $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$ that make a strictly improving transformation possible.

In the full-distribution setting, we obtain an even stronger result: whenever a strictly improving transformation exists, a projection h^* achieves a uniformly positive expected-loss reduction over *all* valid ground truth distributions $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1,Y}$.

2.1 Extension to General Elicitable Properties

In many practical applications, models report summary statistics rather than full probability distributions. We generalize our framework to *elicitable properties*: those for which truthful reporting is incentivized by a proper loss [Lambert et al., 2008]. Our central example is the (mode, max-probability) pair used throughout our LLM experiments.

Example 2.3 (Mode and Confidence). Consider a Question-Answering (QA) system where the predictor outputs a discrete answer and a confidence score.

- **Property** Γ**:** The pair $(\text{mode}, \text{max-probability})$:

$$\Gamma(\boldsymbol{q}) = \left\{(y, c) \mid y \in \arg\max_{y'\in\mathcal{Y}} \boldsymbol{q}(y'),\ c = \max_{y'\in\mathcal{Y}} \boldsymbol{q}(y')\right\}.$$

- **Report** r**:** A tuple $(\widehat{y}, \widehat{c})$ with predicted class $\widehat{y} \in \mathcal{Y}$ and confidence $\widehat{c} \in [0, 1]$.
- **Loss** ℓ**:** A strictly proper loss for this property is [Damani et al., 2026]:

$$\ell((\widehat{y}, \widehat{c}), y) = -\mathbb{I}(y = \widehat{y}) + (\widehat{c} - \mathbb{I}(y = \widehat{y}))^2. \tag{1}$$

We call it the *confidence loss*. This loss function penalizes both incorrect classification (accuracy) and poor confidence assessment (calibration).

We say the reference predictor R_0 is calibrated with respect to property Γ if its report matches the property of the conditional distribution given that report:

$$r \in \Gamma(\mathcal{D}_{Y|R_0=r}).$$

for every possible report r.

For property reports, we allow a random transformation that assigns a finitely supported distribution π_r on the allowable report space $\mathcal{A}_1$ to each $r \in \mathcal{R}_1$, where $\mathcal{R}_1$ denotes the finite support of the primary predictor. Conditional on $R_1 = r$, the transformed report $\widehat{R} \sim \pi_r$ is independent of (R_0, Y). Deterministic rules correspond to point masses. We seek a strictly positive worst-case expected-loss reduction over all extensions of the observed joint report distribution $\mathcal{D}_{R_0,R_1}$ under which the

reference R_0 is calibrated with respect to the property. For each such extension, the expectation is over (R_0, R_1, Y) and the output randomization. Our construction of the optimal transformation combines a distribution-valued projection $h : \mathcal{R}_1 \to \Delta(\mathcal{Y})$ with a suitable report-selection rule supported on $\Gamma(h(r))$. See Appendix D.1 for details.

3 Solution and Theoretical Guarantees

This section presents our method for improving a primary predictor given a calibrated reference. We adopt a geometric perspective, viewing transformed predictors as points. Feasible transformed predictors form a compact, convex region. If a primary predictor is not mutually calibrated with the reference predictor, it lies outside this region. Our approach projects such a predictor back onto this reference-compatible set, yielding a worst-case proper-loss improvement guarantee.

3.1 The Condition for Improvement: Mutual Calibration

We first introduce the central concept determining the possibility of strict improvement: **Mutual Calibration**. This concept characterizes the consistency of the joint distribution of two predictors without requiring access to the ground truth outcome.

Definition 3.1 (Mutual Calibration)**.** Given a joint distribution $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$, predictors $\boldsymbol{Q}_0$ and $\boldsymbol{Q}_1$ are **mutually calibrated** if there exists an extension $\widetilde{\mathcal{D}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1,Y}$ with $(\boldsymbol{Q}_0, \boldsymbol{Q}_1)$-marginal $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$ under which both predictors are calibrated with respect to the outcome Y.

Our main theorem (Theorem 3.3) shows that a strictly improving transformation exists if and only if mutual calibration fails.

Economic Interpretation. Mutually calibrated predictors can be interpreted as Bayesian agents whose reports are posteriors under a common prior, where the extension $\widetilde{\mathcal{D}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1,Y}$ serves as the shared prior distribution. Therefore, characterizing whether two predictors are mutually calibrated given joint distribution $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$ is closely related to checking whether these two predictors share a common prior, which is a fundamental and widely-studied question in the theoretical economic literature. The No-Trade Theorem [Milgrom and Stokey, 1982] and related feasibility results [Arieli et al., 2021, Morris, 2020] characterize when the joint distribution $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$ can be derived from a common prior in terms of **no-arbitrage guarantee**. A failure of mutual calibration implies the existence of an arbitrage opportunity, and in Appendix E we formally quantify the relationship between the maximal arbitrage space and the predictors' calibration errors. This can be viewed as an economic interpretation to our mutual calibration concept.

3.2 Algorithm: Projection onto Reference-Compatible Set

When mutual calibration fails, we propose the following algorithm to generate a strictly improving transformation.

Given a joint distribution between the primary and reference predictors, the algorithm has two steps: first, define a **reference-compatible set** of predictors as the set of all feasible transformations. Second, **project** the primary predictor onto this set by finding the closest reference-compatible predictor in terms of Bregman divergence.

3.2.1 Step 1: The Reference-Compatible Set

Let $\mathcal{H} = \{h \mid h : \mathcal{Q}_1 \mapsto \Delta(\mathcal{Y})\}$ denote the space of all functions mapping the primary predictions to a full prediction over outcome space. Given $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$, define

$$\mathcal{H}_{\text{comp}}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}) = \left\{ h \in \mathcal{H} : \begin{array}{l} \exists\, \widetilde{\mathcal{D}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1,Y} \text{ with marginal } \mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1} \text{ such that} \\ \mathbb{E}_{\widetilde{\mathcal{D}}}[\mathbf{Y} \mid \boldsymbol{Q}_0] = \boldsymbol{Q}_0, \quad \mathbb{E}_{\widetilde{\mathcal{D}}}[\mathbf{Y} \mid \boldsymbol{Q}_1] = h(\boldsymbol{Q}_1) \end{array} \right\}. \tag{2}$$

The definition has an explicit linear representation, implying the compactness and convexity of the set. Let $z(y, \boldsymbol{q}_0, \boldsymbol{q}_1) = \Pr_{\widetilde{\mathcal{D}}}(Y = y, \boldsymbol{Q}_0 = \boldsymbol{q}_0, \boldsymbol{Q}_1 = \boldsymbol{q}_1)$, then $h \in \mathcal{H}_{\text{comp}}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1})$ if and only if

there exists a nonnegative array z satisfying

$$\begin{aligned}
\sum_{y} z(y, \boldsymbol{q}_0, \boldsymbol{q}_1) &= \Pr_{\mathcal{D}}(\boldsymbol{Q}_0 = \boldsymbol{q}_0, \boldsymbol{Q}_1 = \boldsymbol{q}_1) \qquad \forall \boldsymbol{q}_0 \forall \boldsymbol{q}_1, \\
\sum_{\boldsymbol{q}_1} z(y, \boldsymbol{q}_0, \boldsymbol{q}_1) &= \boldsymbol{q}_0(y) \sum_{\boldsymbol{q}_1} \Pr_{\mathcal{D}}(\boldsymbol{Q}_0 = \boldsymbol{q}_0, \boldsymbol{Q}_1 = \boldsymbol{q}_1) \qquad \forall \boldsymbol{q}_0 \forall y, \\
\sum_{\boldsymbol{q}_0} z(y, \boldsymbol{q}_0, \boldsymbol{q}_1) &= h(\boldsymbol{q}_1)(y) \sum_{\boldsymbol{q}_0} \Pr_{\mathcal{D}}(\boldsymbol{Q}_0 = \boldsymbol{q}_0, \boldsymbol{Q}_1 = \boldsymbol{q}_1) \qquad \forall \boldsymbol{q}_1 \forall y.
\end{aligned} \tag{3}$$

Observation 3.2. With finite outcome and report supports, $\mathcal{H}_{\mathrm{comp}}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1})$ is nonempty, compact, and convex. Moreover,

$$\mathrm{id} \in \mathcal{H}_{\mathrm{comp}}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}) \quad \Longleftrightarrow \quad \boldsymbol{Q}_0, \boldsymbol{Q}_1 \text{ are mutually calibrated.}$$

Under the loss assumptions stated below, compactness ensures that the projection objective attains its minimum, while convexity permits standard convex optimization methods.

3.2.2 Step 2: Bregman Projection

To define the projection, we require a notion of "distance". Since we aim to reduce proper loss, the natural choice is the Bregman divergence generated by the proper loss ℓ [Savage, 1971]. Formally, the divergence between a true distribution $\boldsymbol{p}$ and a prediction $\boldsymbol{q}$ is the expected regret:

$$D_\ell(\boldsymbol{p}, \boldsymbol{q}) = \mathbb{E}_{\widetilde{Y} \sim \boldsymbol{p}}[\ell(\boldsymbol{q}, \widetilde{Y}) - \ell(\boldsymbol{p}, \widetilde{Y})].$$

Here $\widetilde{Y}$ is an auxiliary draw from $\boldsymbol{p}$, not an observed ground-truth label.

We seek the reference-compatible transformation $h \in \mathcal{H}_{\mathrm{comp}}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1})$ that remains closest to the primary predictor $\boldsymbol{Q}_1$. Crucially, this projection is equivalent to minimizing the expected divergence (or loss gain) relative to the primary predictor:

$$\begin{aligned}
h^* &\in \operatorname*{argmin}_{h \in \mathcal{H}_{\mathrm{comp}}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1})} \mathbb{E}_{\mathcal{D}_{\boldsymbol{Q}_1}}[D_\ell(h(\boldsymbol{Q}_1), \boldsymbol{Q}_1)] \\
&= \operatorname*{argmin}_{h \in \mathcal{H}_{\mathrm{comp}}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1})} \mathbb{E}_{\mathcal{D}_{\boldsymbol{Q}_1}}\left[\mathbb{E}_{\widetilde{Y} \sim h(\boldsymbol{Q}_1)}\left[\ell(\boldsymbol{Q}_1, \widetilde{Y}) - \ell(h(\boldsymbol{Q}_1), \widetilde{Y})\right]\right].
\end{aligned} \tag{4}$$

Note that this formulation differs from a standard Bregman projection on the probability simplex because we operate on the space of transformations (functions). We show that this objective induces a valid Bregman divergence on the finite-dimensional space of maps without accessing ground truth labels. Compactness guarantees that a minimizer exists.

3.3 Theoretical Guarantees

In this section, we show that the projection-based transformation in Equation (4) improves the primary predictor under *every* ground-truth extension consistent with the observed reports and a calibrated reference. Specifically, the guaranteed reduction is at least the Bregman distance from the identity to the reference-compatible Set, and this lower bound is attained by a worst-case extension.

We impose mild conditions on proper loss. Every proper loss is associated with a convex generator $G(\boldsymbol{q}) = -\mathbb{E}_{Y \sim \boldsymbol{q}}[\ell(\boldsymbol{q}, Y)]$. We assume finite supports and a strictly proper loss whose strictly convex generator is differentiable with finite gradients on a convex prediction domain containing the identity and the reference-compatible set. Brier loss satisfies these conditions directly; log loss is covered when the relevant forecasts remain in the relative interior of the simplex.

Theorem 3.3. *Given a joint distribution over primary and reference predictors $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$. Let h^* be defined in Equation* (4)*. For every extension $\widetilde{\mathcal{D}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1,Y}$ with $(\boldsymbol{Q}_0, \boldsymbol{Q}_1)$-marginal $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$ under which the reference predictor is calibrated,*

$$\begin{aligned}
&\mathbb{E}_{\widetilde{\mathcal{D}}}[\ell(\boldsymbol{Q}_1, Y)] - \mathbb{E}_{\widetilde{\mathcal{D}}}[\ell(h^*(\boldsymbol{Q}_1), Y)] \geq \delta, \\
&\delta := \min_{h \in \mathcal{H}_{\mathrm{comp}}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1})} \mathbb{E}_{\mathcal{D}_{\boldsymbol{Q}_1}}[D_\ell(h(\boldsymbol{Q}_1), \boldsymbol{Q}_1)].
\end{aligned} \tag{5}$$

Table 1: **Main results on MMLU-Redux and CommonSenseQA. Bold** and underline denote the **best** and second-best results among *label-free* methods, while †indicates the *global* best across all methods. Accuracy for Supervised Baselines is marked with – as these post-hoc methods do not alter final predictions. Detailed results with more baselines and error bars can be found in Appendix G.2.

Methods	MMLU-Redux				CommonSenseQA			
	ACC (↑)	BS (↓)	ECE (↓)	CL (↓)	ACC (↑)	BS (↓)	ECE (↓)	CL (↓)
Qwen3-8B								
Base Model ($\boldsymbol{Q}_0$)	77.19%	0.1227	**0.0305**†	−0.6493	**83.21%**†	**0.1167**†	**0.0611**	**−0.7154**†
Instruct Model ($\boldsymbol{Q}_1$)	82.53%	0.1362	0.1269	−0.6891	82.80%	0.1447	0.1355	−0.6833
TS *(Supervised)*	–	0.1137†	0.0426	−0.7116†	–	0.1178	0.0266†	−0.7102
DACA *(Label-Free)*	–	**0.1225**	0.0985	**−0.7028**	–	0.1226	0.0631	−0.7054
Ours *(Label-Free)*	**82.56%**†	0.1236	0.0348	−0.7020	82.80%	0.1309	0.0638	−0.6971
LLaMA-3.1-8B								
Base Model ($\boldsymbol{Q}_0$)	65.35%	**0.1574**†	**0.0208**†	−0.4961	71.33%	0.1708	**0.0363**†	−0.5425
Instruct Model ($\boldsymbol{Q}_1$)	**70.91%**†	0.2550	0.2526	−0.4542	**78.30%**†	0.2099	0.2097	−0.5731
TS *(Supervised)*	–	0.1873	0.0364	−0.5218†	–	0.1671†	0.0471	−0.6158†
DACA *(Label-Free)*	–	0.1878	0.0453	**−0.5213**	–	**0.1703**	0.0740	**−0.6126**
Ours *(Label-Free)*	**70.91%**†	0.2115	0.1340	−0.4976	**78.30%**†	0.1776	0.1054	−0.6053
Ministral-3-8B								
Base Model ($\boldsymbol{Q}_0$)	75.72%	**0.1281**†	0.0534	−0.6291	69.94%	0.1948	0.1349	−0.5046
Instruct Model ($\boldsymbol{Q}_1$)	80.30%	0.1726	0.1656	−0.6304	75.35%	0.2050	0.1691	−0.5485
TS *(Supervised)*	–	0.1476	0.0445	−0.6553	–	0.1930	0.1111	−0.5605
DACA *(Label-Free)*	–	0.1534	0.0914	−0.6495	–	0.2098	0.1674	−0.5437
Ours *(Label-Free)*	**80.48%**†	0.1475	**0.0218**†	**−0.6573**†	**75.41%**†	**0.1779**†	**0.0571**†	**−0.5762**†

The value δ is strictly positive if and only if $\boldsymbol{Q}_0$ and $\boldsymbol{Q}_1$ are not mutually calibrated. Consequently, a strictly improving deterministic transformation exists if and only if mutual calibration fails. Moreover, some admissible extension attains equality in Equation (5)*, so δ is the exact worst-case gain.*

The proof is short and directly follows from the generalized Pythagorean Theorem for Bregman Divergence defined on the functional space of transformations. A related projection argument appears in Mohri et al. [2025] for self-coherence constraints. We provide a complete proof in Appendix B for reference. The total expected proper loss is lower bounded by the sum of two parts: the "distance" between the current prediction and the reference-compatible set and the expected proper loss of the projected predictor. The projected predictor will reduce the loss by at least this "distance", and thus the improvement is lower bounded by this "distance".

Assumption relaxation We relax the assumption that the reference is perfectly calibrated in Theorem C.1. Our analysis shows that the loss improvement achieved by h^* admits a three-term decomposition into a geometric gain, a nonnegative compatibility bonus, and a signed miscalibration residual. Dropping the compatibility bonus and controlling the residual yields

$$\mathbb{E}[\ell(\boldsymbol{Q}_1, Y)] - \mathbb{E}[\ell(h^*(\boldsymbol{Q}_1), Y)] \geq A - B\sqrt{\epsilon_{\mathrm{ref}}},$$

where $\epsilon_{\mathrm{ref}} = \mathbb{E}\,\|\,\mathbb{E}[\mathbf{Y} \mid \boldsymbol{Q}_0] - \boldsymbol{Q}_0\|_1$ characterizes the calibration error of the reference predictor, and $A = D_\ell(h^*, \mathrm{id})$ and B is the root-mean-square gradient displacement. Thus, strict improvement is preserved whenever the geometric gain dominates the miscalibration penalty, i.e. $A > B\sqrt{\epsilon_{\mathrm{ref}}}$.

3.4 Extension to General Elicitable Properties

The framework extends to elicitable reports such as means, medians, and (answer, confidence) pairs. Under the regularity conditions in Appendix D, an optimal distribution-valued projection $h^* : \mathcal{R}_1 \to \Delta(\mathcal{Y})$ together with a suitable, possibly randomized report-selection rule supported on $\Gamma(h^*(r))$ attains the exact maximum worst-case expected-loss improvement δ_Γ under the original property loss. This value is positive if and only if mutual calibration fails. When the ambient Bayes risk is differentiable at every projected distribution, any deterministic Bayes-optimal report selection attains the same guarantee. Appendix D gives the full construction and proof.

4 Experimental Setup

Datasets. We evaluate our framework on three multiple-choice question answering benchmarks: **MMLU-Redux** [Gema et al., 2025], **CommonsenseQA** [Talmor et al., 2019], **MMLU** [Hendrycks et al., 2021]. For each benchmark, we use two disjoint sets: a calibration set for learning the recalibration mapping and an evaluation set for assessing performance. Dataset-specific splits are detailed in Appendix F.1.

Models. We consider the following open-source LLMs: `QWen3-8B` [Yang et al., 2025], `LLaMA-3.1-8B` [Grattafiori et al., 2024], and `Ministral-3-8B` [Liu et al., 2026]. Consistent with recent findings [Achiam et al., 2023, Leng et al., 2025], we observe an alignment trade-off:

- **Primary Predictor (Q_1):** We select the **Instruct** checkpoints, which offer superior accuracy but suffer from miscalibration.
- **Reference Predictor (Q_0):** We select the corresponding **Base** checkpoints, which retain the better-calibrated distributions of the pre-training corpus.

Our preliminary analysis (Table 1) confirms that **Base** models consistently yield lower ECE than **Instruct** models in our main settings, motivating the choice of the reference predictor in Section 2. When prior knowledge about model families is unavailable, Appendix G.1 further demonstrates that this calibration gap can be checked using only a small labeled diagnostic set, providing a practical criterion for selecting the reference model. Note here the small labeled diagnostic set is optional and only used for comparing calibration performance. These labels are *not* used to fit the map.

Implementation Details. While our framework is compatible with any proper loss under the stated regularity conditions, LLM applications typically involve eliciting an (answer, confidence) pair. Therefore, we instantiate our optimization objective using the **Confidence Loss (CL)** defined in Equation (1), which is designed to truthfully elicit such pairs. The reports are obtained from normalized answer-token log-probabilities of the open-source LLMs. We approximate $\mathcal{D}_{R_0,R_1}$ by the empirical distribution over discretized paired reports on the calibration set, and restrict the transformation h to the observed support of the primary reports. For Confidence Loss, the finite-support projection yields a second-order cone formulation with linear reference-compatibility constraints. Given LLM inference results, we solve this lightweight problem with a standard convex solver.

At deployment, we first discretize the primary report's confidence using the same grid as in calibration. Then we apply the learned mapping to observed discretized reports and retain unseen discretized reports unchanged, without querying the reference model. Refer to Appendix F.2 for details and Appendix G.4 for the discretization analysis.

Baselines. Unlike fine-tuning approaches which are computationally intensive, our framework operates entirely as a *lightweight post-processing* over frozen LLMs. We compare against both supervised and label-free post-hoc calibration methods. Table 1 compares our method with supervised **Temperature Scaling (TS)** [Guo et al., 2017] and label-free **Disagreement-Aware Confidence Alignment (DACA)** [Luo et al., 2025], which fits the temperature by minimizing the KL divergence between the reference and the primary model's outputs on data where both models agree on the final answer. Further details and comparisons appear in Appendix F.3 and Appendix G.2.

Evaluation Metrics. We report evaluations from three perspectives: 1) **Prediction Accuracy (Acc)** for LLM capabilities 2) **Expected Calibration Error (ECE)** [Guo et al., 2017] with $M = 10$ bins and **Brier Score (BS)** [BRIER, 1950] as standard calibration metrics 3) the instantiated **Confidence Loss (CL)** in Eq. 1 as the optimization objective. To verify robustness to the choice of calibration-error estimator, we additionally evaluate alternative binned ECE estimators and automatic-bandwidth SmoothECE [Błasiok and Nakkiran, 2024] (Appendices F.4 and G.5).

5 Results

Table 1 summarizes our main experimental results on **MMLU-Redux** and **CommonSenseQA**. Due to space constraints, we defer the results on **MMLU** to Appendix G.3.

