A Comprehensive Evaluation of Post-hoc Calibration Methods Across Modern Vision Architectures and Datasets

SHUJIAN LI, University of Twente, The Netherlands

This study presents a comprehensive evaluation of five post-hoc calibration methods (*Temperature Scaling*, *Isotonic Regression*, *Histogram Binning*, *Dirichlet Calibration*, and a composite *Dirichlet-Isotonic* method) across a variety of modern vision architectures (ViT, ResNet, ConvNeXt) and datasets. Our multi-perspective analysis reveals that a one-size-fits-all approach to calibration is insufficient. We demonstrate that simple methods like *Temperature Scaling* are unreliable and can completely fail on certain architectures, whereas flexible non-parametric and composite methods provide statistically significant improvements in calibration. The composite *Dirichlet-Isotonic* method consistently proves to be the most effective and robust choice, successfully correcting complex, architecture-specific error patterns. Furthermore, our work uncovers critical trade-offs, such as the potential to improve average calibration error (ECE) at the cost of worst-case error (MCE), highlighting the need for careful assessment. The primary contribution of this work is an evidence-based decision framework that guides practitioners in selecting the optimal calibration method based on their specific model architecture and application requirements, thus contributing to more trustworthy and reliable AI systems.

Additional Key Words and Phrases: Model calibration, deep neural networks, post-hoc calibration, temperature scaling, isotonic regression, Dirichlet calibration, vision transformers, convolutional neural networks, uncertainty quantification, reliability diagrams, expected calibration error.

1 Introduction

Psychologist Fischhoff found in his research that people tend to be overconfident in difficult tasks (confidence > accuracy), while they tend to be underconfident in very simple tasks (confidence < accuracy). This is known as the "hard-easy effect"[21]. Psychological research classifies this type of confidence bias as metacognitive bias, which means that when an individual assesses their confidence level in their knowledge or performance, they systematically deviate from the actual level[4]. In Fischhoff's *probability calibration model*, calibration refers to the degree of agreement between a person's subjective confidence and their actual accuracy in decision-making or judgment tasks. Good calibration is critical in high-stakes fields such as medicine[2][5], law, and finance[1]. Miscalibration of confidence can sometimes have serious consequences, leading to flawed judgments and poor decision-making.

Recently, a similar phenomenon has been observed in the field of artificial intelligence. Like humans, DNNs obtain their "confidence" through exposure to vast amounts of information, though in different forms. Whereas human metacognition is influenced by years of life experience, social feedback, and cognitive reflection, the confidence of a neural network is molded through optimization during supervised learning. A DNN training process involves minimizing a loss function (often cross-entropy) for better classification. Specifically, the network first outputs raw scores (logits) that are subsequently fed into a softmax function to obtain a probability distribution over classes. Softmax outputs (values 0-1) are often interpreted as the model's confidence[6, 9]. Consequently, the model learns not only to predict the correct class, but also to output a high probability for the predicted label. However, similar to humans, DNNs are also susceptible to metacognitive biases such as the hard-easy effect; they can be overconfident, particularly when dealing with out-of-distribution data or when the input does not favor any one class. These models can produce predictions with high confidence scores even when they are incorrect, a problem known as miscalibration[6, 8].

This issue is particularly obvious in contemporary high-capacity architectures, such as Vision Transformers (ViTs)[3] and large Convolutional Neural Networks (CNNs). As these models are increasingly deployed in high-stakes, safety-critical applications like autonomous driving[10] and medical diagnostics[2], ensuring that their confidence scores are a faithful representation of their likelihood of being correct is paramount. This miscalibration arises from several factors inherent to modern deep learning, including model capacity, the nature of standard training objectives, the quantity and quality of training data, and specific architectural choices[6, 12, 15, 24], which we will detail further in Section 2.3.

To address this, researchers have developed a variety of calibration techniques, which can be broadly categorized into two families: in-training and post-hoc methods[6, 24]. In-training methods modify the model's architecture or training regime, for example by using alternate loss functions (e.g., focal loss[13]) or confidence-penalizing regularization[16]. On the other hand, post-hoc methods operate on the outputs of a pre-trained model without retraining, and thus are often highly practical in many real-world settings[6, 11]. We focus our study on post-hoc methods due to their compelling practical advantages. They are easy to implement, computationally efficient, and can maintain the original model's accuracy. Most importantly, in an era dominated by large-scale, pre-trained base models, post-hoc techniques offer a valuable and feasible way to improve the reliability of models, which are resource-intensive to retrain[11].