Significant Reduction in Calibration Error. Our method achieves substantial reductions in calibration error compared to Instruct models across all benchmarks, with improvements exceeding

8% in Brier Score (BS) and 40% in Expected Calibration Error (ECE). Compared to the supervised baseline TS and the label-free baseline DACA, our method achieves lower ECE for `Qwen3-8B` on MMLU-Redux and for `Ministral-3-8B` on both main benchmarks. On CommonsenseQA with `Qwen3-8B`, TS has lower ECE, while DACA and our method obtain similar values (0.0631 and 0.0638, respectively). Another exception occurs for `LLaMA-3.1-8B`, which we analyze below. These gains correlate directly with the optimization of Confidence Loss (CL), confirming that our framework effectively minimizes the proper loss to balance prediction correctness and confidence reliability.

Preservation/Marginal Gains of Accuracy. Unlike standard post-hoc calibration methods that only modify confidence and thus fix accuracy (marked with – in Table 1), our method re-estimates the full probability vector, allowing for corrections in the final answer. Results indicate that our approach effectively preserves the high predictive accuracy of the Instruct models. Furthermore, on MMLU-Redux with `Ministral-3-8B`, we observe an accuracy improvement from 80.30% to 80.48%, suggesting that leveraging the Base model may correct instances where the Instruct model is confidently incorrect.

Robustness. Additional results in Appendix G.2 compare our method with other supervised and label-free baselines, showing similar trends. Specifically, we create a *dummy reference* by randomizing the option order and always predicting a uniform distribution. This *dummy reference* yields much smaller reductions in ECE, confirming that our framework indeed leverages meaningful information from the reference model. Appendices G.4 and G.5 examine different discretization steps and calibration-error estimators. Across the alternative binned ECE variants and non-binned SmoothECE, our method consistently improves over the Instruct model.

Failure Case Analysis. We call `LLaMA-3.1-8B` a "failure case" only in the relative sense that TS and DACA obtain lower ECE. Our method still reduces ECE from 0.2526 to 0.1340 on MMLU-Redux and from 0.2097 to 0.1054 on CommonsenseQA, while also reducing Confidence Loss relative to the Instruct model. In this case, the Base model places unusually large mass in low-confidence regions, indicating lower-quality reference information and limiting calibration transfer from Base model to Instruct model.

6 Discussion

Limitations. Our results establish a label-free method for improving a primary model when a (approximately) calibrated reference is available. Consequently, the method relies on prior model or domain knowledge that supports the reference-quality assumption. When such knowledge is unavailable, a small labeled diagnostic set can be used to guide reference selection, but these labels do not enter the recalibration-map fitting process.

Another limitation comes from the number of constraints and variables when solving the recalibration map. Although the finite-support multiple-choice problems in our experiments reduce to lightweight convex programs after LLM inference, their size grows with the number of possible LLM reports. Therefore, directly applying our algorithm to open-ended scenarios may be computationally infeasible.

In practice, we recommend applying our algorithm only when reference selection is supported by prior knowledge or can be guided empirically. Further, implementing our algorithm involves discretizing input confidence scores. This discretization plays the usual bias–variance role in nonparametric estimation: coarser grids give more stable estimates, while finer grids retain more information about the reports. Our experiments show that the method is robust across a range of discretization choices, with $s = 0.01$ providing a strong default in our main settings. More generally, one can choose the finest grid for which the empirical joint distribution remains sufficiently populated.

Future Work. The framework points to several extensions. First, the projection view naturally suggests online and adaptive variants, where the reference-compatible set is updated as new unlabeled examples arrive. Second, the same mutual-calibration principle can be applied beyond multiple-choice QA to richer elicitable reports, including open-ended generation with semantic confidence, human-AI decision support, and subjective annotation tasks. Third, because the method transfers a property rather than a label, it may be useful for aligning other distributional properties, such as uncertainty, abstention behavior, or fairness-related summaries, whenever a trusted reference for that property is available.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 62372007 and 6247070125.

Competing Interests. The authors declare no competing interests.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The twelfth international conference on learning representations*, 2024.

Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *International conference on machine learning*, pages 222–232. PMLR, 2020.

Yasemin Altun and Alex Smola. Unifying divergence minimization and statistical inference via convex duality. In *Proceedings of the 19th Annual Conference on Learning Theory (COLT 2006)*, pages 139–153, 2006.

Itai Arieli, Yakov Babichenko, Fedor Sandomirskiy, and Omer Tamuz. Feasible joint posterior beliefs. *Journal of Political Economy*, 129(9):2546–2594, 2021.

Robert J Aumann. Agreeing to disagree. *The Annals of Statistics*, pages 1236–1239, 1976.

A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and Dharmendra S. Modha. A generalized maximum entropy approach to Bregman co-clustering and matrix approximation. *Journal of Machine Learning Research*, 8:1919–1986, 2007.

Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

Jarosław Błasiok and Preetum Nakkiran. Smooth ECE: Principled reliability diagrams via kernel smoothing. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=XwiA1nDahv`.

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

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: eliciting strong capabilities with weak supervision. In *Proceedings of the 41st International Conference on Machine Learning*, pages 4971–5012, 2024.

Moses Charikar, Chirag Pabbaraju, and Kirankumar Shiragur. Quantifying the gain in weak-to-strong generalization. *Advances in neural information processing systems*, 37:126474–126499, 2024.

Michael Collins, Robert E. Schapire, and Yoram Singer. Logistic regression, Adaboost and Bregman distances. *Machine Learning*, 48(1-3):253–285, 2002.

Corinna Cortes, Vitaly Kuznetsov, Mehryar Mohri, and Umar Syed. Structural maxent models. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 391–399, Lille, France, 07–09 Jul 2015. PMLR.

André F Cruz, Moritz Hardt, and Celestine Mendler-Dünner. Evaluating language models as risk scores. *Advances in Neural Information Processing Systems*, 37:97378–97407, 2024.

Hamid Dadkhahi, Firas Trabelsi, Parker Riley, Juraj Juraska, and Mehdi Mirzazadeh. Distribution-calibrated inference time compute for thinking LLM-as-a-judge, 2025. URL https://arxiv.org/abs/2512.03019. arXiv:2512.03019.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond binary rewards: Training LMs to reason about their uncertainty. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ASQ649zdHm.

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.

Miroslav Dudík, Steven J. Phillips, and Robert E. Schapire. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *Journal of Machine Learning Research*, 8:1217–1260, 2007.

Béla A Frigyik, Santosh Srivastava, and Maya R Gupta. Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Transactions on Information Theory*, 54(11):5130–5139, 2008.

Rafael Frongillo and Ian Kash. General truthfulness characterizations via convex analysis. In *Web and Internet Economics: 10th International Conference, WINE 2014, Beijing, China, December 14-17, 2014. Proceedings 10*, pages 354–370. Springer, 2014.

Masahiro Fujisawa and Futoshi Futami. PAC-bayes analysis for recalibration in classification. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 17986–18023, 2025.

Futoshi Futami and Masahiro Fujisawa. Information-theoretic generalization analysis for expected calibration error. In *Advances in Neural Information Processing Systems*, volume 37, pages 84246–84297, 2024.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, et al. Are we done with mmlu? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5069–5096, 2025.

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

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

P. Harremoës. Maximum entropy and sufficiency. *arXiv preprint arXiv:1607.02259*, 2016.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014.

Jyrki Kivinen and Manfred K. Warmuth. Boosting as entropy projection. In *Proceedings of the 12th Annual Conference on Computational Learning Theory (COLT '99)*, pages 134–144, Santa Cruz, CA, USA, 1999. URL https://mwarmuth.bitbucket.io/pubs/C51.pdf. PDF available from authors' webpages.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32, 2019.

Nicolas S Lambert, David M Pennock, and Yoav Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pages 129–138, 2008.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in LLMs: Reward calibration in RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=l0tg0jzsdL.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026.

Beier Luo, Shuoyuan Wang, Sharon Li, and Hongxin Wei. Your pre-trained LLM is secretly an unsupervised confidence calibrator. In *Advances in Neural Information Processing Systems*, volume 38, pages 163514–163543. Curran Associates, Inc., 2025. doi: 10.52202/085713-5458. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/ef6b667990416111e3ebbeeba4a8f283-Abstract-Conference.html.

Paul Milgrom and Nancy Stokey. Information, trade and common knowledge. *Journal of economic theory*, 26(1):17–27, 1982.

Mehryar Mohri, Jon Schneider, and Yifan Wu. Coherence mechanisms for provable self-improvement. *arXiv preprint arXiv:2511.08440*, 2025.

Stephen Morris. No trade and feasible joint posterior beliefs, 2020. URL https://economics.mit.edu/sites/default/files/inline-files/no%20trade%206.pdf. Working paper.

Azadeh Sadat Mozafari, Hugo Siqueira Gomes, Wilson Leão, Steeven Janny, and Christian Gagné. Attended temperature scaling: a practical approach for calibrating deep neural networks. *arXiv preprint arXiv:1810.11586*, 2018.

Abhijeet Mulgund and Chirag Pabbaraju. Relating misfit to gain in weak-to-strong generalization beyond the squared loss. In *International Conference on Machine Learning*, pages 45129–45150. PMLR, 2025.

R. Nock, A. K. Menon, and C. S. Ong. A scaled Bregman theorem with applications. *arXiv preprint arXiv:1607.00360*, 2016.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Sangdon Park, Osbert Bastani, James Weimer, and Insup Lee. Calibrated prediction with covariate shift via unsupervised domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, pages 3219–3229. PMLR, 2020.

John C. Platt. Probabilities for SV machines. In *Advances in Large-Margin Classifiers*, pages 61–74. The MIT Press, 2000. doi: 10.7551/mitpress/1113.003.0008. URL https://doi.org/10.7551/mitpress/1113.003.0008.

Teodora Popordanoska, Gorjan Radevski, Tinne Tuytelaars, and Matthew B Blaschko. Estimating calibration error under label shift without labels. *arXiv preprint arXiv:2312.08586*, 2023.

R Tyrrell Rockafellar. *Convex Analysis*, volume 28. Princeton University Press, 1997.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. URL https://arxiv.org/abs/1910.01108.

Leonard J Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, 2019.

Hexiang Tan, Wanli Yang, Junwei Zhang, Xin Chen, Rui Tang, Du Su, Jingang Wang, Yuanzhuo Wang, Fei Sun, and Xueqi Cheng. BaseCal: Unsupervised confidence calibration via base model signals. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5172–5187. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.234. URL https://aclanthology.org/2026.acl-long.234/.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Position: Will we run out of data? limits of llm scaling based on human-generated data. In *Forty-first International Conference on Machine Learning*, 2024.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.441. URL https://aclanthology.org/2024.findings-acl.441/.

Ziming Wang, Zeyu Shi, Haoyi Zhou, Shiqi Gao, Qingyun Sun, and Jianxin Li. Towards objective fine-tuning: How LLMs' prior knowledge causes potential poor calibration? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14830–14853, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.722. URL https://aclanthology.org/2025.acl-long.722/.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Jonathan Wenger, Hedvig Kjellström, and Rudolph Triebel. Non-parametric calibration for classification. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 178–190, 2020.

Jiancong Xiao, Bojian Hou, Zhanliang Wang, Ruochen Jin, Qi Long, Weijie J Su, and Li Shen. Restoring calibration for aligned large language models: A calibration-aware fine-tuning approach. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 68364–68390. PMLR, 2025. URL https://proceedings.mlr.press/v267/xiao25b.html.

Gengze Xu, Wei Yao, Ziqiao Wang, and Yong Liu. Weak-to-strong generalization via bregman bias–variance decomposition. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2505.24313.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Wei Yao, Wenkai Yang, Ziqiao Wang, Yankai Lin, and Yong Liu. Revisiting weak-to-strong generalization in theory and practice: Reverse kl vs. forward kl. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2860–2888, 2025a.

Wei Yao, Wenkai Yang, Gengze Xu, Ziqiao Wang, Yankai Lin, and Yong Liu. The capabilities and limitations of weak-to-strong generalization: Generalization and calibration. *arXiv preprint arXiv:2502.01458*, 2025b. URL https://arxiv.org/abs/2502.01458.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the 18th International Conference on Machine Learning*, pages 609–616, 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699, 2002.

A More Related Work

Bregman Projection Bregman projection has been widely studied in a number of learning theory problems, e.g, maximum entropy models [Collins et al., 2002, Dudík et al., 2007, Altun and Smola, 2006, Banerjee et al., 2007, Frongillo and Kash, 2014, Harremoës, 2016, Nock et al., 2016, Cortes et al., 2015], boosting [Collins et al., 2002, Kivinen and Warmuth, 1999], clustering [Banerjee et al., 2005], Bayesian estimation [Frigyik et al., 2008].

Model Distillation Model distillation trains a model with the output of a reference model and aims at better performance than training on hard ground-truth labels alone (see, e.g., Hinton et al. 2015, Sanh et al. 2019, Gu et al. 2024, Agarwal et al. 2024). Our work is similar to model distillation in the sense that we "distill" the calibration of the weak model. Distillation is most commonly used with a reference model at least as good as the trainable model, while our work is closer to the weak-to-strong generalization.

Calibration under Distribution Shift The classical literature of calibration focuses on post-processing model outputs with ground truth [Platt, 2000, Zadrozny and Elkan, 2002, Kull et al., 2019]. The closest work to ours improves model calibration under distribution shifts. The goal of this literature is similar to our work, which is to measure or to improve calibration when there are no ground-truth labels available for the shifted distribution. This literature usually makes assumptions of the distribution shift. The proposed solution relies on the assumptions. For example, Alexandari et al. [2020], Popordanoska et al. [2023] assume that only the marginal distribution of the true outcomes is shifted while the conditional distribution of the features remains the same. Park et al. [2020] assumes that the shifted distribution is sufficiently close to the training data. Our work is orthogonal to the work on distribution shift. We use a weak calibrated model as a reference and correct contradictions in predicted probabilities.

Calibration of Language Models A growing literature studies the calibration of confidence from language models. Achiam et al. [2023] observe that the next-token probability of a language model can be miscalibrated, especially after post-training. Some existing approaches study or improve semantic-level confidence calibration by prompting the model to output a confidence score, e.g., Lin et al. [2022], Tian et al. [2023], Leng et al. [2025], Damani et al. [2026]. While this line of work considers classical calibration with ground truth labels, our work does not assume access to labeled data for fitting the map. Instead, we leverage the empirical observation that base pretrained checkpoints are often better calibrated than their instruct counterparts [Achiam et al., 2023, Luo

et al., 2025, Tan et al., 2026], and we transfer this calibration as a distributional property from a weak reference to a stronger primary model via label-free post-processing.

Calibration with Inference-Time Compute Dadkhahi et al. [2025] study distribution-calibrated inference-time compute for thinking LLM-as-a-judge systems. Their setting calibrates aggregation over repeated judge samples using labeled or human-consensus evaluation targets. Our setting is different: we learn a post-processing map from the joint predictions of a primary and a reference model without target labels, and the learned map does not require repeated reference queries at deployment.

B Proofs of the Main Results

In this section, we provide the proofs for the main result. We first introduce the necessary preliminaries linking strictly proper losses to Bregman divergences.

B.1 More Preliminaries: Proper Losses and Bregman Divergence

Lemma B.1 (Convex representation of proper losses; Savage 1971, Rockafellar 1997)**.** *For a finite outcome space, a loss ℓ is proper if and only if there is a convex generator $G : \Delta(\mathcal{Y}) \to \mathbb{R}$ and a (possibly subgradient) mapping $\psi(\boldsymbol{q}) \in \partial G(\boldsymbol{q})$ such that, for all forecasts $\boldsymbol{q}$ and outcomes y,*

$$\ell(\boldsymbol{q}, y) = -\big(G(\boldsymbol{q}) + \langle \psi(\boldsymbol{q}), \mathbf{e}_y - \boldsymbol{q}\rangle\big).$$

where $\mathbf{e}_y$ is the one-hot vector corresponding to outcome y. If ℓ is strictly proper, G is strictly convex on the relevant affine hull.

Remark B.2 (Regret as Bregman divergence)*.* The expected regret of forecasting $\boldsymbol{q}$ when the true distribution is $\boldsymbol{p}$ is

$$\mathbb{E}_{Y\sim\boldsymbol{p}}[\ell(\boldsymbol{q}, Y) - \ell(\boldsymbol{p}, Y)] = G(\boldsymbol{p}) - G(\boldsymbol{q}) - \langle \psi(\boldsymbol{q}), \boldsymbol{p} - \boldsymbol{q}\rangle =: D_\ell(\boldsymbol{p}, \boldsymbol{q}). \tag{6}$$

Throughout the differentiable results, the displayed gradients and divergences are assumed finite on the prediction domain.

As in Mohri et al. [2025], we extend this divergence to the space of transformations $\mathcal{H}$, by taking expectations over the marginal distribution of the primary predictor's reports.

Observation B.3*.* On the positive-probability support of the primary predictor, a Bregman divergence on $\Delta(\mathcal{Y})$ induces a Bregman divergence on the space of transformations $\mathcal{H}$:

$$D_\ell(f, g) = \mathbb{E}_{\mathcal{D}_{\boldsymbol{Q}_1}}\left[D_\ell(f(\boldsymbol{Q}_1), g(\boldsymbol{Q}_1))\right].$$

If G is strictly convex and differentiable with finite gradients on the prediction domain, this is the Bregman divergence generated by a strictly convex and differentiable generator function $\mathcal{G}(f) = \sum_{\boldsymbol{q}_1} \Pr_{\mathcal{D}}(\boldsymbol{Q}_1 = \boldsymbol{q}_1) G(f(\boldsymbol{q}_1))$.

Proof. Expanding Equation (6) inside the expectation gives

$$D_\ell(f, g) = \mathbb{E}_{\mathcal{D}_{\boldsymbol{Q}_1}}\left[G(f(\boldsymbol{Q}_1)) - G(g(\boldsymbol{Q}_1)) - \langle \psi(g(\boldsymbol{Q}_1)), f(\boldsymbol{Q}_1) - g(\boldsymbol{Q}_1)\rangle\right],$$

which is exactly the Bregman divergence of $\mathcal{G}$. Positive support weights preserve strict convexity. □

Lemma B.4 (Generalized Pythagorean inequality Rockafellar 1997)**.** *Let Bregman divergence D_ℓ be induced by a strictly convex, differentiable generator G. Let $\mathcal{C}$ be a nonempty closed convex set, and suppose the Bregman projection of a point v_0 onto $\mathcal{C}$*

$$v^* \in \arg\min_{v\in\mathcal{C}} D_\ell(v, v_0)$$

exists and G has finite gradients at v_0 and v^*. Then, for every $v \in \mathcal{C}$ and any point v_0,

$$D_\ell(v, v_0) \geq D_\ell(v, v^*) + D_\ell(v^*, v_0).$$

In particular, existence is automatic when $\mathcal{C}$ is compact and $D_\ell(\cdot, v_0)$ is continuous.

Proof. The first-order variational inequality for the convex projection problem gives

$$\langle \nabla G(v^*) - \nabla G(v_0), v - v^*\rangle \geq 0 \qquad \text{for all } v \in \mathcal{C}.$$

The Bregman three-point identity gives

$$D_G(v, v_0) - D_G(v, v^*) - D_G(v^*, v_0) = \langle \nabla G(v^*) - \nabla G(v_0), v - v^*\rangle,$$

which proves the claim. □

B.2 Proof of Observation 3.2

Observation 3.2. With finite outcome and report supports, $\mathcal{H}_{\mathrm{comp}}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1})$ is nonempty, compact, and convex. Moreover,

$$\mathrm{id} \in \mathcal{H}_{\mathrm{comp}}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}) \quad \Longleftrightarrow \quad \boldsymbol{Q}_0, \boldsymbol{Q}_1 \text{ are mutually calibrated.}$$

Proof. Write

$$p_{01}(q_0, q_1) := \mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}(q_0, q_1), \qquad p_i(q_i) := \mathcal{D}_{\boldsymbol{Q}_i}(q_i), \quad i \in \{0, 1\}.$$

Because $\mathcal{Q}_i$ is the support of $\boldsymbol{Q}_i$, $p_i(q_i) > 0$ for every $q_i \in \mathcal{Q}_i$. Let $\mathcal{Z}$ be the set of nonnegative arrays $z = (z(y, q_0, q_1))_{y,q_0,q_1}$ satisfying

$$\sum_y z(y, q_0, q_1) = p_{01}(q_0, q_1) \qquad \text{for every } q_0 \in \mathcal{Q}_0,\ q_1 \in \mathcal{Q}_1,$$

$$\sum_{q_1} z(y, q_0, q_1) = q_0(y)p_0(q_0) \qquad \text{for every } q_0 \in \mathcal{Q}_0,\ y \in \mathcal{Y}.$$

These are precisely the fixed-report-marginal and reference-calibration constraints in the first two lines of Equation (3). For each $z \in \mathcal{Z}$, define, on the positive-probability support $\mathcal{Q}_1$,

$$h_z(q_1)(y) := \frac{\sum_{q_0} z(y, q_0, q_1)}{p_1(q_1)}. \tag{7}$$

The denominator is positive, and the first constraint gives

$$\sum_y h_z(q_1)(y) = \frac{\sum_{q_0,y} z(y, q_0, q_1)}{p_1(q_1)} = \frac{\sum_{q_0} p_{01}(q_0, q_1)}{p_1(q_1)} = 1.$$

Thus $h_z(q_1) \in \Delta(\mathcal{Y})$.

Every $z \in \mathcal{Z}$ defines an extension by setting $\widetilde{\mathcal{D}}_z(q_0, q_1, y) := z(y, q_0, q_1)$. Its report marginal is $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$, and the second constraint together with Equation (7) gives

$$\Pr_{\widetilde{\mathcal{D}}_z}(Y = y \mid \boldsymbol{Q}_0 = q_0) = q_0(y), \qquad \Pr_{\widetilde{\mathcal{D}}_z}(Y = y \mid \boldsymbol{Q}_1 = q_1) = h_z(q_1)(y).$$

Hence the extension calibrates the reference and induces h_z as its primary-bin posterior map. Conversely, the joint mass function of every extension appearing in Equation (2) belongs to $\mathcal{Z}$ and induces its map through Equation (7). Hence

$$\mathcal{H}_{\mathrm{comp}}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}) = \{h_z : z \in \mathcal{Z}\}.$$

The set $\mathcal{Z}$ is nonempty: one feasible array is

$$z^0(y, q_0, q_1) := p_{01}(q_0, q_1)q_0(y),$$

which corresponds to drawing $Y \sim q_0$ conditional on $(\boldsymbol{Q}_0, \boldsymbol{Q}_1) = (q_0, q_1)$. The set is convex and closed, and nonnegativity together with the first constraint implies $0 \le z(y, q_0, q_1) \le p_{01}(q_0, q_1)$. Thus $\mathcal{Z}$ is bounded and, in this finite-dimensional space, compact. The map $z \mapsto h_z$ is linear and continuous because the denominators $p_1(q_1)$ are fixed and positive. Its image $\mathcal{H}_{\mathrm{comp}}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1})$ is consequently nonempty, compact, and convex.

It remains to identify when the identity belongs to this set. If $\mathrm{id} \in \mathcal{H}_{\mathrm{comp}}$, its witnessing extension satisfies

$$\mathbb{E}[\mathbf{Y} \mid \boldsymbol{Q}_0] = \boldsymbol{Q}_0, \qquad \mathbb{E}[\mathbf{Y} \mid \boldsymbol{Q}_1] = \mathrm{id}(\boldsymbol{Q}_1) = \boldsymbol{Q}_1.$$

Thus both predictors are calibrated under the same extension, so they are mutually calibrated. Conversely, any extension witnessing mutual calibration has primary-bin posterior map $q_1 \mapsto \mathbb{E}[\mathbf{Y} \mid \boldsymbol{Q}_1 = q_1] = q_1$, which is the identity and therefore belongs to the Reference-Compatible Set. □

B.3 Proof of Theorem 3.3

Theorem 3.3. *Given a joint distribution over primary and reference predictors $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$. Let h^* be defined in Equation* (4)*. For every extension $\widetilde{\mathcal{D}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1,Y}$ with $(\boldsymbol{Q}_0,\boldsymbol{Q}_1)$-marginal $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$ under which the reference predictor is calibrated,*

$$\begin{aligned} &\mathbb{E}_{\widetilde{\mathcal{D}}}[\ell(\boldsymbol{Q}_1,Y)] - \mathbb{E}_{\widetilde{\mathcal{D}}}[\ell(h^*(\boldsymbol{Q}_1),Y)] \geq \delta, \\ &\delta := \min_{h\in\mathcal{H}_{\mathrm{comp}}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1})} \mathbb{E}_{\mathcal{D}_{\boldsymbol{Q}_1}}[D_\ell(h(\boldsymbol{Q}_1),\boldsymbol{Q}_1)]. \end{aligned} \tag{5}$$

The value δ is strictly positive if and only if $\boldsymbol{Q}_0$ and $\boldsymbol{Q}_1$ are not mutually calibrated. Consequently, a strictly improving deterministic transformation exists if and only if mutual calibration fails. Moreover, some admissible extension attains equality in Equation (5)*, so δ is the exact worst-case gain.*

Proof. Fix an arbitrary admissible extension $\widetilde{\mathcal{D}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1,Y}$ and define its posterior map on the original primary-report bins by

$$\eta_{\widetilde{\mathcal{D}}}(\boldsymbol{q}_1) = \mathbb{E}_{\widetilde{\mathcal{D}}}[\mathbf{Y} \mid \boldsymbol{Q}_1 = \boldsymbol{q}_1].$$

The extension $\widetilde{\mathcal{D}}$ itself witnesses $\eta_{\widetilde{\mathcal{D}}} \in \mathcal{H}_{\mathrm{comp}}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1})$ under Equation (2).

For every primary report $\boldsymbol{q}_1$ with positive probability, the conditional distribution of Y given $\boldsymbol{Q}_1 = \boldsymbol{q}_1$ under the extension $\widetilde{\mathcal{D}}$ is $\eta_{\widetilde{\mathcal{D}}}(\boldsymbol{q}_1)$. Hence,

$$\begin{aligned} \mathbb{E}_{\widetilde{\mathcal{D}}}[\ell(\boldsymbol{q}_1,Y) - \ell(h^*(\boldsymbol{q}_1),Y) \,|\, \boldsymbol{Q}_1 = \boldsymbol{q}_1] &= \mathbb{E}_{Y\sim\eta_{\widetilde{\mathcal{D}}}(\boldsymbol{q}_1)}[\ell(\boldsymbol{q}_1,Y) - \ell(h^*(\boldsymbol{q}_1),Y)] \\ &= \mathbb{E}_{Y\sim\eta_{\widetilde{\mathcal{D}}}(\boldsymbol{q}_1)}\big[\ell(\boldsymbol{q}_1,Y) - \ell(\eta_{\widetilde{\mathcal{D}}}(\boldsymbol{q}_1),Y)\big] \\ &\quad - \mathbb{E}_{Y\sim\eta_{\widetilde{\mathcal{D}}}(\boldsymbol{q}_1)}\big[\ell(h^*(\boldsymbol{q}_1),Y) - \ell(\eta_{\widetilde{\mathcal{D}}}(\boldsymbol{q}_1),Y)\big] \\ &= D_\ell\big(\eta_{\widetilde{\mathcal{D}}}(\boldsymbol{q}_1),\boldsymbol{q}_1\big) - D_\ell\big(\eta_{\widetilde{\mathcal{D}}}(\boldsymbol{q}_1),h^*(\boldsymbol{q}_1)\big), \end{aligned} \tag{8}$$

where the last equality follows from Equation (6). Applying the law of total expectation therefore gives

$$\begin{aligned} \mathbb{E}_{\widetilde{\mathcal{D}}}[\ell(\boldsymbol{Q}_1,Y)] - \mathbb{E}_{\widetilde{\mathcal{D}}}[\ell(h^*(\boldsymbol{Q}_1),Y)] &= \mathbb{E}_{\mathcal{D}_{\boldsymbol{Q}_1}}\big[D_\ell\big(\eta_{\widetilde{\mathcal{D}}}(\boldsymbol{Q}_1),\boldsymbol{Q}_1\big) - D_\ell\big(\eta_{\widetilde{\mathcal{D}}}(\boldsymbol{Q}_1),h^*(\boldsymbol{Q}_1)\big)\big] \\ &= \mathbb{E}_{\mathcal{D}_{\boldsymbol{Q}_1}}\big[D_\ell\big(\eta_{\widetilde{\mathcal{D}}}(\boldsymbol{Q}_1),\mathrm{id}(\boldsymbol{Q}_1)\big)\big] - \mathbb{E}_{\mathcal{D}_{\boldsymbol{Q}_1}}\big[D_\ell\big(\eta_{\widetilde{\mathcal{D}}}(\boldsymbol{Q}_1),h^*(\boldsymbol{Q}_1)\big)\big] \\ &= D_\ell(\eta_{\widetilde{\mathcal{D}}},\mathrm{id}) - D_\ell(\eta_{\widetilde{\mathcal{D}}},h^*). \end{aligned} \tag{9}$$

By Observation 3.2, the compatible posterior set is nonempty, compact, and convex. Hence the projection h^* exists. Applying Lemma B.4 to the comparison point $\eta_{\widetilde{\mathcal{D}}}$ gives

$$D_\ell(\eta_{\widetilde{\mathcal{D}}},\mathrm{id}) \geq D_\ell(\eta_{\widetilde{\mathcal{D}}},h^*) + D_\ell(h^*,\mathrm{id}).$$

Substitution into Equation (9) proves the uniform lower bound Equation (5).

We next prove strictness and necessity.

- If the predictors are not mutually calibrated, Observation 3.2 implies that $\mathrm{id} \notin \mathcal{H}_{\mathrm{comp}}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1})$. The attained projection h^* therefore differs from the identity id on a positive-mass report, and strictness yields $\delta = D_\ell(h^*,\mathrm{id}) > 0$. Thus h^* strictly improves the primary predictor $\boldsymbol{Q}_1$ under every possible ground truth extension.
- Conversely, suppose the predictors are mutually calibrated. Then there exists an admissible extension $\widetilde{\mathcal{D}}^0$ under which both predictors are calibrated. In particular,

  $$\mathbb{E}_{\widetilde{\mathcal{D}}^0}[\mathbf{Y} \mid \boldsymbol{Q}_1] = \boldsymbol{Q}_1.$$

  Therefore, conditional on each primary report, $\boldsymbol{Q}_1$ is the true conditional distribution of the outcome. For any deterministic transformation $h : \mathcal{Q}_1 \to \Delta(\mathcal{Y})$, strictness gives

  $$\mathbb{E}_{\widetilde{\mathcal{D}}^0}[\ell(h(\boldsymbol{Q}_1),Y) \,|\, \boldsymbol{Q}_1] \geq \mathbb{E}_{\widetilde{\mathcal{D}}^0}[\ell(\boldsymbol{Q}_1,Y) \,|\, \boldsymbol{Q}_1]\,.$$

  Taking expectations over $\boldsymbol{Q}_1$ yields

  $$\mathbb{E}_{\widetilde{\mathcal{D}}^0}[\ell(h(\boldsymbol{Q}_1),Y)] \geq \mathbb{E}_{\widetilde{\mathcal{D}}^0}[\ell(\boldsymbol{Q}_1,Y)],$$

with equality only if $h(\boldsymbol{Q}_1) = \boldsymbol{Q}_1$ almost surely.

On the other hand, any transformation that weakly improves the primary predictor under every admissible extension must satisfy the reverse inequality under $\widetilde{\mathcal{D}}^0$. Hence equality is necessary, and strict propriety implies that such a transformation must coincide with the identity on the support of $\mathcal{D}_{\boldsymbol{Q}_1}$. The identity cannot yield a strict improvement under any extension. Consequently, no strictly improving deterministic transformation exists when the predictors are mutually calibrated.

It remains to establish exact worst-case equality. Since $h^* \in \mathcal{H}_{\mathrm{comp}}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1})$, the definition of the reference-compatible set guarantees the existence of an admissible extension $\widetilde{\mathcal{D}}^*$ satisfying

$$\mathbb{E}_{\widetilde{\mathcal{D}}^*}[\mathbf{Y} \mid \boldsymbol{Q}_1] = h^*(\boldsymbol{Q}_1).$$

Thus, conditional on $\boldsymbol{Q}_1$, the true outcome distribution under $\widetilde{\mathcal{D}}^*$ is $h^*(\boldsymbol{Q}_1)$. The pointwise regret definition therefore gives

$$\mathbb{E}_{\widetilde{\mathcal{D}}^*}[\ell(\boldsymbol{Q}_1, Y) - \ell(h^*(\boldsymbol{Q}_1), Y) \,|\, \boldsymbol{Q}_1] = D_\ell\big(h^*(\boldsymbol{Q}_1), \boldsymbol{Q}_1\big).$$

Taking expectations over $\boldsymbol{Q}_1$ and using the fact that every admissible extension has the fixed marginal $\mathcal{D}_{\boldsymbol{Q}_1}$ yields

$$\mathbb{E}_{\widetilde{\mathcal{D}}^*}[\ell(\boldsymbol{Q}_1, Y)] - \mathbb{E}_{\widetilde{\mathcal{D}}^*}[\ell(h^*(\boldsymbol{Q}_1), Y)] = \mathbb{E}_{\mathcal{D}_{\boldsymbol{Q}_1}}\big[D_\ell\big(h^*(\boldsymbol{Q}_1), \boldsymbol{Q}_1\big)\big] = \delta,$$

where the final equality follows from the definitions of h^* and δ in Equations (4) and (5).

We have already shown that the loss improvement is at least δ under every admissible extension, while the admissible extension $\widetilde{\mathcal{D}}^*$ constructed above attains equality. Hence δ is the exact worst-case loss improvement. □

C Performance Guarantees under an Imperfectly Calibrated Reference

We now allow the actual reference to be imperfectly calibrated. To distinguish the probability laws used below, write

$$\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1} := \mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$$

for the observed joint report law. The reference-compatible set in Equation (2) is constructed from this law and generic reference-calibrated extensions $\widetilde{\mathcal{D}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1,Y}$. In contrast, let $\mathcal{D}^{\mathrm{act}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1,Y}$ denote the actual joint law, whose $(\boldsymbol{Q}_0, \boldsymbol{Q}_1)$-marginal is $\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$ but under which the reference need not be calibrated. The projection h^* is computed only from $\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$.

Identify each outcome with its one-hot vector $\mathbf{Y}$. For $i \in \{0, 1\}$, define the actual posterior map

$$\eta_i^{\mathrm{act}}(\boldsymbol{q}_i) := \mathbb{E}_{\mathcal{D}^{\mathrm{act}}}[\mathbf{Y} \mid \boldsymbol{Q}_i = \boldsymbol{q}_i],$$

and define the actual reference calibration error by

$$\epsilon_{\mathrm{ref}} := \mathbb{E}_{\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_0}}\left[\left\|\eta_0^{\mathrm{act}}(\boldsymbol{Q}_0) - \boldsymbol{Q}_0\right\|_1\right].$$

Theorem C.1. *Assume finite outcome and report supports and a strictly proper loss ℓ generated by a differentiable, strictly convex function G with finite gradients on the prediction domain. Let h^* be the projection in Equation* (4) *computed from the observed $\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$, and $\mathcal{D}^{\mathrm{act}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1,Y}$ be the actual joint distribution involving ground truth Y with $(\boldsymbol{Q}_0, \boldsymbol{Q}_1)$-margin $\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$. Define*

$$A := \mathbb{E}_{\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_1}}\big[D_\ell(h^*(\boldsymbol{Q}_1), \boldsymbol{Q}_1)\big], \tag{10}$$

$$B := \left(\mathbb{E}_{\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_1}}\left[\|\nabla G(h^*(\boldsymbol{Q}_1)) - \nabla G(\boldsymbol{Q}_1)\|_2^2\right]\right)^{1/2}. \tag{11}$$

Then the actual loss improvement satisfies

$$\begin{aligned} \Delta L^{\mathrm{act}} &:= \mathbb{E}_{\mathcal{D}^{\mathrm{act}}}[\ell(\boldsymbol{Q}_1, Y)] - \mathbb{E}_{\mathcal{D}^{\mathrm{act}}}[\ell(h^*(\boldsymbol{Q}_1), Y)] \\ &\geq A - B\sqrt{\epsilon_{\mathrm{ref}}}. \end{aligned} \tag{12}$$

Consequently, h^* strictly improves expected loss whenever $A > B\sqrt{\epsilon_{\mathrm{ref}}}$.

Proof. The proof proceeds in three steps: derive an exact decomposition around an arbitrary auxiliary reference-compatible map, construct a nearby map by repairing the outcome, and specialize the decomposition to that repaired posterior map to control the residue.

Step 1: derive the exact auxiliary decomposition. Define the gradient-displacement field

$$\xi^*(\boldsymbol{q}_1) := \nabla G(h^*(\boldsymbol{q}_1)) - \nabla G(\boldsymbol{q}_1).$$

Following Equation (8), we can first express the actual loss improvement as

$$\Delta L^{\mathrm{act}} = \mathbb{E}_{\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_1}}\left[D_\ell(\eta_1^{\mathrm{act}}(\boldsymbol{Q}_1), \boldsymbol{Q}_1) - D_\ell(\eta_1^{\mathrm{act}}(\boldsymbol{Q}_1), h^*(\boldsymbol{Q}_1))\right]. \tag{13}$$

Choose any auxiliary map

$$h_{\mathrm{aux}} \in \mathcal{H}_{\mathrm{comp}}(\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}).$$

Applying the Bregman three-point identity and splitting

$$\eta_1^{\mathrm{act}} - h^* = (h_{\mathrm{aux}} - h^*) + (\eta_1^{\mathrm{act}} - h_{\mathrm{aux}})$$

gives the following three-term decomposition:

$$\begin{aligned}\Delta L^{\mathrm{act}} = & \underbrace{A}_{\text{Geometric Gain (Always } \geq 0)} \\ & + \underbrace{\mathbb{E}_{\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_1}}[\langle \xi^*(\boldsymbol{Q}_1), h_{\mathrm{aux}}(\boldsymbol{Q}_1) - h^*(\boldsymbol{Q}_1)\rangle]}_{\text{Compatibility Bonus (Always } \geq 0)} \\ & + \underbrace{\mathbb{E}_{\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_1}}\left[\left\langle \xi^*(\boldsymbol{Q}_1), \eta_1^{\mathrm{act}}(\boldsymbol{Q}_1) - h_{\mathrm{aux}}(\boldsymbol{Q}_1)\right\rangle\right]}_{\text{Miscalibration Risk (Can be } \leq 0)}.\end{aligned} \tag{14}$$

Indeed, since h^* minimizes the Bregman displacement over the convex reference-compatible set, its first-order variational inequality in the feasible direction $h_{\mathrm{aux}} - h^*$ is

$$\mathbb{E}_{\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_1}}[\langle \xi^*(\boldsymbol{Q}_1), h_{\mathrm{aux}}(\boldsymbol{Q}_1) - h^*(\boldsymbol{Q}_1)\rangle] \geq 0.$$

This proves the non-negativity of the compatibility bonus in Equation (14).

Step 2: construct a nearby reference-compatible posterior map. For each reference report $\boldsymbol{q}_0$, choose a maximal coupling $\pi_{\boldsymbol{q}_0}$ whose first marginal is $\eta_0^{\mathrm{act}}(\boldsymbol{q}_0)$ and whose second marginal is $\boldsymbol{q}_0$. Thus

$$\pi_{\boldsymbol{q}_0}\left(\{(y, y^{\mathrm{rep}}) : y \neq y^{\mathrm{rep}}\}\right) = \frac{1}{2}\|\eta_0^{\mathrm{act}}(\boldsymbol{q}_0) - \boldsymbol{q}_0\|_1.$$

Let $K^{\mathrm{rep}}_{\boldsymbol{q}_0}(y^{\mathrm{rep}} \mid y)$ be a transition kernel realizing this coupling. We make the construction explicit by defining the augmented coupling law

$$\begin{aligned}&\mathcal{D}^{\mathrm{cpl}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1,Y,Y^{\mathrm{rep}}}(\boldsymbol{q}_0, \boldsymbol{q}_1, y, y^{\mathrm{rep}}) \\ &\quad := \mathcal{D}^{\mathrm{act}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1,Y}(\boldsymbol{q}_0, \boldsymbol{q}_1, y) K^{\mathrm{rep}}_{\boldsymbol{q}_0}(y^{\mathrm{rep}} \mid y),\end{aligned}$$

and its repaired marginal

$$\begin{aligned}&\mathcal{D}^{\mathrm{rep}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1,Y^{\mathrm{rep}}}(\boldsymbol{q}_0, \boldsymbol{q}_1, y^{\mathrm{rep}}) \\ &\quad := \sum_{y \in \mathcal{Y}} \mathcal{D}^{\mathrm{cpl}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1,Y,Y^{\mathrm{rep}}}(\boldsymbol{q}_0, \boldsymbol{q}_1, y, y^{\mathrm{rep}}).\end{aligned}$$

The coupling law has $\mathcal{D}^{\mathrm{act}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1,Y}$ as its $(\boldsymbol{Q}_0, \boldsymbol{Q}_1, Y)$-marginal, while the repaired law has $\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$ as its report marginal. Moreover, conditional on $\boldsymbol{Q}_0 = \boldsymbol{q}_0$, the (Y, Y^{rep})-marginal under $\mathcal{D}^{\mathrm{cpl}}$ is $\pi_{\boldsymbol{q}_0}$. The coupling construction therefore gives

$$\mathbb{E}_{\mathcal{D}^{\mathrm{rep}}}[\mathbf{Y}^{\mathrm{rep}} \mid \boldsymbol{Q}_0] = \boldsymbol{Q}_0.$$

Define the posterior induced by the repaired law on each original primary-report bin by

$$\eta_1^{\mathrm{rep}}(\boldsymbol{q}_1) := \mathbb{E}_{\mathcal{D}^{\mathrm{rep}}}[\mathbf{Y}^{\mathrm{rep}} \mid \boldsymbol{Q}_1 = \boldsymbol{q}_1].$$

When Y^{rep} is treated as the outcome coordinate, the law $\mathcal{D}^{\mathrm{rep}}$ is a reference-calibrated extension of the observed report law and witnesses the compatibility of η_1^{rep}. By Equation (2), we have

$$\eta_1^{\mathrm{rep}} \in \mathcal{H}_{\mathrm{comp}}(\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}). \tag{15}$$

Because $\mathcal{D}^{\mathrm{cpl}}$ has the actual and repaired laws as the corresponding marginals, conditional Jensen's inequality gives