This study systematically evaluates five different post-hoc calibration methods ranging from simple, well-known methods to more complex, state-of-the-art methods. We begin with Temperature Scaling, a simple yet surprisingly effective method that serves as a strong baseline[6]. We then investigate two classic non-parametric methods, Isotonic Regression[26] and Histogram Binning[25], which offer greater flexibility to model complex calibration errors[26]. Realizing the limitations of these methods in multiclass settings, we also include Dirichlet Calibration, a more recent and powerful parametric method designed to handle class-wise calibration issues[11]. Finally, inspired by recent work[27], we explored composite methods, where we apply Dirichlet followed by Isotonic Regression to leverage the strengths of both methods.

Despite the importance of calibration, there is a significant gap in the current literature. Although several studies have evaluated calibration, none of them have provided a systematic comparison across diverse modern vision architectures (e.g., Transformers, CNNs, and

Author's Contact Information: ShuJian Li, University of Twente, Enschede, The Netherlands.

hybrid models). Furthermore, the evaluations of the methods' performance were not comprehensive enough to capture the full picture of a method's performance. In this paper, we aim to fill this gap by conducting a comprehensive, multi-perspective evaluation of post-hoc calibration methods. We leverage an evaluation framework that analyzes performance from four dimensions: calibration quality, predictive performance, reliability, and robustness. By testing across a variety of models and datasets, we gain a deeper understanding of how these methods perform under different settings. Our main contribution is not only a thorough empirical analysis but also the development of a systematic decision framework to guide practitioners in selecting the optimal calibration method for their specific use case.

In this study, our research seeks to answer the following questions:

1. How do different post-hoc calibration methods perform across modern vision architectures like Vision Transformers, ResNets, and ConvNeXt?
2. What is the trade-off between improvements in calibration quality and the preservation of predictive performance?
3. Which calibration methods are most robust and generalize best across a diverse range of datasets and model architectures?
4. Can composite calibration methods systematically outperform their individual components?
5. How can we develop a systematic framework to guide the selection of an optimal post-hoc calibration method?

2 Background and Theoretical Foundations

To systematically evaluate post-hoc calibration methods, it is essential to first establish a clear theoretical foundation. This section defines the core concepts of model calibration, details the primary metrics and diagrams used for its measurement, discusses the theoretical basis of probabilistic predictions, and explores the sources of miscalibration in modern neural networks.

2.1 Defining and Measuring Calibration

Imagine a weather forecaster who predicts a "70% chance of rain." We intuitively trust this forecast if, over many days with such a prediction, it actually rains about 70% of the time. This is the essence of calibration: to ensure the model's predicted probabilities reflect the true likelihood of events.

In machine learning, a model predicts a class $\hat{Y}$ with a certain confidence score $\hat{P}$. We say the model is *perfectly calibrated* if the confidence score truly reflects the probability of being correct. Mathematically, this simple idea is expressed as:

$$\mathbb{P}(\hat{Y} = Y \mid \hat{P} = p) = p, \quad \forall p \in [0, 1]$$

where $\hat{Y}$ is the predicted class, Y is the true label, and $\hat{P}$ is the confidence score. Crucially, calibration differs from accuracy: while accuracy measures prediction correctness ($\hat{Y} = Y$), calibration assesses confidence reliability. A highly accurate model assigning uniform 99.9% confidence to all predictions remains poorly calibrated.

In practice, the condition for perfect calibration cannot be verified for every possible confidence value p. Instead, miscalibration is estimated by partitioning the predictions into M confidence bins. This is commonly visualized using a *Reliability Diagram*[17] (Figure 1), which plots the average accuracy within each bin against the average confidence within that same bin[6]. For a perfectly calibrated model, these points would lie on the diagonal identity line. Deviations from this line signify miscalibration.

To quantify this deviation, we use several metrics:

- **Expected Calibration Error (ECE)**: This is the most common metric. It measures the average gap between confidence and accuracy across all the bins, weighted by the number of samples in each. A lower ECE indicates better calibration[6].

  $$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|$$

  where n is the total number of samples, B_m is the set of indices of samples whose prediction confidence falls into the m-th interval, $\text{acc}(B_m)$ is the accuracy of bin B_m, and $\text{conf}(B_m)$ is the average confidence in bin B_m. These are formally defined as:

  $$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{Y}_i = Y_i)$$

  $$\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{P}_i$$

  While widely used, ECE's value can be sensitive to the number of bins M and the binning strategy (e.g., equal-width vs. equal-mass)[19].
- **Maximum Calibration Error (MCE)**: For high-stakes applications, the average error is insufficient; we need to know the worst-case scenario. MCE identifies the single bin with the largest confidence-accuracy gap[6]:

  $$\text{MCE} = \max_{m \in \{1,\ldots,M\}} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|$$

- **Adaptive Calibration Error (ACE)**: A refinement of Expected Calibration Error (ECE) that uses equal-mass bins to compute calibration error, improving robustness against skewed confidence score distributions[19].

  $$\text{ACE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|$$

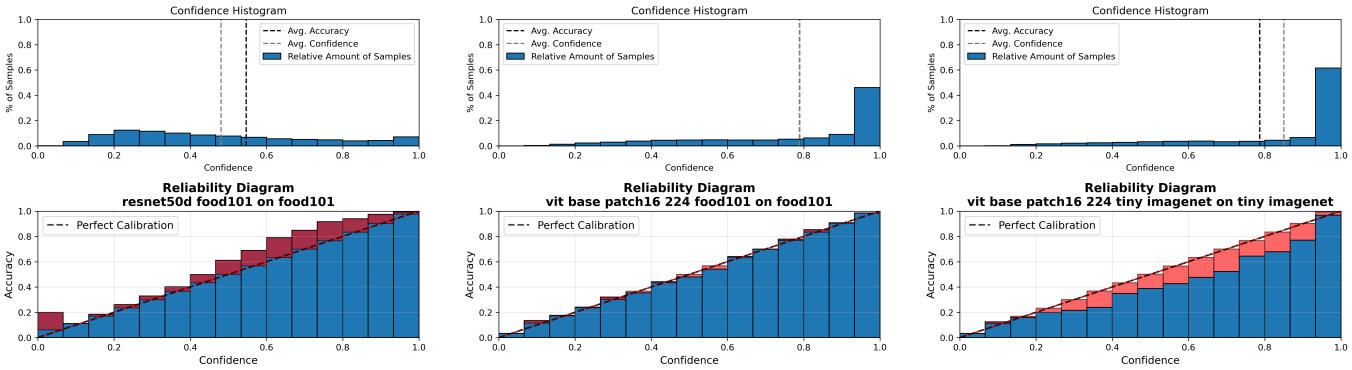
- **Classwise ECE**: Crucial for multiclass problems, this metric calculates the ECE for each class separately. This is important because a model might be well-calibrated for common classes but poorly calibrated for rare ones[11].

  $$\text{Classwise ECE} = \sum_{c=1}^{C} \text{ECE}(c)$$

  where C is the number of classes, and $\text{ECE}(c)$ is the ECE for class c.

2.2 Decomposition of Probabilistic Forecasts

A deeper theoretical understanding of calibration can be gained by decomposing proper scoring rules, such as the Brier score. The Brier



(a) Underconfident Model ResNet50-D on Food-101

(b) Well-Calibrated Model ViT on Food-101

(c) Overconfident Model ViT on Tiny-ImageNet

Fig. 1. Representative reliability diagrams illustrating different calibration behaviors across model-dataset combinations. The diagonal dashed line represents perfect calibration. (a) An underconfident model where predicted confidence is systematically lower than actual accuracy, demonstrating conservative behavior. (b) A well-calibrated model where confidence closely matches accuracy across all confidence bins, representing the ideal calibration state. (c) An overconfident model where predicted confidence consistently exceeds actual accuracy, particularly in high-confidence regions where most practical decisions are made.