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_1}} \|\eta_1^{\mathrm{act}}(\boldsymbol{Q}_1) - \eta_1^{\mathrm{rep}}(\boldsymbol{Q}_1)\|_1 \\
&= \mathbb{E}_{\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_1}} \|\mathbb{E}_{\mathcal{D}^{\mathrm{cpl}}}[\mathbf{Y} - \mathbf{Y}^{\mathrm{rep}} \mid \boldsymbol{Q}_1]\|_1 \\
&\leq \mathbb{E}_{\mathcal{D}^{\mathrm{cpl}}} \|\mathbf{Y} - \mathbf{Y}^{\mathrm{rep}}\|_1 \\
&= 2\,\mathbb{E}_{\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_0}} \left[\Pr_{\mathcal{D}^{\mathrm{cpl}}}(Y \neq Y^{\mathrm{rep}} \mid \boldsymbol{Q}_0) \right] \\
&= \mathbb{E}_{\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_0}} \|\eta_0^{\mathrm{act}}(\boldsymbol{Q}_0) - \boldsymbol{Q}_0\|_1 = \epsilon_{\mathrm{ref}}.
\end{aligned} \tag{16}$$

For probability vectors $\boldsymbol{p}, \boldsymbol{q}$, $\|\boldsymbol{p} - \boldsymbol{q}\|_2^2 \leq \|\boldsymbol{p} - \boldsymbol{q}\|_1$. Therefore,

$$\mathbb{E}_{\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_1}} \|\eta_1^{\mathrm{act}}(\boldsymbol{Q}_1) - \eta_1^{\mathrm{rep}}(\boldsymbol{Q}_1)\|_2^2 \leq \epsilon_{\mathrm{ref}}. \tag{17}$$

Step 3: specialize the auxiliary map and control the residual. By Equation (15), the repaired posterior from Step 2 is a valid choice $h_{\mathrm{aux}} = \eta_1^{\mathrm{rep}}$ in Equation (14). Cauchy–Schwarz in the space of transformations weighted by $\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_1}$ and Equation (17) give

$$\begin{aligned}
&\left| \mathbb{E}_{\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_1}} \left[\left\langle \xi^*(\boldsymbol{Q}_1), \eta_1^{\mathrm{act}}(\boldsymbol{Q}_1) - \eta_1^{\mathrm{rep}}(\boldsymbol{Q}_1) \right\rangle \right] \right| \\
&\quad \leq \left(\mathbb{E}_{\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_1}} \|\xi^*(\boldsymbol{Q}_1)\|_2^2 \right)^{1/2} \left(\mathbb{E}_{\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_1}} \|\eta_1^{\mathrm{act}}(\boldsymbol{Q}_1) - \eta_1^{\mathrm{rep}}(\boldsymbol{Q}_1)\|_2^2 \right)^{1/2} \\
&\quad \leq B\sqrt{\epsilon_{\mathrm{ref}}}.
\end{aligned} \tag{18}$$

The compatible bonus term in Equation (14) is nonnegative by Step 1, whereas its signed residual (miscalibration risk) is bounded below by the negative of the left-hand side of Equation (18). Therefore,

$$\begin{aligned}
\Delta L^{\mathrm{act}} &\geq A - \left| \mathbb{E}_{\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_1}} \left[\left\langle \xi^*(\boldsymbol{Q}_1), \eta_1^{\mathrm{act}}(\boldsymbol{Q}_1) - \eta_1^{\mathrm{rep}}(\boldsymbol{Q}_1) \right\rangle \right] \right| \\
&\geq A - B\sqrt{\epsilon_{\mathrm{ref}}},
\end{aligned}$$

which proves Equation (12). □

Corollary C.2 (Strongly convex and smooth generator)**.** *In the setting of Theorem C.1, suppose additionally that G is μ_G-strongly convex and L_G-smooth on the prediction domain. Then*

$$B \leq L_G \sqrt{\frac{2A}{\mu_G}},$$

and hence

$$\Delta L^{\mathrm{act}} \geq A - L_G \sqrt{\frac{2A\epsilon_{\mathrm{ref}}}{\mu_G}}. \tag{19}$$

In particular, the right-hand side is positive whenever $A > 2L_G^2 \epsilon_{\mathrm{ref}}/\mu_G$.

Proof. Smoothness and strong convexity imply, pointwise,

$$\begin{aligned}
\|\nabla G(h^*(\boldsymbol{Q}_1)) - \nabla G(\boldsymbol{Q}_1)\|_2 &\leq L_G \|h^*(\boldsymbol{Q}_1) - \boldsymbol{Q}_1\|_2, \\
D_\ell(h^*(\boldsymbol{Q}_1), \boldsymbol{Q}_1) &\geq \frac{\mu_G}{2} \|h^*(\boldsymbol{Q}_1) - \boldsymbol{Q}_1\|_2^2.
\end{aligned}$$

Square the first inequality, average under $\mathcal{D}^{\mathrm{obs}}_{\boldsymbol{Q}_1}$, and apply the second to obtain the bound on B. The remaining claims follow from Equation (12). □

D Extension to Elicitable Properties

D.1 Setup and Definitions

Models often report summary statistics, such as an answer and its confidence, rather than full predictive distributions. We therefore extend our framework to elicitable properties Lambert et al. [2008]. We show that in this setting, we can still obtain an optimal worst-case improvement guarantee under the original property loss by combining a convex projection over conditional distributions with a suitably chosen, possibly randomized report-selection rule. The guarantee applies to both differentiable and nondifferentiable Bayes risks.

Let $\mathcal{Y}$ be a finite nonempty outcome space. For $i \in \{0, 1\}$, let $\mathcal{A}_i$ be the allowable report space and let $\Gamma_i : \Delta(\mathcal{Y}) \to 2^{\mathcal{A}_i}$ be a nonempty-valued property. The observed predictor R_i has finite support $\mathcal{R}_i := \mathrm{supp}(R_i) \subseteq \mathcal{A}_i$. We fix the observed joint distribution $\mathcal{D}_{R_0,R_1}$ throughout and write $\mu_r := \Pr(R_1 = r) > 0$ for each $r \in \mathcal{R}_1$. We equip the primary report space $\mathcal{A}_1$ with a Hausdorff topology. Probability simplices and spaces of posterior maps carry their usual finite-dimensional Euclidean topologies. These topologies allow us to state the compactness and continuity assumptions used to establish the existence of optimal projections and report-selection rules. The level set of a report r is

$$\Gamma_i^{-1}(r) := \{\boldsymbol{q} \in \Delta(\mathcal{Y}) : r \in \Gamma_i(\boldsymbol{q})\}.$$

Under a joint distribution $\widetilde{\mathcal{D}}$, predictor R_i is calibrated with respect to Γ_i if

$$r_i \in \Gamma_i\Big(\Pr_{\widetilde{\mathcal{D}}}(Y = \cdot \mid R_i = r_i)\Big) \qquad \text{for every } r_i \in \mathcal{R}_i.$$

All conditional constraints below refer to original positive-probability report bins.

Definition D.1 (Mutual calibration for elicitable properties)**.** Reports R_0, R_1 are mutually calibrated with respect to Γ_0, Γ_1 if there exists an extension $\widetilde{\mathcal{D}}_{R_0,R_1,Y}$ of $\mathcal{D}_{R_0,R_1}$ that calibrates both predictors with respect to their properties. Such an extension is a witness to mutual calibration.

Loss and property regret. Let $\ell_i : \mathcal{A}_i \times \mathcal{Y} \to \mathbb{R}$ be finite-valued and strictly proper for Γ_i, meaning that

$$\operatorname*{argmin}_{r' \in \mathcal{A}_i} \mathbb{E}_{Y \sim \boldsymbol{q}}[\ell_i(r', Y)] = \Gamma_i(\boldsymbol{q}) \qquad \text{for every } \boldsymbol{q} \in \Delta(\mathcal{Y}).$$

Write $\mathbf{L}_{i,r} := (\ell_i(r, y))_{y \in \mathcal{Y}}$ for the loss vector of a report, and abbreviate $\mathbf{L}_r := \mathbf{L}_{1,r}$. The primary loss is continuous in the report if, for each fixed outcome y, the map $r \mapsto \ell_1(r, y)$ is continuous on $\mathcal{A}_1$: for every open set $U \subseteq \mathbb{R}$, its preimage $\{r \in \mathcal{A}_1 : \ell_1(r, y) \in U\}$ is open in the chosen topology on $\mathcal{A}_1$. Since $\mathcal{Y}$ is finite, this coordinatewise continuity is equivalent to continuity of the loss-vector map $r \mapsto \mathbf{L}_r$ into $\mathbb{R}^{|\mathcal{Y}|}$ equipped with its Euclidean topology. We impose this continuity condition, together with compactness of $\mathcal{A}_1$, in Theorem D.3. The primary loss has Bayes risk

$$\underline{L}_1(\boldsymbol{q}) := \min_{r' \in \mathcal{A}_1} \langle \mathbf{L}_{r'}, \boldsymbol{q} \rangle$$

and property regret

$$\begin{aligned} \rho_{\ell_1}(\boldsymbol{q}, r) &:= \langle \mathbf{L}_r, \boldsymbol{q} \rangle - \underline{L}_1(\boldsymbol{q}) \\ &= \max_{r' \in \mathcal{A}_1} \langle \mathbf{L}_r - \mathbf{L}_{r'}, \boldsymbol{q} \rangle. \end{aligned} \tag{20}$$

The Bayes risk is a pointwise infimum of linear functions of $\boldsymbol{q}$ and is therefore concave. Property regret is convex because it subtracts this concave Bayes risk from a report's linear expected loss, and nonnegative because the Bayes risk is the minimum expected loss over all reports. Strict properness gives

$$\rho_{\ell_1}(\boldsymbol{q}, r) = 0 \quad \iff \quad r \in \Gamma_1(\boldsymbol{q}).$$

Report-selection rules and gain. Let $\mathfrak{D}_{\Gamma_0}$ be the extensions of the observed joint distribution under which R_0 is calibrated with respect to Γ_0. A report-selection rule $\pi = (\pi_r)_{r \in \mathcal{R}_1}$ assigns a finitely supported probability distribution on $\mathcal{A}_1$ to each primary report. On input $R_1 = r$, the rule outputs $\widehat{R} \sim \pi_r$, using randomness independent of (R_0, Y) conditional on R_1. Deterministic rules correspond to point masses. Let Π denote the collection of these rules and define

$$\mathrm{Gain}(\pi; \widetilde{\mathcal{D}}) := \mathbb{E}_{\widetilde{\mathcal{D}}, \pi}\Big[\ell_1(R_1, Y) - \ell_1(\widehat{R}, Y)\Big]. \tag{21}$$

This comparison uses the original property loss for both reports. We call a rule uniformly strictly improving if there exists a constant $\gamma > 0$ such that $\mathrm{Gain}(\pi; \widetilde{\mathcal{D}}) \geq \gamma$ for every $\widetilde{\mathcal{D}} \in \mathfrak{D}_{\Gamma_0}$.

D.2 Generalized Reference-Compatible Set and Algorithm

For $\widetilde{\mathcal{D}} \in \mathfrak{D}_{\Gamma_0}$, define its primary posterior map by

$$\eta_{\widetilde{\mathcal{D}}}(r) := \Pr_{\widetilde{\mathcal{D}}}(Y = \cdot \mid R_1 = r).$$

The reference-compatible set consists of all such maps,

$$\mathcal{C}_\Gamma := \mathcal{H}^{\Gamma_0}_{\mathrm{comp}}(\mathcal{D}_{R_0,R_1}) = \{\eta_{\widetilde{\mathcal{D}}} : \widetilde{\mathcal{D}} \in \mathfrak{D}_{\Gamma_0}\}. \tag{22}$$

Thus membership of h requires a single extension that calibrates the reference and satisfies $h(r) = \Pr_{\widetilde{\mathcal{D}}}(Y = \cdot \mid R_1 = r)$ on each original primary-report bin.

Observation D.2. Suppose the outcome and observed-report supports are finite, the losses ℓ_i are finite-valued and strictly proper for Γ_i, and $\Gamma_i^{-1}(r) \neq \varnothing$ for every $r \in \mathcal{R}_i$ and $i \in \{0, 1\}$. Every observed level set $\Gamma_i^{-1}(r_i)$ is nonempty, closed, and convex. Consequently, $\mathcal{C}_\Gamma$ is nonempty, compact, and convex.

Proof. Strict properness gives

$$\Gamma_i^{-1}(r_i) = \bigcap_{r' \in \mathcal{A}_i} \{\boldsymbol{q} \in \Delta(\mathcal{Y}) : \langle \mathbf{L}_{i,r_i} - \mathbf{L}_{i,r'}, \boldsymbol{q} \rangle \leq 0\}.$$

For each fixed pair of reports, the function $\boldsymbol{q} \mapsto \langle \mathbf{L}_{i,r_i} - \mathbf{L}_{i,r'}, \boldsymbol{q} \rangle$ is linear and continuous in the probability vector $\boldsymbol{q}$. The intersection is therefore closed and convex relative to the simplex, and it is nonempty by assumption.

An extension of the fixed report marginal is specified by a kernel $K(y \mid r_0, r_1)$. Choose $\boldsymbol{q}_{r_0} \in \Gamma_0^{-1}(r_0)$ for each reference report and set $K(\cdot \mid r_0, r_1) = \boldsymbol{q}_{r_0}$. This constructs a reference-calibrated extension and hence a point in $\mathcal{C}_\Gamma$. The reference-calibrated kernels form a closed convex subset of a finite product of simplices. Mixing kernels preserves the report marginal, and all conditional distributions are linear in the kernel because the bin probabilities are fixed and positive. The map from a kernel to its primary posterior map is therefore linear and continuous. Its image $\mathcal{C}_\Gamma$ is compact and convex. This argument uses finite loss vectors and the linear dependence on conditional distributions, so it applies without assuming continuity of the losses in their report arguments. □

We first project onto $\mathcal{C}_\Gamma$ by minimizing expected property regret,

$$\begin{aligned} \Phi(h) &:= \mathbb{E}_{\mathcal{D}_{R_1}}\left[\rho_{\ell_1}(h(R_1), R_1)\right], \\ h^* &\in \operatorname*{argmin}_{h \in \mathcal{C}_\Gamma} \Phi(h), \qquad \delta_\Gamma := \Phi(h^*). \end{aligned} \tag{23}$$

We then select output reports from $\Gamma_1(h^*(r))$. The theorem below establishes a selection rule with an exact uniform improvement guarantee. The rule may depend on the original report bin and the entire reference-compatible set.

D.3 Exact Improvement Guarantee

Theorem D.3 (Optimal improvement for elicitable properties)**.** *Assume the hypotheses of Observation D.2. Suppose additionally that $\mathcal{A}_1$ is compact in its specified topology and, for every fixed $y \in \mathcal{Y}$, the function $r \mapsto \ell_1(r, y)$ is continuous on $\mathcal{A}_1$. The solution set in Equation* (23) *is nonempty. For every solution h^*, there exists a rule $\pi^* \in \Pi$ such that*

$$\mathrm{supp}(\pi^*_r) \subseteq \Gamma_1(h^*(r)) \qquad \textit{for every } r \in \mathcal{R}_1 \tag{24}$$

and, for every $\widetilde{\mathcal{D}} \in \mathfrak{D}_{\Gamma_0}$,

$$\mathrm{Gain}(\pi^*; \widetilde{\mathcal{D}}) \geq \delta_\Gamma. \tag{25}$$

The bound is attained by an admissible extension and is optimal among all rules in Π,

$$\max_{\pi \in \Pi} \min_{\widetilde{\mathcal{D}} \in \mathfrak{D}_{\Gamma_0}} \mathrm{Gain}(\pi; \widetilde{\mathcal{D}}) = \min_{\widetilde{\mathcal{D}} \in \mathfrak{D}_{\Gamma_0}} \mathrm{Gain}(\pi^*; \widetilde{\mathcal{D}}) = \delta_\Gamma. \tag{26}$$

Moreover, $\delta_\Gamma > 0$ if and only if R_0, R_1 are not mutually calibrated with respect to Γ_0, Γ_1. Thus a uniformly strictly improving rule in Π exists exactly when mutual calibration fails.

Proof. Define the set of report loss vectors and its convex hull by

$$\mathcal{S} := \{\mathbf{L}_{r'} : r' \in \mathcal{A}_1\}, \qquad \mathcal{K} := \operatorname{conv}(\mathcal{S}).$$

Each of the finitely many coordinates of the loss-vector map is continuous by assumption. Thus $\mathcal{S}$ is the continuous image of the compact space $\mathcal{A}_1$ and is compact in $\mathbb{R}^{|\mathcal{Y}|}$. The convex hull of a compact set in a finite-dimensional Euclidean space is compact, so $\mathcal{K}$ is compact and convex. Every vector in $\mathcal{K}$ is the expected loss vector of a finitely supported distribution on reports. A rule $\pi \in \Pi$ is therefore represented, for purposes of expected loss, by $\mathbf{z} = (\mathbf{z}_r)_{r\in\mathcal{R}_1} \in \mathcal{K}^{\mathcal{R}_1}$, where $\mathbf{z}_r := \mathbb{E}_{r'\sim\pi_r}[\mathbf{L}_{r'}]$. Conversely, every such $\mathbf{z}$ is realized by a rule in Π.

For a feasible posterior map h and expected loss vectors $\mathbf{z}$, define

$$F(h, \mathbf{z}) := \sum_{r\in\mathcal{R}_1} \mu_r \langle \mathbf{L}_r - \mathbf{z}_r, h(r)\rangle. \tag{27}$$

Conditional expectation gives $\operatorname{Gain}(\pi; \widetilde{\mathcal{D}}) = F(\eta_{\widetilde{\mathcal{D}}}, \mathbf{z})$. For every fixed h,

$$\begin{aligned} \max_{\mathbf{z}\in\mathcal{K}^{\mathcal{R}_1}} F(h, \mathbf{z}) &= \sum_{r\in\mathcal{R}_1} \mu_r \left[\langle \mathbf{L}_r, h(r)\rangle - \min_{\mathbf{z}_r\in\mathcal{K}} \langle \mathbf{z}_r, h(r)\rangle \right] \\ &= \sum_{r\in\mathcal{R}_1} \mu_r \left[\langle \mathbf{L}_r, h(r)\rangle - \underline{L}_1(h(r)) \right] = \Phi(h). \end{aligned} \tag{28}$$

To establish existence of the projection, let $B := \max_{\mathbf{u}\in\mathcal{S}} \|\mathbf{u}\|_2 < \infty$. For any distributions $\boldsymbol{q}, \boldsymbol{p}$,

$$\begin{aligned} \left|\underline{L}_1(\boldsymbol{q}) - \underline{L}_1(\boldsymbol{p})\right| &\le \max_{\mathbf{u}\in\mathcal{S}} \left|\langle \mathbf{u}, \boldsymbol{q} - \boldsymbol{p}\rangle\right| \\ &\le B\|\boldsymbol{q} - \boldsymbol{p}\|_2. \end{aligned}$$

The Bayes risk is therefore Lipschitz continuous, and Φ is continuous. By Observation D.2, Φ attains its minimum on $\mathcal{C}_\Gamma$.

The two strategy spaces $\mathcal{C}_\Gamma$ and $\mathcal{K}^{\mathcal{R}_1}$ are compact and convex. The payoff F is a finite sum of inner products, so it is continuous and affine in each argument. The minimax theorem [Sion, 1958] yields

$$\begin{aligned} \max_{\mathbf{z}\in\mathcal{K}^{\mathcal{R}_1}} \min_{h\in\mathcal{C}_\Gamma} F(h, \mathbf{z}) &= \min_{h\in\mathcal{C}_\Gamma} \max_{\mathbf{z}\in\mathcal{K}^{\mathcal{R}_1}} F(h, \mathbf{z}) \\ &= \min_{h\in\mathcal{C}_\Gamma} \Phi(h) = \delta_\Gamma. \end{aligned} \tag{29}$$

The minimum over h is a continuous function of $\mathbf{z}$ because F is continuous on the compact product of the strategy spaces. It therefore attains its maximum on $\mathcal{K}^{\mathcal{R}_1}$. Choose an optimal $\mathbf{z}^*$ on the left. Then $F(h, \mathbf{z}^*) \ge \delta_\Gamma$ for every $h \in \mathcal{C}_\Gamma$.

Fix any minimizer h^* of Φ. Since

$$\delta_\Gamma \le F(h^*, \mathbf{z}^*) \le \max_{\mathbf{z}} F(h^*, \mathbf{z}) = \Phi(h^*) = \delta_\Gamma,$$

we have

$$0 = \sum_{r\in\mathcal{R}_1} \mu_r \left[\langle \mathbf{z}^*_r, h^*(r)\rangle - \underline{L}_1(h^*(r)) \right].$$

Each summand is nonnegative, so each bracket is zero. Represent $\mathbf{z}^*_r$ as a finite convex combination of report loss vectors. Every report with positive weight must attain the Bayes risk at $h^*(r)$, since the average attains that risk and each individual risk is at least as large. Strict properness therefore places every such report in $\Gamma_1(h^*(r))$. These combinations define a rule π^* satisfying Equations (24) and (25).

An extension $\widetilde{\mathcal{D}}^*$ witnessing h^* satisfies $\eta_{\widetilde{\mathcal{D}}^*} = h^*$, hence $\operatorname{Gain}(\pi^*; \widetilde{\mathcal{D}}^*) = \delta_\Gamma$. For any other rule π, its conditional expected loss under this extension is at least the Bayes risk in each bin. Consequently,

$$\operatorname{Gain}(\pi; \widetilde{\mathcal{D}}^*) \le \sum_{r\in\mathcal{R}_1} \mu_r \left[\langle \mathbf{L}_r, h^*(r)\rangle - \underline{L}_1(h^*(r)) \right] = \delta_\Gamma,$$

which proves Equation (26).