score measures the mean squared error between predicted probabilities and one-hot encoded true labels. For a set of probabilistic forecasts, it can be additively decomposed into three components[17]:

$$\text{Brier Score} = \text{Reliability} - \text{Resolution} + \text{Uncertainty}$$

- **Reliability** measures the weighted average of the squared differences between the mean forecast probability and the true conditional probability for each bin. It is a direct measure of miscalibration; a perfectly calibrated model has a reliability term of zero.
- **Resolution** measures the ability of the model to separate samples into subpopulations with different outcomes. A higher resolution indicates a more informative model that can confidently distinguish between easy and hard cases.
- **Uncertainty** reflects the inherent variability of the outcomes in the dataset and is an irreducible error component that is independent of the model.

This decomposition exposes a critical calibration-performance trade-off[17]. Post-hoc methods tend to focus mainly on improving reliability (pushing it towards zero via probability smoothing), at the cost of potentially degrading resolution—the capacity of the model to output distinct, discriminative predictions[6]. Sharpness, measured by the negative entropy of predictive distributions, represents a related consideration[12]. While we desire clear predictions (low entropy), these must also be well-calibrated. The optimal calibration method should maximize reliability gains while preserving as much resolution and predictive accuracy as possible.

2.3 Sources of Miscalibration in Modern Architectures

Modern vision models achieve impressive results but face new calibration challenges. While older models like AlexNet often had well-calibrated confidence scores, today's powerful models tend to be overconfident[6]. Four key factors explain this trend:

First, the training process itself contributes. Models trained with standard cross-entropy loss are incentivized to maximize accuracy. With millions of parameters, they learn to be highly confident (pushing probabilities toward 1.0) since this most quickly reduces training loss[15]. The system prioritizes being right over being honest about uncertainty. Although techniques like using focal loss[13] or label smoothing[18] can help during training, they do not address already-trained models[13, 18, 22].

Second, specific design choices affect confidence scores. Batch Normalization—crucial for stable training—unintentionally affects confidence scores in ways that hurt calibration[6]. Newer approaches like Layer Normalization (used in Transformers) have different but equally important impacts, showing that every architectural choice influences calibration[15].

Third, the quality and quantity of training data fundamentally impact a model's ability to produce reliable confidence estimates. When complex models are trained on limited data, they tend to memorize individual examples instead of learning general patterns, leading to overconfident predictions for unusual inputs. Techniques like Mixup/CutMix data augmentation and pretraining on large datasets before fine-tuning help models generalize better[23]. Class imbalance introduces additional challenges—models frequently become overconfident in common classes while struggling with rare ones. Researchers address this issue by adjusting the loss function to focus more on harder examples during training[13]. Even with perfectly balanced data, differences between training and real-world environments can undermine reliability, as models may encounter data patterns they were not trained on[20].

Finally, a model's built-in assumptions matter profoundly. Traditional CNNs start with strong biases about how images work (focusing on local patterns), while Vision Transformers (ViTs)[3] instead process images as sequences of patches, without such innate biases. Although this allows ViTs to discover global patterns, it also enables them to develop unnatural confidence patterns that

CNNs avoid. This fundamental difference explains why no single calibration solution works for all architectures—each needs tailored approaches[15].

3 Post-hoc Calibration Methods

The term "post-hoc" means "after the fact." Post-hoc methods improve a model's confidence scores by learning a transformation on its outputs (logits or probabilities) using a separate calibration set, thus avoiding the need for costly retraining. This section details the five methods evaluated in this study, which range from simple parametric techniques to more flexible non-parametric and composite methods.

3.1 Temperature Scaling (TS)

Temperature Scaling is the classic "less is more" approach in calibration. It is an effective extension of Platt Scaling[26] to the multiclass setting[6]. It operates on the principle that miscalibration is often due to systemic over- or under-confidence, which can be corrected by "softening" or "sharpening" the softmax function. This is done by adding a single scalar parameter, the temperature $T > 0$, which divides the logits $\mathbf{z}$ before the softmax operation. The calibrated probability $\hat{q}_i$ for class i is given by:

$$\hat{q}_i = \frac{\exp(z_i/T)}{\sum_{j=1}^{K} \exp(z_j/T)}$$

The temperature T is obtained by minimizing the Negative Log-Likelihood (NLL) on a held-out calibration set. If $T > 1$, the resulting probability distribution becomes softer (less confident), correcting for overconfidence. If $T < 1$, the distribution becomes sharper (more confident). Crucially, it does not alter the ranking of the predictions; the class with the highest probability remains the same. Its simplicity and low computational overhead make it a strong and widely used baseline.

3.2 Isotonic Regression (ISO)

However, if the calibration error is not that simple—e.g., the model is overconfident for some predictions but underconfident for others—then this is where a more flexible, non-parametric method like Isotonic Regression comes in. It provides greater flexibility than Temperature Scaling by not assuming a fixed functional form for the calibration map[26]. It learns a non-decreasing, piecewise-constant function that maps the model's original confidence scores to calibrated probabilities. This power comes with a price: it requires more data and is more prone to overfitting compared to TS. We included it as a representative of a significant step up in flexibility from simple parametric models.

3.3 Histogram Binning (HB)

Histogram Binning is another non-parametric classic, perhaps the most intuitive of all. It divides the confidence space into a set of M bins and learns a simple correction for each bin[25]. For a prediction with confidence p that falls into bin B_m, the calibrated confidence $\hat{p}$ is set to the empirical accuracy of all samples within that bin:

$$\hat{p} = \text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i)$$

The primary design choice in Histogram Binning is the binning strategy. Equal-width binning divides the $[0, 1]$ interval into bins of the same size, while equal-mass (or equal-frequency) binning creates bins with an equal number of samples. Intuitive and easily interpretable, Histogram Binning's performance is sensitive to the number of bins, and it can produce unstable results if bins are sparsely populated. We selected it as a representative of this fundamental "bin-and-correct" philosophy.

3.4 Dirichlet Calibration (Dir)

The global nature of Temperature Scaling becomes a critical weakness in complex multiclass problems where miscalibration patterns differ across classes. Dirichlet Calibration is a more powerful parametric method designed specifically to address this limitation[11]. It learns a class-wise affine transformation in the log-probability space. Given an uncalibrated probability vector $\mathbf{p}$, the calibrated logits $\mathbf{z}'$ are computed as:

$$\mathbf{z}' = \mathbf{W} \log(\mathbf{p}) + \mathbf{b}$$

The final calibrated probabilities are then obtained by applying a softmax function to $\mathbf{z}'$. The parameters, a weight matrix $\mathbf{W}$ (often constrained to be diagonal) and a bias vector $\mathbf{b}$, are optimized on the calibration set. This formulation is equivalent to fitting a Dirichlet distribution to the model's posterior probabilities and provides a flexible framework for correcting intricate, class-dependent calibration errors that simpler methods cannot capture. This method was selected because it achieves such a balance: significantly more expressive than TS, yet still parametric, making it more data-efficient and less prone to overfitting than non-parametric methods like ISO.

3.5 Composite Methods (Dir+Iso)

However, what if the best solution is not one method, but a combination of them? This is the compelling idea behind composite calibration, in which methods are stacked sequentially. Recent work, particularly the "Mix-n-Match" paradigm explored by Zhang et al.[27], has shown that such compositional recipes can create calibration maps that are more expressive and robust than any single method alone.

Our curiosity drove us to investigate a powerful pairing: Dirichlet followed by Isotonic Regression (Dir+Iso). The logic here is compelling and speaks directly to a classic bias-variance trade-off.

(1) **Stage 1 (The Broad-Stroke Correction):** The Dirichlet calibrator acts as a low-variance, parametric tool. Its job is to perform the initial, heavy lifting—fixing the large, structural, class-specific biases that a simple method like TS would miss.
(2) **Stage 2 (The Fine-Tuning):** After Dirichlet has fixed the gross errors, the Isotonic Regressor takes over. It acts as a high-variance, non-parametric "finisher," mopping up any subtle, non-monotonic residual errors that remain.