If mutual calibration holds, the primary posterior map of its witness has zero property regret, so $\delta_\Gamma = 0$. Conversely, if $\delta_\Gamma = 0$, nonnegative regret and positive bin probabilities imply $r \in \Gamma_1(h^*(r))$ for every $r \in \mathcal{R}_1$. The extension witnessing h^* then calibrates both original predictors. This proves the equivalence. When $\delta_\Gamma = 0$, Equation (26) and its attaining witness also exclude a rule that strictly improves under every feasible extension. □

Selecting reports after projection. For a fixed optimal h^*, the required selection can be expressed as a variational condition. Choose

$$\mathbf{z}_r \in \operatorname{conv}\{\mathbf{L}_{r'} : r' \in \Gamma_1(h^*(r))\}.$$

Since $F(h^*, \mathbf{z}) = \delta_\Gamma$, such a selection attains the theorem's guarantee if and only if

$$\mathbb{E}_{\mathcal{D}_{R_1}}\left[\langle \mathbf{L}_{R_1} - \mathbf{z}_{R_1}, h(R_1) - h^*(R_1)\rangle\right] \geq 0 \quad \text{for every } h \in \mathcal{C}_\Gamma. \tag{30}$$

The proof establishes existence of such a selection. It can be found by solving the maximization in Equation (29) with each $\mathbf{z}_r$ restricted to the displayed convex hull. The optimal value remains δ_Γ.

D.4 Differentiable Case

We now show that the optimal worst-case guarantee can be attained by a deterministic report-selection rule when the Bayes risk is differentiable at the projected conditional distributions. Under this condition, all Bayes-optimal reports at each projected distribution have the same loss vector, so any deterministic selection preserves the guarantee of Theorem D.3.

Here, differentiability concerns the Bayes risk as a function of its distribution argument. To make this precise, including at simplex boundaries, extend the Bayes risk to the ambient vector space:

$$\underline{L}_1(\mathbf{v}) := \min_{r' \in \mathcal{A}_1} \langle \mathbf{L}_{r'}, \mathbf{v}\rangle, \qquad \mathbf{v} \in \mathbb{R}^{|\mathcal{Y}|}.$$

Under the theorem's assumptions, this extension is finite and concave. All derivatives below are taken with respect to the ambient coordinates $\mathbf{v}$ and evaluated at probability vectors. The deterministic guarantee requires differentiability only at the projected distributions $h^*(r)$ for $r \in \mathcal{R}_1$.

At a distribution $\boldsymbol{q}$ where this extension is differentiable, define

$$\ell_1^\Delta(\boldsymbol{q}, y) := \underline{L}_1(\boldsymbol{q}) + \langle \nabla \underline{L}_1(\boldsymbol{q}), \mathbf{e}_y - \boldsymbol{q}\rangle. \tag{31}$$

For any $r' \in \Gamma_1(\boldsymbol{q})$, Bayes optimality gives $\underline{L}_1(\boldsymbol{q}) = \langle \mathbf{L}_{r'}, \boldsymbol{q}\rangle$. Moreover, for every $\mathbf{v} \in \mathbb{R}^{|\mathcal{Y}|}$,

$$\underline{L}_1(\mathbf{v}) \leq \langle \mathbf{L}_{r'}, \mathbf{v}\rangle = \underline{L}_1(\boldsymbol{q}) + \langle \mathbf{L}_{r'}, \mathbf{v} - \boldsymbol{q}\rangle.$$

Thus $\mathbf{L}_{r'}$ is a supergradient of the ambient Bayes risk at $\boldsymbol{q}$. Differentiability makes this supergradient unique, so

$$\mathbf{L}_{r'} = \nabla \underline{L}_1(\boldsymbol{q}), \qquad \ell_1^\Delta(\boldsymbol{q}, y) = \ell_1(r', y).$$

Consequently, all Bayes-optimal reports at $\boldsymbol{q}$ have the same loss for every outcome.

Concavity also gives

$$\mathbb{E}_{Y \sim \boldsymbol{p}}[\ell_1^\Delta(\boldsymbol{q}, Y)] = \underline{L}_1(\boldsymbol{q}) + \langle \nabla \underline{L}_1(\boldsymbol{q}), \boldsymbol{p} - \boldsymbol{q}\rangle \geq \underline{L}_1(\boldsymbol{p}),$$

with equality when $\boldsymbol{p} = \boldsymbol{q}$. Thus ℓ_1^Δ is a proper loss for distribution-valued reports on its differentiability domain, and

$$\rho_{\ell_1}(\boldsymbol{q}, r) = \mathbb{E}_{Y \sim \boldsymbol{q}}[\ell_1(r, Y) - \ell_1^\Delta(\boldsymbol{q}, Y)]. \tag{32}$$

Corollary D.4 (Deterministic improvement at differentiable projections)**.** *Under the assumptions of Theorem D.3, let h^* be a minimizer such that the ambient Bayes risk $\underline{L}_1$ is differentiable at $h^*(r)$ for every $r \in \mathcal{R}_1$. Any deterministic choice $f^*(r) \in \Gamma_1(h^*(r))$ attains the optimal worst-case gain δ_Γ. Equivalently, defining*

$$\operatorname{Gain}_\Delta(h^*; \widetilde{\mathcal{D}}) := \mathbb{E}_{\widetilde{\mathcal{D}}}[\ell_1(R_1, Y) - \ell_1^\Delta(h^*(R_1), Y)], \tag{33}$$

we have

$$\operatorname{Gain}_\Delta(h^*; \widetilde{\mathcal{D}}) \geq \delta_\Gamma \quad \textit{for every } \widetilde{\mathcal{D}} \in \mathfrak{D}_{\Gamma_0}, \qquad \min_{\widetilde{\mathcal{D}} \in \mathfrak{D}_{\Gamma_0}} \operatorname{Gain}_\Delta(h^*; \widetilde{\mathcal{D}}) = \delta_\Gamma.$$

Proof. Every report in $\Gamma_1(h^*(r))$ has loss vector $\nabla \underline{L}_1(h^*(r))$. The randomized rule supplied by Theorem D.3 is supported on this set. It therefore has the same loss as any deterministic choice $f^*(r)$ for every outcome in every report bin. The claimed guarantee and its exact worst-case value follow from that theorem and Equation (31).

In this case, the report-selection condition becomes

$$\mathbb{E}_{\mathcal{D}_{R_1}}\left[\langle \mathbf{L}_{R_1} - \nabla \underline{L}_1(h^*(R_1)), h(R_1) - h^*(R_1)\rangle\right] \geq 0 \quad \text{for every } h \in \mathcal{C}_\Gamma, \tag{34}$$

which is the first-order optimality condition for minimizing the convex projection objective

$$\Phi(h) = \mathbb{E}_{\mathcal{D}_{R_1}}\left[\langle \mathbf{L}_{R_1}, h(R_1)\rangle - \underline{L}_1(h(R_1))\right]$$

over $\mathcal{C}_\Gamma$. □

D.5 Confidence Reports

We now specialize the general construction to confidence reports, verifying the required regularity conditions and describing the resulting report-selection rule. The rule is deterministic when every projected distribution has a unique mode; otherwise, a suitable mixture over the tied modes attains the same optimal worst-case guarantee.

For the confidence loss in Equation (1), the report space is $\mathcal{A}_1 = \mathcal{Y} \times [0,1]$, and

$$\ell_1((a,c),y) = c^2 - 2c\,\mathbb{I}(y=a), \qquad \mathbf{L}_{(a,c)} = c^2\mathbf{1} - 2c\,\mathbf{e}_a.$$

Equip the finite answer set $\mathcal{Y}$ with the discrete topology and $[0,1]$ with its usual topology, and give $\mathcal{A}_1$ the product topology. This space is compact and Hausdorff. For each fixed outcome y and answer a, the loss is a polynomial in c and hence continuous on $\{a\} \times [0,1]$. Since these components are open in the product topology, the loss is continuous on the entire report space. Thus the compactness and report-continuity assumptions of Theorem D.3 hold.

For a fixed answer a, the expected loss is

$$\mathbb{E}_{Y\sim\boldsymbol{q}}[\ell_1((a,c),Y)] = c^2 - 2c\,\boldsymbol{q}(a) = (c-\boldsymbol{q}(a))^2 - \boldsymbol{q}(a)^2.$$

Minimizing over c gives $c = \boldsymbol{q}(a)$ and value $-\boldsymbol{q}(a)^2$. Minimizing over a therefore yields

$$\underline{L}_1(\boldsymbol{q}) = -\big(\max_{y\in\mathcal{Y}} \boldsymbol{q}(y)\big)^2, \qquad \Gamma_1(\boldsymbol{q}) = \left\{(a,c) : a \in \operatorname*{argmax}_y \boldsymbol{q}(y),\ c = \max_y \boldsymbol{q}(y)\right\}.$$

For this property, the level set of a report (a,c) is

$$\Gamma_1^{-1}(a,c) = \{\boldsymbol{q} \in \Delta(\mathcal{Y}) : \boldsymbol{q}(a) = c,\ \boldsymbol{q}(y) \le c \text{ for every } y \in \mathcal{Y}\}.$$

It is nonempty precisely when $c \in [1/|\mathcal{Y}|, 1]$. Thus the standing assumption that observed reports have nonempty level sets requires their confidence values to lie in this range.

The projection objective takes the explicit form

$$\Phi(h) = \sum_{r=(a,c)\in\mathcal{R}_1} \mu_r \left[c^2 - 2c\,h(r)(a) + \big(\max_y h(r)(y)\big)^2\right].$$

This objective is convex: the first two terms are affine in $h(r)$, and on the probability simplex the final term equals $\max_y h(r)(y)^2$, a maximum of convex functions.

Fix an optimal projection and write

$$\boldsymbol{q}_r^* := h^*(r), \qquad c_r^* := \max_y \boldsymbol{q}_r^*(y), \qquad M_r := \operatorname*{argmax}_y \boldsymbol{q}_r^*(y).$$

Under the standing assumptions of Theorem D.3, that theorem supplies probability vectors α_r supported on M_r such that the output (A, c_r^*), with $A \sim \alpha_r$, attains worst-case expected gain δ_Γ under the original confidence loss. Its expected loss vector is

$$\mathbf{z}_r = (c_r^*)^2\mathbf{1} - 2c_r^*\alpha_r.$$

The weights must satisfy Equation (30). Equivalently, they can be chosen by solving the following optimization with the projection held fixed:

$$\max_{\substack{\alpha_r\in\Delta(\mathcal{Y})\ \forall r\in\mathcal{R}_1 \\ \operatorname{supp}(\alpha_r)\subseteq M_r\ \forall r\in\mathcal{R}_1}} \ \min_{h\in\mathcal{C}_\Gamma} \sum_{r\in\mathcal{R}_1} \mu_r \left\langle \mathbf{L}_r - (c_r^*)^2\mathbf{1} + 2c_r^*\alpha_r, h(r)\right\rangle. \tag{35}$$

Its optimal value is δ_Γ.

If every M_r is a singleton, the report-selection rule is deterministic. The support constraint then uniquely determines each α_r, eliminating the report-selection optimization. To connect this case to Corollary D.4, consider a distribution $\boldsymbol{q}$ with unique mode a^* and maximum probability c. Its Bayes-optimal report (a^*, c) is unique. Since the loss-vector set is compact and the minimizing loss vector is unique, the ambient Bayes risk defined in the preceding subsection is differentiable at $\boldsymbol{q}$, with gradient

$$\nabla\underline{L}_1(\boldsymbol{q}) = c^2\mathbf{1} - 2c\,\mathbf{e}_{a^*}.$$

Consequently, Corollary D.4 applies, and

$$\ell_1^{\Delta}(\boldsymbol{q}, y) = c^2 - 2c\,\mathbb{I}(y = a^*) = \ell_1((a^*, c), y).$$

At a tied mode, distinct optimal answers have different loss vectors and hence give distinct supergradients of the ambient Bayes risk, which is therefore nondifferentiable at that distribution. The assumptions of Theorem D.3 still hold, so Equation (35) yields a suitable mixture over the tied answers, all with confidence c_r^*. Uniform averaging or a deterministic tie-breaking rule attains the same bound if its loss vectors satisfy Equation (30); membership in the set of Bayes-optimal reports alone is insufficient. The guarantee includes expectation over the output randomization and holds for every feasible ground-truth extension.

E Mutual Calibration, Common Priors, and No-Trade Theorem

In this section, we build the connection between our **mutual calibration** concept and the well-known economic literature on the No-Trade Theorem Milgrom and Stokey [1982], Arieli et al. [2021], Morris [2020]. We first rephrase the results in Arieli et al. [2021], Morris [2020] within our mutual-calibration context and then establish a quantitative relationship between the predictors' calibration errors and the arbitrage opportunities studied in that literature.

E.1 Common Priors and the No-Trade Characterization

Fix the observed joint report distribution $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$; as in the main text, index 0 denotes the reference predictor and index 1 the primary predictor. In Arieli et al. [2021], this joint distribution is called *feasible joint posterior beliefs* if there exists an extension $\widetilde{\mathcal{D}}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1,Y}$ with marginal $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$ such that Bayesian predictor i's posterior will coincide with $\boldsymbol{Q}_i$, i.e.

$$\mathbb{E}_{\widetilde{\mathcal{D}}}[\mathbf{Y} \mid \boldsymbol{Q}_i] = \boldsymbol{Q}_i, \qquad i \in \{0, 1\}. \tag{36}$$

This is exactly mutual calibration (Definition 3.1). Indeed, one may take $\boldsymbol{Q}_i$ itself as predictor i's signal; then Equation (36) says that the reported vector is the predictor's Bayesian posterior under the shared law $\widetilde{\mathcal{D}}$. Conversely, posterior rationality under a common prior gives the two calibration identities in Equation (36). Thus checking mutual calibration is the same feasibility problem as checking whether the reported posteriors admit a common-prior rationalization.

The classical No-Trade Theorem shows that a common prior rules out the arbitrage opportunity between predictors [Milgrom and Stokey, 1982]. The feasibility results of Arieli et al. [2021], Morris [2020] formalize the corresponding no-trade or money-pump characterization for a distribution of posterior reports. We next state that characterization in our notation and introduce a normalization that makes the size of an arbitrage opportunity finite.

Let $\mathcal{K}$ be the set of conditional kernels $K(\cdot \mid \boldsymbol{q}_0, \boldsymbol{q}_1) \in \Delta(\mathcal{Y})$, one for every $(\boldsymbol{q}_0, \boldsymbol{q}_1) \in \mathcal{Q}_0 \times \mathcal{Q}_1$. Each $K \in \mathcal{K}$ induces the candidate extension

$$\widetilde{\mathcal{D}}^K(\boldsymbol{q}_0, \boldsymbol{q}_1, y) := \mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}(\boldsymbol{q}_0, \boldsymbol{q}_1) K(y \mid \boldsymbol{q}_0, \boldsymbol{q}_1). \tag{37}$$

Only a kernel for which $\widetilde{\mathcal{D}}^K$ satisfies Equation (36) for both predictors witnesses mutual calibration.

For predictor $i \in \{0, 1\}$, a state-contingent payment $t_i : \mathcal{Q}_i \times \mathcal{Y} \to \mathbb{R}$ is a *zero-value trade* if its value under every reported belief is zero. We use the normalized class

$$\mathcal{T}_i := \left\{ t_i : \sum_{y \in \mathcal{Y}} \boldsymbol{q}_i(y) t_i(\boldsymbol{q}_i, y) = 0, \quad \|t_i(\boldsymbol{q}_i, \cdot)\|_\infty \le 1 \quad \text{for every } \boldsymbol{q}_i \in \mathcal{Q}_i \right\}. \tag{38}$$

Here $t_i(\boldsymbol{q}_i, y)$ is the payment from an external arbitrageur to predictor i. Hence each trade is weakly acceptable to the predictor, while the arbitrageur's realized payoff is $-\sum_{i=0}^{1} t_i(\boldsymbol{q}_i, y)$ when the reported pair is $(\boldsymbol{q}_0, \boldsymbol{q}_1)$ and the realized outcome is y.

Because the real outcome conditional kernel is unobserved, define the normalized robust arbitrage value achieved by the outsider as

$$\begin{aligned}\mathrm{Arb}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}) &:= \max_{t_0\in\mathcal{T}_0,\, t_1\in\mathcal{T}_1} \min_{K\in\mathcal{K}} \mathbb{E}_{\widetilde{\mathcal{D}}^K}\left[-\sum_{i=0}^{1} t_i(\boldsymbol{Q}_i, Y)\right] \\ &= -\min_{t_0\in\mathcal{T}_0,\, t_1\in\mathcal{T}_1} \mathbb{E}_{\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}}\left[\max_{y\in\mathcal{Y}} \sum_{i=0}^{1} t_i(\boldsymbol{Q}_i, y)\right]. \end{aligned} \tag{39}$$

For fixed trades, the inner minimization may choose an outcome separately at each report pair, which gives the second equality. The zero trades ($t_i(\boldsymbol{q}_i, y) \equiv 0$) are feasible, so the value is nonnegative. Moreover, Equation (39) is zero exactly when every normalized zero-value scheme satisfies

$$\mathbb{E}_{\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}}\left[\max_{y\in\mathcal{Y}} \sum_{i=0}^{1} t_i(\boldsymbol{Q}_i, y)\right] \geq 0. \tag{40}$$

This is the no-money-pump condition [Arieli et al., 2021, Morris, 2020] in mutual-calibration notation with normalization. The following example illustrates how a failure of mutual calibration creates a concrete arbitrage opportunity.

Example E.1 (Arbitrage in Contradictory Predictions). Consider the setting of Example 1.1, where the primary predictor outputs 10% (Group A) and 70% (Group B), and the reference predictor outputs a calibrated base rate of 50%. An arbitrageur offers the following contracts ($0 < \epsilon < 0.1$):

- **To the reference predictor:** pays $(0.5 + \epsilon)$ if the investment doubles; receives 0.5 if not.
- **To the primary predictor:** when predicting 0.7, receives 0.3 if the investment doubles and pays $(0.7 + \epsilon)$ otherwise; when predicting 0.1, receives 0.9 if the investment doubles and pays $(0.1 + \epsilon)$ otherwise.

Both predictors accept these contracts because their subjective expected profits are strictly positive. The arbitrageur's payoff is $0.4 - \epsilon$ in Group A and $-0.2 - \epsilon$ in Group B, independently of the realized outcome. Since the groups have equal probability, the expected payoff over the fixed report distribution $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$ is $0.1 - \epsilon > 0$ and holds uniformly over all possible outcome, indicating an arbitrage opportunity through the contradictory predictions.

Theorem E.2 (No-Trade Theorem in mutual-calibration form [Arieli et al., 2021, Morris, 2020])**.** *Under finite report and outcome supports, the following statements are equivalent:*

1. $\boldsymbol{Q}_0$ *and* $\boldsymbol{Q}_1$ *are mutually calibrated;*
2. $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$ *admits a common-prior rationalization; and*
3. $\mathrm{Arb}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}) = 0$.

Consequently, failure of mutual calibration is equivalent to a strictly positive normalized robust arbitrage value.

Proof. The equivalence of the first two statements follows from Equation (36). By Equation (39), the zero-trade scheme is feasible and

$$\mathrm{Arb}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}) = 0$$

if and only if the no-money-pump condition in Equation (40) holds for every normalized zero-value trade. Under finite supports, the normalization does not change this qualitative condition: any zero-value trade violating the condition can be multiplied by a positive constant so that it belongs to $\mathcal{T}_0 \times \mathcal{T}_1$, while preserving the sign of its expected payoff. The no-money-pump characterization of Arieli et al. [2021], Morris [2020] states that this condition holds if and only if the joint report distribution admits a common-prior rationalization. Combining these equivalences proves the theorem. □

E.2 A Quantitative Calibration–Arbitrage Relationship

The no-trade characterization records whether a common-prior extension exists. In this section, we give a quantitative condition, and build the connection with calibration errors.

For each $K \in \mathcal{K}$, define the posterior map on predictor i's original report bins by

$$\eta_i^K(\boldsymbol{q}_i) := \mathbb{E}_{\widetilde{\mathcal{D}}^K}[\mathbf{Y} \mid \boldsymbol{Q}_i = \boldsymbol{q}_i], \qquad i \in \{0,1\}, \tag{41}$$

Therefore, its predictor-wise and total calibration gaps can be represented by

$$\begin{aligned} \mathrm{CalGap}_i(K) &:= \mathbb{E}_{\widetilde{\mathcal{D}}^K}\left[\|\boldsymbol{Q}_i - \eta_i^K(\boldsymbol{Q}_i)\|_1\right], \qquad i \in \{0,1\}, \\ \mathrm{CalGap}(K) &:= \sum_{i=0}^{1} \mathrm{CalGap}_i(K). \end{aligned} \tag{42}$$

Definition E.3 (Minimal full-distribution calibration error (MinECE))**.** For the observed report distribution $\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}$, define

$$\begin{aligned} \mathrm{MinECE}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}) &:= \frac{1}{2} \min_{K \in \mathcal{K}} \mathrm{CalGap}(K) \\ &= \frac{1}{2} \min_{K \in \mathcal{K}} \sum_{i=0}^{1} \mathbb{E}_{\widetilde{\mathcal{D}}^K}\left[\left\|\boldsymbol{Q}_i - \mathbb{E}_{\widetilde{\mathcal{D}}^K}[\mathbf{Y} \mid \boldsymbol{Q}_i]\right\|_1\right]. \end{aligned} \tag{43}$$

The factor $1/2$ converts each ℓ_1 discrepancy into total variation. To relate this calibration distance to arbitrage, we use the following lemma to bridge ℓ_1 distance to maximal inner product.