This two-stage process yields a strictly more expressive calibration map. The composite function, $f_{\text{comp}}(p) = g_{\text{iso}}(f_{\text{DC}}(p))$, can capture complex error landscapes that are inaccessible to a single

calibrator[15, 27]. This is especially powerful in low-data regimes; rather than relying on one powerful but data-hungry method, we use a simpler tool to get most of the way there, leaving less work for the more flexible second stage. While stacking methods can potentially alter the model's accuracy, the implementation pipeline, formalized in Algorithm 1, is straightforward and adds negligible inference cost, making it a highly practical strategy for pushing the boundaries of calibration performance.

Algorithm 1 Composite Calibration Pipeline

Require: Model f, calibration set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^{n}$, test set $\mathcal{D}_{\text{test}}$
Ensure: Calibrated predictions P_{cal}

// **Extract uncalibrated predictions**
$P_{\text{uncal}}^{\text{cal}} \leftarrow \{f(x_i) : (x_i, y_i) \in \mathcal{D}_{\text{cal}}\}$
$P_{\text{uncal}}^{\text{test}} \leftarrow \{f(x_i) : (x_i, y_i) \in \mathcal{D}_{\text{test}}\}$
$Y_{\text{cal}} \leftarrow \{y_i : (x_i, y_i) \in \mathcal{D}_{\text{cal}}\}$
// **Fit Dirichlet calibrator**
$\theta_{\text{dir}} \leftarrow \arg\max_{\theta} \sum_i \log p_{\text{dir}}(y_i | P_{\text{uncal}}^{\text{cal}}[i], \theta)$
$P_{\text{dir}}^{\text{cal}} \leftarrow \text{ApplyDirichlet}(P_{\text{uncal}}^{\text{cal}}, \theta_{\text{dir}})$
$P_{\text{dir}}^{\text{test}} \leftarrow \text{ApplyDirichlet}(P_{\text{uncal}}^{\text{test}}, \theta_{\text{dir}})$
// **Fit Isotonic regression on Dirichlet-calibrated probabilities**
$\theta_{\text{iso}} \leftarrow \text{FitIsotonic}(P_{\text{dir}}^{\text{cal}}, Y_{\text{cal}})$
$P_{\text{cal}} \leftarrow \text{ApplyIsotonic}(P_{\text{dir}}^{\text{test}}, \theta_{\text{iso}})$
return P_{cal}

4 Experimental Methodology

Our experimental design was created to answer a fundamental question: how do we fairly and comprehensively compare post-hoc calibration methods across the diverse landscape of modern vision models? We developed a systematic framework that moves beyond single-metric evaluations to capture the nuanced ways these methods perform under different architectural and data-driven pressures.

4.1 Experimental Setup

To create a robust testbed, our experiments encompass 12 model-dataset combinations. Our model lineup spans three distinct paradigms in modern computer vision, chosen to provide a comprehensive architectural analysis. We selected ViT-Base[3] as a representative of Transformer architectures that process images as a sequence of patches, avoiding the convolutional inductive biases of traditional CNNs. To represent mature, highly optimized CNNs, we chose ResNet50-D[7], an improved variant of the classic Residual Network. Lastly, for modern hybrid designs, we included ConvNeXt-Base[14], which adapts standard CNNs with principles from Vision Transformers. This selection, with all models initialized with pre-trained weights from the `timm` library, allows us to test calibration on models with fundamentally different inductive biases. These architectures were evaluated on datasets representing a spectrum of difficulty: *CIFAR-10 and CIFAR-100* serve as controlled benchmarks for studying the effect of class count, *Tiny-ImageNet* increases the complexity, and *Food-101* presents a real-world, fine-grained classification challenge where visual similarities between classes can confound even human experts.

Experimental Protocol: To mimic real-world scenarios, we used a parameter-efficient fine-tuning strategy; that is, we froze the pre-trained backbones and only trained the final classification head for 5 epochs. For every experiment, the dataset was strictly split into an 80% fine-tuning set, a 20% calibration set (used solely for fitting the calibrators), and the original held-out test set for final evaluation. This strict separation prevents any information leakage and ensures the integrity of our results[6].

4.2 The Four-Perspective Evaluation Framework

A single metric like ECE is insufficient to capture the full impact of a calibration method. We therefore developed a multi-perspective evaluation framework that first analyzes performance using a suite of foundational metrics and then synthesizes these into a unified visual tool—the Assessment Matrix—for holistic comparison.