Lemma E.4 (Centered-trade duality)**.** *For any reported distribution $\boldsymbol{q} \in \Delta(\mathcal{Y})$ and any outcome distribution $\boldsymbol{p} \in \Delta(\mathcal{Y})$, let*

$$\Phi(\boldsymbol{q}, \boldsymbol{p}) := \max_{\substack{\|\mathbf{v}\|_\infty \le 1 \\ \langle \mathbf{v}, \boldsymbol{q} \rangle = 0}} \langle \mathbf{v}, \boldsymbol{p} \rangle.$$

Then

$$\frac{1}{2}\|\boldsymbol{p} - \boldsymbol{q}\|_1 \le \Phi(\boldsymbol{q}, \boldsymbol{p}) \le \|\boldsymbol{p} - \boldsymbol{q}\|_1. \tag{44}$$

Proof. For every feasible $\mathbf{v}$, the centering constraint and Hölder's inequality give

$$\langle \mathbf{v}, \boldsymbol{p} \rangle = \langle \mathbf{v}, \boldsymbol{p} - \boldsymbol{q} \rangle \le |\langle \mathbf{v}, \boldsymbol{p} - \boldsymbol{q} \rangle| \le \|\mathbf{v}\|_\infty \|\boldsymbol{p} - \boldsymbol{q}\|_1 \le \|\boldsymbol{p} - \boldsymbol{q}\|_1,$$

which proves the upper bound. For the lower bound, set $\mathbf{u}(y) = \mathrm{sign}(\boldsymbol{p}(y) - \boldsymbol{q}(y))$, with $\mathrm{sign}(0) = 0$, and let $\alpha = \langle \mathbf{u}, \boldsymbol{q} \rangle$. Then

$$\mathbf{v}^* := \frac{\mathbf{u} - \alpha \mathbf{1}}{2}$$

satisfies $\|\mathbf{v}^*\|_\infty \le 1$ and $\langle \mathbf{v}^*, \boldsymbol{q} \rangle = 0$. Since both distributions have unit mass,

$$\langle \mathbf{v}^*, \boldsymbol{p} \rangle = \frac{1}{2}\langle \mathbf{u}, \boldsymbol{p} - \boldsymbol{q} \rangle = \frac{1}{2}\|\boldsymbol{p} - \boldsymbol{q}\|_1,$$

which proves the lower bound. □

For a fixed report $\boldsymbol{q}_i$, apply Lemma E.4 with $\boldsymbol{q} = \boldsymbol{q}_i$, $\boldsymbol{p} = \eta_i^K(\boldsymbol{q}_i)$, and $\mathbf{v}(y) = -t_i(\boldsymbol{q}_i, y)$. Trades may be selected independently across report bins and predictors, so averaging and summing gives the fixed-kernel bridge

$$\frac{1}{2}\,\mathrm{CalGap}(K) \le \mathrm{TradeVal}(K) \le \mathrm{CalGap}(K). \tag{45}$$

where

$$\mathrm{TradeVal}(K) := \max_{t_0 \in \mathcal{T}_0,\, t_1 \in \mathcal{T}_1} \mathbb{E}_{\widetilde{\mathcal{D}}^K}\left[-\sum_{i=0}^{1} t_i(\boldsymbol{Q}_i, Y)\right]. \tag{46}$$

is the maximal normalized payoff under this specific kernel K.

Proposition E.5 (Quantitative calibration–arbitrage relation)**.** *Under the finite report and outcome supports,*

$$\mathrm{MinECE}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}) \le \mathrm{Arb}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}) \le 2\,\mathrm{MinECE}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}). \tag{47}$$

Hence the arbitrage value and the minimal calibration error vanish simultaneously, precisely under mutual calibration.

Proof. Under finite supports, $\mathcal{T}_0 \times \mathcal{T}_1$ and $\mathcal{K}$ are compact convex sets, and the payoff is bilinear. The finite-dimensional minimax theorem [Sion, 1958] therefore yields

$$\mathrm{Arb}(\mathcal{D}_{\boldsymbol{Q}_0,\boldsymbol{Q}_1}) = \min_{K\in\mathcal{K}} \mathrm{TradeVal}(K). \tag{48}$$

Minimizing Equation (45) over $K \in \mathcal{K}$ and using Equations (43) and (48) gives

$$\frac{1}{2} \min_{K\in\mathcal{K}} \mathrm{CalGap}(K) \le \min_{K\in\mathcal{K}} \mathrm{TradeVal}(K) \le \min_{K\in\mathcal{K}} \mathrm{CalGap}(K),$$

which is exactly Equation (47). The final equivalence follows from Theorem E.2. □

Remark E.6. MinECE here is a population, full-distribution calibration quantity, not the binned top-label or confidence ECE commonly reported in experiments. It evaluates the complete multiclass probability vector on each original report bin and measures the discrepancy in total variation, matching the stronger *multi-class calibration* discussed by Kull et al. [2019].

F Experiment Details

F.1 Dataset Details

We evaluate our method on three established benchmarks. All three are formulated as multiple-choice tasks, though they differ in the number of answer options:

- **MMLU [Hendrycks et al., 2021]** (MIT License). This dataset is a **4-option** multiple-choice benchmark containing questions from various subjects. We evaluate our method on the *test* split, which contains 14,042 samples. To estimate the joint distribution $\mathcal{D}_{R_0,R_1}$ or tune parameters, we use the *validation* split, which contains 1,531 samples.
- **MMLU-Redux [Gema et al., 2025]** (CC BY 4.0). This dataset is a **4-option** multiple-choice benchmark containing a manually curated subset of **MMLU**, designed to remove annotation errors and ambiguous questions. We utilize the *test* split for evaluation, which contains 5,700 samples. Since MMLU-Redux lacks a distinct validation set, we employ the *validation split of the original MMLU dataset* as the validation set to estimate the joint distribution $\mathcal{D}_{R_0,R_1}$ for our method or tune parameters for supervised baselines, which contains 1,531 samples.
- **CommonSenseQA (CSQA) [Talmor et al., 2019]** (MIT License). This dataset features **5-option** multiple-choice questions requiring various types of commonsense knowledge. We evaluate our method on the *validation* split, which contains 1,221 samples, as the labels for the test split are unavailable. To estimate the joint distribution $\mathcal{D}_{R_0,R_1}$ or tune parameters, we use the *training* split, which contains 9,741 samples.

Asset Sources and Licenses. Table 2 lists the datasets and model checkpoints used in our experiments, together with their repository identifiers, original sources, and licenses. The `LLaMA-3.1-8B` checkpoints are also subject to the accompanying Acceptable Use Policy, as specified in their linked model cards.

The implementation uses PyTorch 2.11.0 (BSD 3-Clause) and vLLM 0.26.0 (Apache 2.0). Convex optimization uses CVXPY 1.9.3 and Clarabel 0.11.1, both under Apache 2.0. Other principal dependencies are NumPy 2.3.5, SciPy 1.17.1, and scikit-learn 1.9.1 (BSD 3-Clause), together with Transformers 5.13.0, Datasets 5.0.1, and mistral-common 1.11.7 (Apache 2.0). The adapted SmoothECE implementation is derived from Apple's `ml-calibration` and retains its upstream Apple license. The corresponding license notice is included in our public repository.

F.2 Implementation Details

Code and Data Availability. The implementation and reproduction instructions are available at `https://github.com/smyft/calibration-without-ground-truth`. The datasets and model checkpoints are obtained from their original sources, subject to their respective licenses and terms of use.

Table 2: Repository identifiers, sources, and licenses of existing experimental assets. Each model row identifies the checkpoint used in our experiments.

Asset	Repository identifier and source	License
MMLU [Hendrycks et al., 2021]	`cais/mmlu`	MIT
MMLU-Redux 2.0 [Gema et al., 2025]	`edinburgh-dawg/mmlu-redux-2.0`	CC BY 4.0
CommonsenseQA [Talmor et al., 2019]	`tau/commonsense_qa`	MIT
Qwen3 reference [Yang et al., 2025]	`Qwen/Qwen3-8B-Base`	Apache 2.0
Qwen3 primary [Yang et al., 2025]	`Qwen/Qwen3-8B`	Apache 2.0
Llama reference [Grattafiori et al., 2024]	`meta-llama/Llama-3.1-8B`	Llama 3.1 Community License
Llama primary [Grattafiori et al., 2024]	`meta-llama/Llama-3.1-8B-Instruct`	Llama 3.1 Community License
Ministral reference [Liu et al., 2026]	`mistralai/Ministral-3-8B-Base-2512`	Apache 2.0
Ministral primary [Liu et al., 2026]	`mistralai/Ministral-3-8B-Instruct-2512-BF16`	Apache 2.0

Reproduction Configuration. The experiments use Python 3.11.15, PyTorch `2.11.0+cu130` (the CUDA 13.0 build), and vLLM 0.26.0, together with the dependency versions listed above. The dataset and model repository identifiers, sources, and licenses are listed in Table 2. The exact configuration and execution commands are provided in `https://github.com/smyft/calibration-without-ground-truth`.

Compute Resources. Experiments were run on AutoDL. LLM inference used 1 RTX 6000D with 84 GB of device memory. The host system provided 18 vCPU Intel(R) Xeon(R) Platinum 8470Q and 110 GB of RAM.

As shown in Appendix F.2.1, we use six LLM checkpoints, comprising three Base/Instruct pairs. Base inference uses one stage, whereas Instruct inference uses a reasoning stage followed by an answer-extraction stage. The inference stages used to produce the final predictions took approximately 6 GPU hours. After inference, fitting and deploying the calibration map required only a few seconds on the above CPU configuration.

F.2.1 Prompting Strategy and Confidence Extraction

We describe the prompting strategies and the subsequent procedure for extracting confidence scores.

Prompting Strategy. Following previous studies [Hendrycks et al., 2021, Wang et al., 2024], we use standard zero-shot prompts to directly elicit the answer from the first token log probabilities. This works well for **Base** models, but **Instruct** models typically generate reasoning contents before the final answer, making direct next-token probability extraction unreliable. To employ their CoT capabilities [Wei et al., 2022], we adopt a two-step approach widely used in instructing LLMs to verbalize confidence scores [Lin et al., 2022, Tian et al., 2023]. We first prompt the model to generate a reasoning process using the standard zero-shot prompts. The generated reasoning process is filtered before being used in the second stage. Subsequently, we append this reasoning process along with an answer-triggering phrase to extract the next-token probability. Specific prompt templates are detailed in Figure 2 and Figure 3.

Standard zero-shot prompts for *base* model

```
The following is a multiple-choice question about {subject}.

Question:  {task_query}
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
E. {choiceE} % only for CommonSenseQA

Answer:
```

Figure 2: Standard zero-shot prompt template utilized for *Base* models to directly elicit the first token probability following the question. The `{subject}` placeholder is instantiated with the specific domain for MMLU/MMLU-Redux or `common sense` for CommonsenseQA, and `{task_query}`,`{choice}` placeholders are instantiated with the question and choices of the specific sample.

Confidence Extraction. After extracting the log-probabilities from LLMs, we construct the (answer, confidence) pair through the following renormalization process. Let $\mathcal{Y}$ denote the set of valid answer options. We construct an option log-score $\mathcal{P}_{\mathcal{M}}(y \mid \mathbf{x})$ from the returned top-20 next-token log-probabilities and normalize these scores over $\mathcal{Y}$:

$$\widehat{\mathcal{P}}_{\mathcal{M}}(y|\mathbf{x}) = \frac{\exp(\mathcal{P}_{\mathcal{M}}(y|\mathbf{x}))}{\sum_{y' \in \mathcal{Y}} \exp(\mathcal{P}_{\mathcal{M}}(y'|\mathbf{x}))},$$

The answer token with the highest re-normalized probability is selected as the predicted answer, with $\widehat{\mathcal{P}}_{\mathcal{M}}(y|\mathbf{x})$ serving as the associated confidence score.

F.2.2 Joint Distribution Estimation

To compute the recalibration map, we approximate the joint distribution $\mathcal{D}_{R_0,R_1}$ using the empirical distribution of paired predictor reports from the validation dataset. Let the report from a predictor be denoted as a pair $r = (\widehat{y}, \widehat{c})$, consisting of an answer $\widehat{y} \in \mathcal{Y}$ and a continuous confidence score $\widehat{c} \in [0, 1]$.

Given the continuous nature of confidence scores, direct estimation of the joint distribution on finite samples is theoretically sparse. Therefore, we utilize a discretization process with a step size $s \in (0, 1]$ with $1/s \in \mathbb{N}$. We define a quantization function $\psi_s(c) = s \left\lfloor \frac{c}{s} + \frac{1}{2} \right\rfloor$ that maps a continuous confidence score to the nearest value on the discrete grid $\mathcal{C}_s = \{0, s, 2s, \ldots, 1\}$. Specifically, for an observed report $r_i = (\widehat{y}_i, \widehat{c}_i)$, the discretized report is given by $\widetilde{r}_i = (\widehat{y}_i, \psi_s(\widehat{c}_i))$. This discretization applies only to the input reports, but does not constrain the coordinates of the projected outcome distributions to grid values.

Let $\{(\widetilde{r}_{0i}, \widetilde{r}_{1i})\}_{i=1}^N$ denote the set of discretized report pairs generated by the *Base* and *Instruct* models for N input samples in the validation dataset. The estimated joint distribution is then formulated as the normalized frequency of these discretized pairs:

$$\widehat{\mathcal{D}}_{R_0,R_1} = \frac{1}{N} \sum_{i=1}^{N} \delta_{(\widetilde{r}_{0i}, \widetilde{r}_{1i})},$$

where $\delta_{(\cdot)}$ is the Dirac delta mass. In our main experiments, we set $s = 0.01$, ensuring a fine-grained yet computationally tractable grid for probability estimation. We also report results under discretization step $s \in \{0.1, 0.05, 0.02, 0.01, 0.005, 0.002\}$ in Appendix G.4 to check the sensitivity of our method to the confidence discretization step s.

Two-step zero-shot CoT prompts for *instruct* model

```
Step 1: Reasoning Generation

System Prompts:
The following is a multiple-choice question about {subject}.
---
User Prompt:
Question: {task_query}
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
E. {choiceE} % only for CommonSenseQA

Model Response: {reasoning_process}

Step 2: Answer Extraction

System Prompts:
The following is a multiple-choice question about {subject}.
---
User Prompt:
Question: {task_query}
A. {choiceA}
B. {choiceB}
C. {choiceC}
D. {choiceD}
E. {choiceE} % only for CommonSenseQA

---
Assistant Prompt:
{filtered_reasoning}
So the final answer is (answer with an option letter):
```

Figure 3: Two-step zero-shot CoT prompt templates designed for *Instruct* models. We first prompt the model to generate a reasoning chain (Step 1), and then filter the reasoning process and subsequently append an answer-triggering phrase to extract the next-token probability (Step 2). The `{subject}` placeholder is instantiated with the specific domain for MMLU/MMLU-Redux or `common sense` for CommonsenseQA, and `{task_query}`,`{choice}` placeholders are instantiated with the question and choices of the specific sample.

F.2.3 Algorithm Details

In practice, we first solve the projection in Equation (23) using the empirical distribution of discretized reports, restricting h to the observed primary-report support. For Confidence Loss, the finite-support optimization yields a second-order cone formulation with linear reference-compatibility constraints. The resulting finite-dimensional convex program is solved using Clarabel 0.11.1 through CVXPY 1.9.3 in double precision. We use the default solver settings, including a maximum of 200 iterations and target tolerances of 10^{-8} for feasibility and absolute and relative duality gaps. The default reduced-accuracy termination criteria are also retained. Across the nine default-grid settings, four runs were reported as `optimal` and five as `optimal_inaccurate`. All returned primal solutions, with no solver errors or infeasibility reports. In the 54-setting discretization study, the corresponding counts were 23 and 31. At the default discretization step $s = 0.01$, the CVXPY solve calls took 0.10–5.51 seconds per (dataset, model) setting on the CPU.

Then, we convert each computed projection vector into an (answer, confidence) report by selecting the first maximizing class and its probability. If every exact projected distribution has a unique mode, the report-selection problem in Equation (35) reduces to this deterministic choice, requiring no

additional mixture optimization or sampling. Our implementation always uses this first-maximizer rule, without optimizing mixtures over tied modes. Across the normalized projection vectors from the nine default discretization step $s = 0.01$ settings and 54 discretization sensitivity analysis settings, no two largest entries were exactly equal in floating-point arithmetic. The smallest top-two gaps before output-confidence rounding were 3.17×10^{-8} and 1.55×10^{-15}, respectively.

At deployment, the learned map is applied to the primary reports without another reference-model query. For an unseen discretized report, we set the transformation to the identity map. Specifically, let $r = (a, c)$ be a primary report and $\widetilde{r} = (a, \psi_s(c))$ its discretized version. Let $\mathcal{S}$ denote the observed support of discretized primary reports, and let F_s denote the learned report-valued transformation on this support. At deployment, we use

$$\overline{F}_s(\widetilde{r}) = \begin{cases} F_s(\widetilde{r}), & \widetilde{r} \in \mathcal{S}, \\ \widetilde{r}, & \widetilde{r} \notin \mathcal{S}. \end{cases}$$

Support membership requires an exact match of both the predicted answer and the discretized confidence. For an unseen discretized report, both components are retained unchanged. For mapped reports, answer selection precedes output rounding. Only the confidence component of the final report is rounded to four decimal places, independently of s. At our default $s = 0.01$, such out-of-support predictions account for at most 4–6% of test instances on MMLU-Redux and under 1% on CommonSenseQA, confirming that the learned recalibration map covers the vast majority of test-time inputs.

F.3 Baseline Details

We compare our method against three categories of baselines that differ in their information requirements and deployment costs: (i) a **dummy reference** that provides no useful information, (ii) **supervised** calibration methods that require a labeled validation set, and (iii) **label-free** baselines that, like our method, require no labels but instead leverage a reference model. Within the label-free category, we further distinguish between *test-time* heuristics that query the reference model on every test input, and *validation-time* methods that, like our own, use the reference model only to learn a recalibration mapping on a held-out calibration split and require no access to the reference model at test time. All **supervised** and **label-free** baselines preserve the original predicted answer $\widehat{y}$ and only modify the associated confidence $\widehat{c}$, producing a calibrated score $\widetilde{c}$. In contrast, our method re-estimates the whole probability vector, providing opportunities to update the final answer.

Dummy Reference. As a sanity check, we create a **Dummy Reference** by randomizing the option order and replacing the reference model with a uniform distribution over answer positions. This creates a calibrated reference while throwing away all informative signals from the reference. Thus it isolates whether observed gains stem from useful reference information rather than from generic conservative shrinkage of confidences.

Supervised Baselines. Although these baselines are originally developed for full probability distributions over the outcome space $\mathcal{Y}$, we adapt them to our (answer, confidence) setting by formulating calibration as a binary classification task. Specifically, we define the ground truth for the i-th sample as $z_i = \mathbb{I}(\widehat{y}_i = y_i) \in \{0, 1\}$, indicating whether the predicted answer $\widehat{y}_i$ matches the true label y_i. Consequently, the calibration dataset is constructed as pairs of initial confidence and correctness indicators: $\{(\widehat{c}_i, z_i)\}_{i=1}^{N_{\text{val}}}$.

- **Temperature Scaling (TS)** [Guo et al., 2017]. Originally designed for multi-class logits, TS is adapted to binary calibration by scaling the log-odds of the confidence, which can also be viewed as a special case of **Platt Scaling** [Platt, 2000] with no bias. It optimizes a scalar temperature $T > 0$ to minimize the Negative Log Likelihood (NLL) on the calibration set:

$$\min_T -\sum_{i=1}^{N_{\text{val}}} \Big[z_i \log(\sigma(\phi_i/T)) + (1 - z_i) \log(1 - \sigma(\phi_i/T)) \Big],$$

  where $\phi_i = \text{logit}(\widehat{c}_i) = \log(\frac{\widehat{c}_i}{1-\widehat{c}_i})$ denotes the uncalibrated log-odds, and $\sigma(\cdot)$ is the sigmoid function. The resulting calibrated confidence score is given by:

$$\widetilde{c} = \sigma\left(\frac{1}{T} \log\left(\frac{\widehat{c}}{1-\widehat{c}}\right)\right).$$

- **Histogram Binning (HB)** [Zadrozny and Elkan, 2001]. This non-parametric method partitions the confidence interval $[0, 1]$ into B disjoint bins defined by boundaries $0 = a_1 < \cdots < a_{B+1} = 1$. For any sample falling into bin $B_m = (a_m, a_{m+1}]$, the calibrated score θ_m is estimated as the average empirical accuracy (i.e., the mean of z_i) within that bin. The calibration map can be viewed as minimizing the bin-wise squared loss:

$$\min_{\theta_1,\ldots,\theta_B} \sum_{m=1}^{B} \sum_{i=1}^{N_{\text{val}}} \mathbb{I}(\widehat{c}_i \in B_m)(\theta_m - z_i)^2,$$

  assigning $\widetilde{c} = \theta_m$ wherever $\widehat{c} \in B_m$. An empty bin uses the midpoint $(a_m + a_{m+1})/2$ and the bin index is clipped to the valid range at deployment. We use a uniform partition with $B = 10$ bins, and assign confidence zero to the first bin.

- **Isotonic Regression (IR)** [Zadrozny and Elkan, 2002]. IR minimizes squared error subject to non-decreasing fitted values, as specified below. Our implementation uses isotonic least-squares fitting, linearly interpolates between the fitted thresholds at prediction time, and clips predictions outside the fitted confidence range to the corresponding endpoint value. Specifically, we can write the optimization problem as:

$$\min_{f} \sum_{i=1}^{N_{\text{val}}} (f(\widehat{c}_i) - z_i)^2$$
$$\text{s.t.} \quad \forall i, j : \widehat{c}_i \le \widehat{c}_j \implies f(\widehat{c}_i) \le f(\widehat{c}_j).$$

  The calibrated confidence is obtained via $\widetilde{c} = f(\widehat{c})$.

- **Gaussian Process Calibration (GPcalib)** [Wenger et al., 2020]. We use a binary Gaussian-process calibration baseline that maps the scalar confidence c to the correctness indicator $z = \mathbb{I}(\widehat{y} = y)$. Its predictive probability $\Pr(z = 1 \mid c)$ becomes the calibrated confidence, while the original predicted answer is retained. This binary implementation adapts the GP-based calibration approach to the (answer, confidence) setting. On CommonsenseQA, GPcalib uses a subset of 1,500 examples to balance performance and time complexity.

- **PAC-Bayes Recalibration (PBR)** [Fujisawa and Futami, 2025]. We apply the recalibration objective in Eq. (8) of Fujisawa and Futami [2025] to the binary correctness-calibration task defined above. The posterior-predictive mean provides the calibrated confidence, while the original predicted answer is retained. We select the regularization strength by five-fold cross-validation on the calibration set using held-out Brier score, and then refit on the full calibration set.

Label-Free Baselines. These baselines, like our method, require no validation labels but instead exploit a reference model. They differ, however, in *when* the reference model is queried. The first two heuristics (BCT and ConfAvg) query the reference model on *every test input*, incurring the cost of an additional forward pass at inference time. The third baseline (DACA), like our own method, uses the reference model only on the validation split to fit a recalibration mapping and requires no access to the reference model at test time.

- **Best-Confidence Transfer (BCT).** BCT retains the primary model's predicted answer and replaces its confidence with the probability assigned to that answer by the reference model. This baseline requires access to the probability assigned by the reference model to the primary model's chosen answer and a reference-model query for each evaluation input.

- **Confidence Averaging (ConfAvg).** Keep the primary model's predicted answer but use the average value of the primary and reference confidences. Requires a reference query at test time.

- **Disagreement-Aware Confidence Alignment (DACA)** [Luo et al., 2025]. DACA performs unsupervised temperature scaling by minimizing the KL divergence between the reference and primary models on the validation split, restricted to inputs where their final answers agree. We similarly adapt it to our (answer, confidence) setting by formulating calibration as a binary classification task. Specifically, DACA fits the temperature τ through:

$$\min_{\tau>0} \sum_{i=1}^{N_{\text{val}}} \mathbb{I}(\widehat{y}_{ref,i} = \widehat{y}_{pri,i}) \cdot \Big[\widehat{c}_{ref,i} \log \frac{\widehat{c}_{ref,i}}{\sigma(\phi_{pri,i}/\tau)} + (1 - \widehat{c}_{ref,i}) \log \frac{1 - \widehat{c}_{ref,i}}{1 - \sigma(\phi_{pri,i}/\tau)}\Big],$$

where $\phi_{pri,i} = \text{logit}(\widehat{c}_{pri,i}) = \log(\frac{\widehat{c}_{pri,i}}{1-\widehat{c}_{pri,i}})$ denotes the uncalibrated log-odds, and $\sigma(\cdot)$ is the sigmoid function. The resulting calibrated confidence score is given by:

$$\widetilde{c} = \sigma\left(\frac{1}{\tau}\log\left(\frac{\widehat{c}}{1-\widehat{c}}\right)\right).$$

F.4 Evaluation Metrics

Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ denote the evaluation set. For each sample $\boldsymbol{x}_i$, the model predicts an answer $\widehat{y}_i$ with an associated confidence $\widehat{c}_i \in [0,1]$. We evaluate the model's performance using the following three metrics in our main experiments:

- **Prediction Accuracy (Acc).** This metric measures the correctness of the LLM's predictions:

$$\text{Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\widehat{y}_i = y_i),$$

  where $\mathbb{I}(\cdot)$ is the indicator function.

- **Calibration Metrics.** We assess calibration performance using **Expected Calibration Error (ECE)** [Guo et al., 2017] and the **Brier Score (BS)** [BRIER, 1950]. Brier Score measures the mean squared error between the confidence and the prediction correctness:

$$\text{BS} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{c}_i - \mathbb{I}(\widehat{y}_i = y_i))^2.$$

  For ECE, we partition $[0,1]$ into $M = 10$ uniform-width bins. Writing $z_i = \mathbb{I}(\widehat{y}_i = y_i)$, the first bin contains $0 \le \widehat{c}_i \le 1/M$, and bin $m \ge 2$ contains $(m-1)/M < \widehat{c}_i \le m/M$. Let B_m denote the set of sample indices in bin m. We compute

$$\text{ECE} = \sum_{m:|B_m|>0}\frac{|B_m|}{N}\left|\frac{1}{|B_m|}\sum_{i\in B_m} z_i - \frac{1}{|B_m|}\sum_{i\in B_m}\widehat{c}_i\right|.$$

  Empty bins contribute zero.

- **Optimization Objective:** We also report the **Confidence Loss (CL)** as defined in Equation 1, which serves as our primary optimization target:

$$\text{CL} = \frac{1}{N}\sum_{i=1}^{N}\left(-\mathbb{I}(\widehat{y}_i = y_i) + (\widehat{c}_i - \mathbb{I}(\widehat{y}_i = y_i))^2\right).$$

Beyond the standard ECE with 10 uniform-width bins used in our main experiments, we evaluate alternative binned and smooth calibration-error estimators.

For binned ECE, we consider uniform-width binning (UWB) and uniform-mass binning (UMB). UWB uses equal-width intervals with the boundary convention defined above. UMB sorts examples by confidence and partitions the ordered examples into groups whose sizes differ by at most one. We evaluate $M = 10$, $M = 15$, and $M^* = \max\{1, \lfloor n^{1/3}\rfloor\}$, where n is the number of evaluated examples, following the sample-size scaling discussed by Futami and Fujisawa [2024].

For the non-binned check, we compute automatic-bandwidth SmoothECE [Błasiok and Nakkiran, 2024] following the official repo[3] with a reflected Gaussian kernel and no logit scaling. Within each dataset–model setting, all methods use the common sample-ID intersection of the full comparison set, and correctness is recomputed from each method's final answer.

Unless otherwise stated, the $\pm$ entries in the metric tables are normal-approximation 95% interval half-widths, $1.96\,\widehat{\text{SE}}$. For accuracy, we use $\widehat{\text{SE}} = \sqrt{\widehat{a}(1-\widehat{a})/n}$. For Brier Score and Confidence Loss, we use the sample standard deviation of the per-example losses divided by $\sqrt{n}$. The Confidence Loss calculation uses the complete per-example quantity $(\widehat{c}_i - z_i)^2 - z_i$.

[3] `https://github.com/apple/ml-calibration`

For binned ECE, let n_m be the number of examples in bin m and let s_m^2 be the sample variance of $z_i - \widehat{c}_i$ within that bin. We use the within-bin approximation

$$\widehat{\mathrm{SE}}_{\mathrm{ECE}} = \frac{1}{n}\sqrt{\sum_{m:n_m \geq 2} n_m s_m^2}.$$

Bins with fewer than two examples contribute zero to this variance estimate. This approximation conditions on the empirical bin memberships and counts, including the empirical partition for UMB.

For SmoothECE, we report 95% percentile intervals from 1000 paired bootstrap replicates. For each (dataset,model) setting, every replicate draws n examples with replacement from the matched evaluation set, where n is the matched sample count reported in the table. All methods use the same sampled example indices. Within each replicate, the SmoothECE bandwidth is selected anew for each method, while the fitted calibration maps remain fixed. The interval endpoints are the 2.5th and 97.5th percentiles of the bootstrap estimates. The reported point estimate is computed using all matched examples; percentile intervals need not contain this estimate.

G Additional Results

G.1 Few-Label Reference Validation

Our framework assumes access to an (approximately) calibrated reference predictor, which is typically identified via prior knowledge of the model family. When such prior knowledge is unavailable, the better-calibrated model can instead be selected empirically using a small labeled diagnostic set. To assess the practicality of this approach, we quantify the minimum number of samples N^* required to empirically verify that one model is better calibrated than the other. This is a diagnostic analysis on fixed evaluation splits: MMLU test, MMLU-Redux test, and CommonsenseQA validation. These labels are not used to fit our recalibration map.

For each tested sample size

$$N \in \{10, 20, \ldots, 200, 250, 300, 400, 500, 750, 1000\},$$

we independently draw 1000 subsets of size N without replacement from the corresponding evaluation split. On each subset S_r, we recompute 10-bin uniform-width ECE for both models and form

$$\Delta_r(N) = \mathrm{ECE}_{S_r}(\mathrm{Instruct}) - \mathrm{ECE}_{S_r}(\mathrm{Base}).$$

We define N^* as the smallest tested N for which the empirical 2.5th percentile of $\{\Delta_r(N)\}_{r=1}^{1000}$ is positive. We repeat this procedure using 1000 Monte Carlo seeds and report the mean and standard deviation of N^*. Results are shown in Table 3. The mean N^* is below 200 in eight of the nine settings. For `Ministral-3-8B` on CommonsenseQA, all 1000 seeds yield $N^* = 750$ on the tested grid. This is only about 7.7% of the full calibration set of $9,741$ examples.

Table 3: **Minimum number of labeled samples N^* needed to verify that the Base model has lower ECE than their Instruct counterpart.** N^* is the smallest tested sample size for which the empirical 2.5th percentile of the Instruct-minus-Base ECE gap over 1000 independently sampled subsets is positive. Entries report mean $\pm$ standard deviation across 1000 Monte Carlo seeds on fixed evaluation splits. These labels are used only for diagnostic evaluation, not for fitting the recalibration map.

$N^*(\downarrow)$	MMLU	MMLU-Redux	CommonSenseQA
Qwen3-8B	110.1±6.3	159.4±8.3	196.9±20.4
LLaMA-3.1-8B	48.1±3.9	49.2±2.8	77.9±4.4
Ministral-3-8B	74.9±5.2	125.9±7.3	750.0±0.0

G.2 Comparison with Additional Supervised Baselines

Tables 4 and 5 report the full results on **MMLU-Redux** and **CommonSenseQA** respectively, including comparisons against an extended set of baselines (See Appendix F.3). The trends observed in Section 5 are consistent across this broader set of comparisons.

Specifically, among label-free methods that do not require Base model access at test time (except Base itself), our approach attains the best ECE for `Qwen3-8B` on MMLU-Redux and `Ministral-3-8B` on both datasets, while simultaneously preserving or marginally improving the predictive accuracy of the Instruct model. Within this group, our method achieves the best Confidence Loss (CL) for `Ministral-3-8B`, confirming that the gains of our method arise from optimization of the proper loss. The *Dummy Reference* yields much smaller reductions in ECE, supporting the usefulness of information from the Base reference. Consistent with the failure case discussed in the main paper, `LLaMA-3.1-8B` remains the most challenging setting for our method.

G.3 Main Results on MMLU

Table 6 presents the full experimental results on the **MMLU** benchmark. We observe trends similar to the findings on MMLU-Redux and CommonSenseQA reported in Section 5.

G.4 Discretization Sensitivity

When estimating the joint distribution $\mathcal{D}_{R_0,R_1}$ (Appendix F.2), we use a discretization step s to map continuous input confidence scores to the nearest value on a discrete grid. Table 7 reports the sensitivity of our method to s. Our method is robust across choices of s, consistently reducing ECE relative to the Instruct model (primary model). `Qwen3-8B` and `Ministral-3-8B` exhibit a U-shaped dependence on s on MMLU and MMLU-Redux, whereas all three families attain their lowest tested ECE at $s = 0.002$ on CommonsenseQA. CommonsenseQA has a larger calibration set of $9,741$ examples, compared with $1,531$ for the other two benchmarks, providing more observations for estimating the joint distribution on finer grids. We retain $s = 0.01$ as the common default rather than selecting a separate grid for each evaluation setting.

G.5 Robustness to Alternative Calibration-Error Estimators

Tables 8 and 9 evaluate robustness across alternative binned ECE measures and automatic-bandwidth SmoothECE (Appendix F.4). Our method improves over the Instruct model under every reported binning scheme and achieves lower SmoothECE in all nine (dataset,model) settings, indicating that the improvement is robust to the choice of calibration-error estimator.

Table 4: **Main results on MMLU-Redux with more baselines and error bars. Bold** and underline denote the **best** and second-best results among *label-free* methods that *do not* require access to Base Model during test time, while †indicates the *global* best across all methods. Accuracy for some Baselines is marked with – as these post-hoc methods do not alter final predictions.

Methods	**Label?**	**Base Access?**	**ACC** (↑)	**BS** (↓)	**ECE** (↓)	**CL** (↓)
			Qwen3-8B			
Base Model (Q_0)	✗	✗	77.19±1.09%	0.1227±0.0043	**0.0305±0.0091**	−0.6493±0.0140
Instruct Model (Q_1)	✗	✗	82.53±0.99%	0.1362±0.0079	0.1269±0.0089	−0.6891±0.0174
Temperature Scaling	✓	✗	–	0.1137±0.0051	0.0426±0.0086	−0.7116±0.0145
Histogram Binning	✓	✗	–	0.1198±0.0059	0.0208±0.0090	−0.7054±0.0150
Isotonic Regression	✓	✗	–	0.1108±0.0054†	0.0189±0.0086†	−0.7145±0.0144†
GPcalib	✓	✗	–	0.1132±0.0056	0.0470±0.0086	−0.7120±0.0147
PBR	✓	✗	–	0.1146±0.0061	0.0565±0.0087	−0.7107±0.0152
BCT	✗	✓	–	0.1407±0.0055	0.1227±0.0088	−0.6845±0.0122
ConfAvg	✗	✓	–	0.1118±0.0053	0.0346±0.0086	−0.7135±0.0148
DACA	✗	✗	–	**0.1225±0.0041**	0.0985±0.0086	**−0.7028±0.0135**
Dummy Reference	✗	✗	82.53±0.99%	0.1336±0.0077	0.1145±0.0089	−0.6916±0.0172
Ours	✗	✗	**82.56±0.99%**†	0.1236±0.0062	0.0348±0.0089	−0.7020±0.0158
			LLaMA-3.1-8B			
Base Model (Q_0)	✗	✗	65.35±1.24%	**0.1574±0.0045**†	**0.0208±0.0103**	−0.4961±0.0147
Instruct Model (Q_1)	✗	✗	70.91±1.18%	0.2550±0.0107	0.2526±0.0113	−0.4542±0.0222
Temperature Scaling	✓	✗	–	0.1873±0.0053	0.0364±0.0111	−0.5218±0.0166
Histogram Binning	✓	✗	–	0.1927±0.0056	0.0118±0.0114†	−0.5165±0.0167
Isotonic Regression	✓	✗	–	0.1836±0.0056	0.0181±0.0111	−0.5255±0.0164
GPcalib	✓	✗	–	0.1878±0.0055	0.0126±0.0112	−0.5214±0.0166
PBR	✓	✗	–	0.2142±0.0087	0.1657±0.0112	−0.4950±0.0192
BCT	✗	✓	–	0.1783±0.0057	0.1061±0.0104	−0.5308±0.0133
ConfAvg	✗	✓	–	0.1758±0.0062	0.1011±0.0104	−0.5333±0.0176†
DACA	✗	✗	–	0.1878±0.0048	0.0453±0.0111	**−0.5213±0.0162**
Dummy Reference	✗	✗	**70.92±1.18%**†	0.2516±0.0105	0.2454±0.0113	−0.4577±0.0220
Ours	✗	✗	70.91±1.18%	0.2115±0.0080	0.1340±0.0114	−0.4976±0.0196
			Ministral-3-8B			
Base Model (Q_0)	✗	✗	75.72±1.11%	**0.1281±0.0041**†	0.0534±0.0091	−0.6291±0.0135
Instruct Model (Q_1)	✗	✗	80.30±1.03%	0.1726±0.0093	0.1656±0.0099	−0.6304±0.0194
Temperature Scaling	✓	✗	–	0.1476±0.0054	0.0445±0.0098	−0.6553±0.0152
Histogram Binning	✓	✗	–	0.1494±0.0053	0.0537±0.0099	−0.6536±0.0152
Isotonic Regression	✓	✗	–	0.1459±0.0055	0.0460±0.0098	−0.6571±0.0150
GPcalib	✓	✗	–	0.1452±0.0054	0.0472±0.0098	−0.6578±0.0151
PBR	✓	✗	–	0.1590±0.0075	0.1118±0.0099	−0.6439±0.0168
BCT	✗	✓	–	0.1494±0.0052	0.1326±0.0091	−0.6536±0.0123
ConfAvg	✗	✓	–	0.1285±0.0056	0.0440±0.0092	−0.6745±0.0156†
DACA	✗	✗	–	0.1534±0.0045	0.0914±0.0098	−0.6495±0.0143
Dummy Reference	✗	✗	**80.48±1.03%**†	0.1681±0.0090	0.1498±0.0099	−0.6368±0.0191
Ours	✗	✗	**80.48±1.03%**†	0.1475±0.0065	**0.0218±0.0098**†	**−0.6573±0.0167**

Table 5: **Main results on CommonSenseQA with more baselines and error bars. Bold** and underline denote the **best** and second-best results among *label-free* methods that *do not* require access to Base Model during test time, while †indicates the *global* best across all methods. Accuracy for some Baselines is marked with – as these post-hoc methods do not alter final predictions.

Methods	**Label?**	**Base Access?**	**ACC** (↑)	**BS** (↓)	**ECE** (↓)	**CL** (↓)
			Qwen3-8B			
Base Model (Q_0)	✗	✗	**83.21**$\pm$**2.10%**†	**0.1167**$\pm$**0.0095**	**0.0611**$\pm$**0.0188**	**−0.7154**$\pm$**0.0286**†
Instruct Model (Q_1)	✗	✗	82.80$\pm$2.12%	0.1447$\pm$0.0181	0.1355$\pm$0.0199	−0.6833$\pm$0.0387
Temperature Scaling	✓	✗	–	0.1178$\pm$0.0113	0.0266$\pm$0.0192	−0.7102$\pm$0.0317
Histogram Binning	✓	✗	–	0.1278$\pm$0.0126	0.0270$\pm$0.0200	−0.7002$\pm$0.0325
Isotonic Regression	✓	✗	–	0.1171$\pm$0.0112	0.0234$\pm$0.0191	−0.7109$\pm$0.0306
GPcalib	✓	✗	–	0.1209$\pm$0.0123	0.0228$\pm$0.0195†	−0.7071$\pm$0.0315
PBR	✓	✗	–	0.1228$\pm$0.0135	0.0596$\pm$0.0194	−0.7052$\pm$0.0331
BCT	✗	✓	–	0.1165$\pm$0.0104†	0.0934$\pm$0.0181	−0.7115$\pm$0.0270
ConfAvg	✗	✓	–	0.1199$\pm$0.0128	0.0510$\pm$0.0191	−0.7082$\pm$0.0335
DACA	✗	✗	–	0.1226$\pm$0.0095	0.0631$\pm$0.0192	−0.7054$\pm$0.0299
Dummy Reference	✗	✗	82.80$\pm$2.12%	0.1447$\pm$0.0181	0.1354$\pm$0.0199	−0.6833$\pm$0.0387
Ours	✗	✗	82.80$\pm$2.12%	0.1309$\pm$0.0153	0.0638$\pm$0.0199	−0.6971$\pm$0.0360
			LLaMA-3.1-8B			
Base Model (Q_0)	✗	✗	71.33$\pm$2.54%	0.1708$\pm$0.0108	**0.0363**$\pm$**0.0231**†	−0.5425$\pm$0.0327
Instruct Model (Q_1)	✗	✗	**78.30**$\pm$**2.31%**†	0.2099$\pm$0.0224	0.2097$\pm$0.0229	−0.5731$\pm$0.0454
Temperature Scaling	✓	✗	–	0.1671$\pm$0.0111	0.0471$\pm$0.0227	−0.6158$\pm$0.0337
Histogram Binning	✓	✗	–	0.1687$\pm$0.0114	0.0430$\pm$0.0229	−0.6143$\pm$0.0342
Isotonic Regression	✓	✗	–	0.1651$\pm$0.0116	0.0400$\pm$0.0227	−0.6179$\pm$0.0338
GPcalib	✓	✗	–	0.1666$\pm$0.0111	0.0487$\pm$0.0228	−0.6164$\pm$0.0338
PBR	✓	✗	–	0.1889$\pm$0.0184	0.1456$\pm$0.0228	−0.5941$\pm$0.0391
BCT	✗	✓	–	0.1933$\pm$0.0134	0.1509$\pm$0.0218	−0.5896$\pm$0.0286
ConfAvg	✗	✓	–	0.1554$\pm$0.0140†	0.0605$\pm$0.0218	−0.6276$\pm$0.0367†
DACA	✗	✗	–	**0.1703**$\pm$**0.0099**	0.0740$\pm$0.0227	**−0.6126**$\pm$**0.0324**
Dummy Reference	✗	✗	**78.30**$\pm$**2.31%**†	0.2028$\pm$0.0216	0.1906$\pm$0.0229	−0.5802$\pm$0.0446
Ours	✗	✗	**78.30**$\pm$**2.31%**†	0.1776$\pm$0.0177	0.1054$\pm$0.0228	−0.6053$\pm$0.0407
			Ministral-3-8B			
Base Model (Q_0)	✗	✗	69.94$\pm$2.57%	0.1948$\pm$0.0085	0.1349$\pm$0.0235	−0.5046$\pm$0.0279
Instruct Model (Q_1)	✗	✗	75.35$\pm$2.42%	0.2050$\pm$0.0200	0.1691$\pm$0.0236	−0.5485$\pm$0.0433
Temperature Scaling	✓	✗	–	0.1930$\pm$0.0096	0.1111$\pm$0.0237	−0.5605$\pm$0.0321
Histogram Binning	✓	✗	–	0.1768$\pm$0.0112	0.0349$\pm$0.0235	−0.5766$\pm$0.0342
Isotonic Regression	✓	✗	–	0.1765$\pm$0.0112	0.0344$\pm$0.0235	−0.5770$\pm$0.0342
GPcalib	✓	✗	–	0.1808$\pm$0.0113	0.0560$\pm$0.0235	−0.5727$\pm$0.0334
PBR	✓	✗	–	0.1944$\pm$0.0158	0.1200$\pm$0.0237	−0.5591$\pm$0.0382
BCT	✗	✓	–	0.2226$\pm$0.0120	0.2458$\pm$0.0221	−0.5308$\pm$0.0248
ConfAvg	✗	✓	–	0.1638$\pm$0.0108†	0.0225$\pm$0.0227†	−0.5897$\pm$0.0339†
DACA	✗	✗	–	0.2098$\pm$0.0053	0.1674$\pm$0.0238	−0.5437$\pm$0.0283
Dummy Reference	✗	✗	**75.41**$\pm$**2.42%**†	0.2051$\pm$0.0200	0.1684$\pm$0.0236	−0.5490$\pm$0.0433
Ours	✗	✗	**75.41**$\pm$**2.42%**†	**0.1779**$\pm$**0.0148**	**0.0571**$\pm$**0.0235**	**−0.5762**$\pm$**0.0380**

Table 6: **Main results on MMLU with more baselines and error bars. Bold** and underline denote the **best** and second-best results among *label-free* methods that *do not* require access to Base Model during test time, while †indicates the *global* best across all methods. Accuracy for some Baselines is marked with – as these post-hoc methods do not alter final predictions.

Methods	**Label?**	**Base Access?**	**ACC** (↑)	**BS** (↓)	**ECE** (↓)	**CL** (↓)
			Qwen3-8B			
Base Model (Q_0)	✗	✗	74.55±0.72%	0.1367±0.0029	**0.0266±0.0061**	−0.6089±0.0093
Instruct Model (Q_1)	✗	✗	79.72±0.66%	0.1579±0.0054	0.1490±0.0060	−0.6392±0.0117
Temperature Scaling	✓	✗	–	0.1275±0.0034	0.0483±0.0058	−0.6696±0.0097
Histogram Binning	✓	✗	–	0.1362±0.0040	0.0179±0.0061†	−0.6610±0.0101
Isotonic Regression	✓	✗	–	0.1244±0.0036†	0.0269±0.0058	−0.6728±0.0096†
GPcalib	✓	✗	–	0.1274±0.0037	0.0533±0.0058	−0.6698±0.0098
PBR	✓	✗	–	0.1288±0.0040	0.0704±0.0058	−0.6683±0.0101
BCT	✗	✓	–	0.1550±0.0037	0.1143±0.0059	−0.6422±0.0083
ConfAvg	✗	✓	–	0.1288±0.0036	0.0471±0.0058	−0.6684±0.0100
DACA	✗	✗	–	**0.1342±0.0027**	0.0902±0.0058	**−0.6630±0.0091**
Dummy Reference	✗	✗	79.73±0.66%	0.1551±0.0052	0.1369±0.0060	−0.6422±0.0116
Ours	✗	✗	**79.75±0.66%**†	0.1412±0.0042	0.0499±0.0060	−0.6563±0.0106
			LLaMA-3.1-8B			
Base Model (Q_0)	✗	✗	64.39±0.79%	**0.1665±0.0029**†	**0.0090±0.0067**†	−0.4775±0.0094
Instruct Model (Q_1)	✗	✗	70.35±0.76%	0.2639±0.0069	0.2613±0.0073	−0.4396±0.0143
Temperature Scaling	✓	✗	–	0.1918±0.0034	0.0388±0.0072	−0.5116±0.0106
Histogram Binning	✓	✗	–	0.1966±0.0036	0.0189±0.0073	−0.5069±0.0108
Isotonic Regression	✓	✗	–	0.1888±0.0035	0.0181±0.0072	−0.5146±0.0105
GPcalib	✓	✗	–	0.1921±0.0035	0.0149±0.0072	−0.5114±0.0107
PBR	✓	✗	–	0.2216±0.0056	0.1697±0.0072	−0.4818±0.0123
BCT	✗	✓	–	0.1907±0.0037	0.1132±0.0068	−0.5128±0.0085
ConfAvg	✗	✓	–	0.1830±0.0040	0.1017±0.0068	−0.5204±0.0113†
DACA	✗	✗	–	0.1920±0.0031	0.0356±0.0072	**−0.5114±0.0104**
Dummy Reference	✗	✗	**70.37±0.76%**†	0.2603±0.0068	0.2540±0.0073	−0.4433±0.0142
Ours	✗	✗	70.35±0.76%	0.2181±0.0052	0.1426±0.0073	−0.4854±0.0126
			Ministral-3-8B			
Base Model (Q_0)	✗	✗	72.57±0.74%	**0.1376±0.0026**†	0.0558±0.0061	−0.5880±0.0089
Instruct Model (Q_1)	✗	✗	76.37±0.70%	0.2143±0.0065	0.2079±0.0068	−0.5494±0.0134
Temperature Scaling	✓	✗	–	0.1694±0.0037	0.0329±0.0068	−0.5943±0.0104
Histogram Binning	✓	✗	–	0.1729±0.0037	0.0113±0.0069†	−0.5909±0.0105
Isotonic Regression	✓	✗	–	0.1683±0.0038	0.0358±0.0068	−0.5954±0.0103
GPcalib	✓	✗	–	0.1686±0.0037	0.0228±0.0068	−0.5951±0.0104
PBR	✓	✗	–	0.1880±0.0052	0.1252±0.0068	−0.5757±0.0115
BCT	✗	✓	–	0.1546±0.0032	0.1069±0.0061	−0.6091±0.0083
ConfAvg	✗	✓	–	0.1507±0.0038	0.0758±0.0062	−0.6130±0.0107†
DACA	✗	✗	–	0.1716±0.0031	**0.0535±0.0068**	**−0.5921±0.0098**
Dummy Reference	✗	✗	**76.50±0.70%**†	0.2086±0.0063	0.1922±0.0069	−0.5564±0.0132
Ours	✗	✗	76.48±0.70%	0.1760±0.0045	0.0634±0.0068	−0.5889±0.0115

Table 7: **ECE sensitivity to discretization step s across benchmarks.** We default $s = 0.01$ in our main experiments. **Bold** denotes the **best** results among different s.

s	Qwen3-8B	LLaMA-3.1-8B	Ministral-3-8B
MMLU-Redux			
Base	0.0305±0.0091	0.0208±0.0103	0.0534±0.0091
Instruct	0.1269±0.0089	0.2526±0.0113	0.1656±0.0099
0.1	0.0795±0.0088	0.2020±0.0113	0.0641±0.0099
0.05	0.0752±0.0088	0.1855±0.0113	0.0548±0.0099
0.02	0.0501±0.0088	0.1622±0.0113	0.0437±0.0098
0.01 (Main)	**0.0348±0.0089**	0.1340±0.0114	**0.0218±0.0098**
0.005	0.0683±0.0089	0.0927±0.0113	0.0396±0.0099
0.002	0.0991±0.0090	**0.0514±0.0114**	0.0852±0.0100
CommonSenseQA			
Base	0.0611±0.0188	0.0363±0.0231	0.1349±0.0235
Instruct	0.1355±0.0199	0.2097±0.0229	0.1691±0.0236
0.1	0.0795±0.0199	0.1147±0.0228	0.0665±0.0235
0.05	0.0759±0.0200	0.1151±0.0228	0.0666±0.0234
0.02	0.0711±0.0199	0.1128±0.0228	0.0613±0.0235
0.01 (Main)	0.0638±0.0199	0.1054±0.0228	0.0571±0.0235
0.005	0.0511±0.0199	0.0920±0.0229	0.0451±0.0235
0.002	**0.0285±0.0201**	**0.0567±0.0229**	**0.0138±0.0236**
MMLU			
Base	0.0266±0.0061	0.0090±0.0067	0.0558±0.0061
Instruct	0.1490±0.0060	0.2613±0.0073	0.2079±0.0068
0.1	0.1023±0.0060	0.2106±0.0073	0.1055±0.0068
0.05	0.0976±0.0059	0.1936±0.0073	0.0962±0.0068
0.02	0.0734±0.0060	0.1707±0.0073	0.0848±0.0068
0.01 (Main)	**0.0499±0.0060**	0.1426±0.0073	0.0634±0.0068
0.005	0.0633±0.0060	0.1000±0.0073	**0.0419±0.0068**
0.002	0.0991±0.0060	**0.0568±0.0073**	0.0477±0.0069

Table 8: **ECE under alternative binning schemes.** UWB/UMB denote uniform-width and uniform-mass binning. OPT uses the sample-size-aware bin count $M^* = \max\{1, \lfloor n^{1/3} \rfloor\}$, where n is the number of evaluated examples. **Bold** denotes the **best** results.

Model	Method	UWB-10	UWB-15	UWB-OPT	UMB-10	UMB-15	UMB-OPT
MMLU-Redux							
Qwen3-8B	Instruct	0.1269±0.0089	0.1269±0.0088	0.1269±0.0088	0.1269±0.0086	0.1269±0.0086	0.1269±0.0086
	DACA	0.0985±0.0086	0.0985±0.0086	0.0988±0.0086	0.0987±0.0086	0.0987±0.0086	0.0979±0.0086
	Ours	**0.0348±0.0089**	**0.0643±0.0088**	**0.0570±0.0089**	**0.0693±0.0087**	**0.0708±0.0087**	**0.0711±0.0088**
LLaMA-3.1-8B	Instruct	0.2526±0.0113	0.2526±0.0113	0.2526±0.0113	0.2526±0.0111	0.2526±0.0111	0.2526±0.0111
	DACA	**0.0453±0.0111**	**0.0453±0.0111**	**0.0471±0.0111**	**0.0550±0.0111**	**0.0588±0.0111**	**0.0569±0.0111**
	Ours	0.1340±0.0114	0.1339±0.0113	0.1339±0.0113	0.1338±0.0112	0.1338±0.0112	0.1338±0.0113
Ministral-3-8B	Instruct	0.1656±0.0099	0.1656±0.0099	0.1656±0.0099	0.1656±0.0098	0.1656±0.0097	0.1656±0.0097
	DACA	0.0914±0.0098	0.0953±0.0098	0.0968±0.0098	0.0922±0.0098	0.0943±0.0098	0.0962±0.0098
	Ours	**0.0218±0.0098**	**0.0221±0.0098**	**0.0321±0.0098**	**0.0474±0.0098**	**0.0453±0.0098**	**0.0447±0.0098**
CommonSenseQA							
Qwen3-8B	Instruct	0.1355±0.0199	0.1357±0.0198	0.1355±0.0199	0.1347±0.0192	0.1347±0.0191	0.1347±0.0192
	DACA	**0.0631±0.0192**	0.0676±0.0192	**0.0631±0.0192**	0.0716±0.0192	**0.0695±0.0192**	0.0716±0.0192
	Ours	0.0638±0.0199	**0.0641±0.0199**	0.0638±0.0199	**0.0708±0.0197**	0.0707±0.0198	**0.0708±0.0197**
LLaMA-3.1-8B	Instruct	0.2097±0.0229	0.2097±0.0228	0.2097±0.0229	0.2075±0.0227	0.2075±0.0227	0.2075±0.0227
	DACA	**0.0740±0.0227**	**0.0740±0.0227**	**0.0740±0.0227**	**0.0776±0.0226**	**0.0750±0.0227**	**0.0776±0.0226**
	Ours	0.1054±0.0228	0.1072±0.0227	0.1054±0.0228	0.1039±0.0230	0.1039±0.0229	0.1039±0.0230
Ministral-3-8B	Instruct	0.1691±0.0236	0.1710±0.0236	0.1691±0.0236	0.1665±0.0235	0.1665±0.0235	0.1665±0.0235
	DACA	0.1674±0.0238	0.1660±0.0237	0.1674±0.0238	0.1660±0.0236	0.1677±0.0235	0.1660±0.0236
	Ours	**0.0571±0.0235**	**0.0573±0.0234**	**0.0571±0.0235**	**0.0524±0.0234**	**0.0655±0.0234**	**0.0524±0.0234**
MMLU							
Qwen3-8B	Instruct	0.1490±0.0060	0.1490±0.0060	0.1490±0.0059	0.1490±0.0058	0.1490±0.0058	0.1490±0.0058
	DACA	0.0902±0.0058	0.0927±0.0058	0.0950±0.0058	0.0944±0.0058	0.0951±0.0058	0.0945±0.0058
	Ours	**0.0499±0.0060**	**0.0641±0.0059**	**0.0613±0.0060**	**0.0691±0.0059**	**0.0818±0.0059**	**0.0820±0.0059**
LLaMA-3.1-8B	Instruct	0.2613±0.0073	0.2613±0.0073	0.2614±0.0073	0.2609±0.0072	0.2609±0.0072	0.2609±0.0072
	DACA	**0.0356±0.0072**	**0.0356±0.0072**	**0.0378±0.0072**	**0.0518±0.0072**	**0.0545±0.0072**	**0.0557±0.0072**
	Ours	0.1426±0.0073	0.1427±0.0073	0.1428±0.0073	0.1417±0.0072	0.1417±0.0072	0.1417±0.0072
Ministral-3-8B	Instruct	0.2079±0.0068	0.2079±0.0068	0.2079±0.0068	0.2079±0.0067	0.2079±0.0067	0.2079±0.0066
	DACA	**0.0535±0.0068**	**0.0573±0.0068**	**0.0574±0.0068**	0.0794±0.0067	0.0858±0.0067	0.0881±0.0067
	Ours	0.0634±0.0068	0.0633±0.0068	0.0633±0.0067	**0.0631±0.0067**	**0.0683±0.0067**	**0.0673±0.0067**

Table 9: **SmoothECE on all datasets (lower is better).** Each entry reports the SmoothECE estimate followed by a 95% bootstrap percentile interval in brackets. Within each (dataset,model) setting, all methods are evaluated on the same examples. *Matched* n counts examples whose sample IDs appear in the valid predictions of all 12 compared methods, including methods not displayed here. Intervals are computed from 1000 paired bootstrap resamples, using the same sampled examples for every method. See Appendix F.4 for details.

Model	Method	MMLU	MMLU-Redux	CommonSenseQA
`Qwen3-8B`	*Matched* n	14,040	5,700	1,221
	Base	0.0262 [0.0221, 0.0322]	0.0291 [0.0254, 0.0377]	0.0552 [0.0407, 0.0736]
	Instruct	0.1751 [0.1683, 0.1825]	0.1510 [0.1402, 0.1619]	0.1634 [0.1397, 0.1855]
	TS *(Supervised)*	0.0486 [0.0453, 0.0518]	0.0445 [0.0393, 0.0501]	0.0295 [0.0246, 0.0459]
	DACA *(Label-Free)*	0.0806 [0.0773, 0.0837]	0.0870 [0.0820, 0.0925]	0.0660 [0.0531, 0.0800]
	Ours *(Label-Free)*	0.0475 [0.0432, 0.0530]	0.0405 [0.0363, 0.0449]	0.0625 [0.0438, 0.0814]
`LLaMA-3.1-8B`	*Matched* n	14,038	5,699	1,221
	Base	0.0116 [0.0119, 0.0179]	0.0211 [0.0178, 0.0312]	0.0318 [0.0270, 0.0497]
	Instruct	0.2860 [0.2781, 0.2938]	0.2793 [0.2662, 0.2914]	0.2233 [0.1975, 0.2491]
	TS *(Supervised)*	0.0373 [0.0332, 0.0417]	0.0374 [0.0304, 0.0436]	0.0459 [0.0347, 0.0663]
	DACA *(Label-Free)*	0.0340 [0.0299, 0.0397]	0.0409 [0.0336, 0.0486]	0.0716 [0.0519, 0.0939]
	Ours *(Label-Free)*	0.1416 [0.1346, 0.1486]	0.1337 [0.1215, 0.1449]	0.1042 [0.0828, 0.1281]
`Ministral-3-8B`	*Matched* n	14,019	5,687	1,220
	Base	0.0493 [0.0449, 0.0536]	0.0529 [0.0473, 0.0616]	0.1352 [0.1097, 0.1579]
	Instruct	0.2275 [0.2202, 0.2349]	0.1827 [0.1728, 0.1936]	0.1630 [0.1408, 0.1903]
	TS *(Supervised)*	0.0344 [0.0290, 0.0396]	0.0441 [0.0367, 0.0517]	0.0970 [0.0804, 0.1139]
	DACA *(Label-Free)*	0.0513 [0.0461, 0.0574]	0.0858 [0.0771, 0.0945]	0.1665 [0.1416, 0.1885]
	Ours *(Label-Free)*	0.0630 [0.0566, 0.0694]	0.0259 [0.0231, 0.0311]	0.0556 [0.0383, 0.0786]

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 main claims about the mutual-calibration condition, the projection-based algorithm, the worst-case improvement guarantee, and the empirical calibration gains, and these are matched by the formal results in Section 3 and the experiments in Sections 4 and 5.

   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**

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

   Answer: [Yes]

   Justification: The paper discusses practical limitations of the method, including the need for an (approximately) calibrated reference, the role of a small labeled diagnostic set for validating that choice, computational limitations in open-ended generation tasks and the discretization bias–variance tradeoff; see Section 6. It also describes an empirical failure case for `LLaMA-3.1-8B` in Section 5.

   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 problem setup and assumptions are stated in Section 2, and the main theorem is presented in Section 3. Full proofs and extensions are provided in Appendix B, Appendix C, Appendix D, and Appendix E.

   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 paper specifies the datasets, model pairs, calibration split/test split usage, optimization objective, baselines, and evaluation metrics in Section 4. Additional implementation details, prompts, confidence extraction, discretization, solver configuration, baseline definitions, and robustness analyses are given in Appendix F and Appendix G.

   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 public repository linked in Appendix F.2 provides our implementation, environment specifications, model-download instructions, dataset sources and split definitions, the required local input layout, and commands for reproducing the main results and baselines. Datasets and model checkpoints are obtained separately from the sources listed in Table 2.

   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 4 describes the datasets, model selection, calibration/evaluation protocol, baselines, and evaluation metrics. Appendix F.1, Appendix F.2, Appendix F.2.3, Appendix F.3, and Appendix F.4 provide the detailed prompting, confidence extraction, discretization step, optimization setup, and metric formulas.

   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: Section G provides confidence estimates for the experimental results in the main paper. The procedures and conditioning assumptions are described in Appendix F.4 and Appendix G.1.

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.2 specifies the compute platform, GPU and CPU configurations, memory, and the computational costs of the reported experiments.

   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 work consists of theoretical analysis and evaluation on existing public multiple-choice benchmarks and publicly available model checkpoints, without crowdsourcing, human-subject studies, or release of new high-risk assets; see Section 4 and Appendix F.1. Existing assets are credited in Appendix F.1 and used subject to their licenses and terms of use.

   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: This paper develops a label-free post-processing method and theoretical framework for improving calibration using a better-calibrated reference model. The work is methodological and is evaluated on standard multiple-choice QA benchmarks, without introducing new models, datasets, deployment systems, or application-specific decision pipelines. Therefore, it does not create a direct path to societal impact beyond the general downstream uses of better-calibrated predictive models. Potential application-level impacts would depend on how such calibrated models are deployed in specific domains, which is outside the scope of this paper.

    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 new pretrained models, scraped datasets, or other high-risk assets. It studies a post-processing method on existing open-source LLM checkpoints and public QA benchmarks; see Section 4 and Appendix F.1.

    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: Appendix F.1 identifies and credits the datasets, model checkpoints, and software used in our experiments, including their versions, original sources, and applicable licenses. Our use of these assets follows their stated licenses and terms of use.

    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: [Yes]

    Justification: The public repository linked in Appendix F.2 provides our implementation under the MIT License, installation instructions, and reproduction commands. External datasets and model checkpoints used in our experiments remain subject to their original licenses and terms of use. The method's assumptions and limitations are discussed in Section 6, with an empirical failure case analyzed in Section 5. We do not release new datasets or model checkpoints.

    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 paper does not involve crowdsourcing or human-subject experiments. It evaluates existing models on established benchmark datasets; see Section 4 and Appendix F.1.

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: No human subjects are involved in this work, so IRB review is not applicable. The study is limited to theory and experiments on public benchmarks with existing model checkpoints; see Section 4 and Appendix F.1.

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: LLMs are the experimental evaluation targets rather than an important or non-standard component of the proposed method itself. The core contribution is the label-free calibration framework and its theoretical analysis in Sections 2 and 3, with LLM experiments used as an application domain in Section 4.

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.