4.2.1 Foundational Analysis Perspectives. Our initial analysis is built on four pillars, each addressing a critical aspect of performance:

(1) **Calibration Quality:** Measures the alignment between confidence and correctness using Expected Calibration Error (ECE), Maximum Calibration Error (MCE), Adaptive Calibration Error (ACE)[19], and Classwise ECE[11].
(2) **Predictive Performance:** Quantifies the impact on the model's core predictive power by tracking Top-1 and Top-5 accuracy, as well as proper scoring rules like Negative Log-Likelihood (NLL) and the Brier Score[17].
(3) **Reliability:** Decomposes the Brier score into its Reliability, Resolution, and Uncertainty components[17], and measures prediction consistency to understand the *nature* of the calibration improvement.
(4) **Robustness:** Assesses generalization by measuring performance consistency across different datasets and architectures, validated with statistical significance testing (paired t-tests).

4.2.2 The Assessment Matrix: A Unified View for Comparison. To synthesize the dozens of metrics from our foundational analysis into a clear and comparable format, we developed the *Assessment Matrix*, visualized as a radar chart. This matrix distills performance into four intuitive axes, providing a holistic profile of each method's strengths and weaknesses.

The four axes of the Assessment Matrix are:

- **Calibration Accuracy:** Derived from ECE, this axis directly measures how well confidence scores match empirical accuracy. A higher score indicates lower calibration error.
- **Predictive Performance:** Based on the model's raw Top-1 accuracy, this axis confirms that the method does not degrade the model's ability to make correct predictions.
- **Accuracy Preservation:** This measures the ratio of the calibrated accuracy to the baseline (uncalibrated) accuracy. A high score signifies that the method improves calibration without harming the original model's performance.
- **Training Stability:** Derived from the consistency metric, this axis reflects the method's stability across different confidence regimes, with higher scores indicating more reliable and less erratic behavior.

Each axis is normalized to a 0–10 scale, where 10 is the ideal score. The resulting radar chart provides an instant visual summary: a method with a larger and more balanced area is superior and more well-rounded. This Assessment Matrix is the primary visual instrument used in our Results section to compare methods across different architectures and datasets.

5 Results and Analysis

Our comprehensive evaluation across three model architectures and four datasets produced a rich set of results that offer a nuanced view of post-hoc calibration. In this section, we present our findings, first comparing the methods at a high level, then examining how performance is influenced by model architecture and dataset difficulty, and finally validating our core claims statistically.

5.1 Overall Performance Comparison

To begin, we evaluate the overall performance of each calibration method averaged over all experimental configurations. Figure 2 offers a high-level summary. While all methods offer some improvement over the `uncalibrated` baseline, two methods quickly stand out: both Isotonic Regression (`iso`) and the composite `dir_iso` methods achieve the lowest average Expected Calibration Error (ECE). Importantly, the Predictive Performance panel confirms a crucial point: no method significantly harms the model's accuracy.

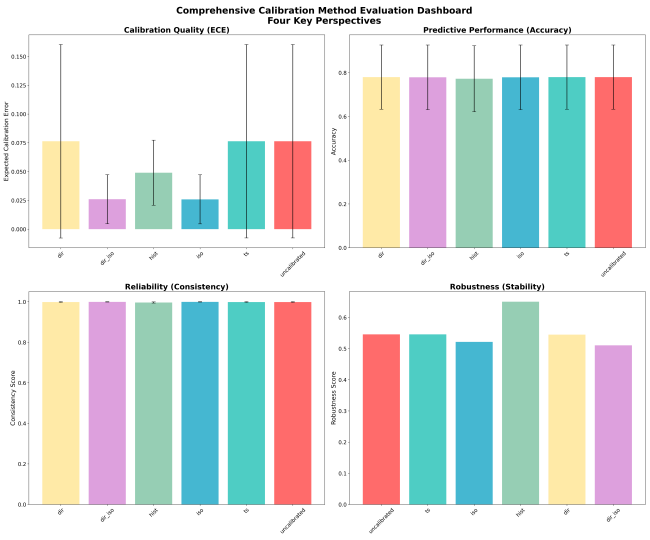


Fig. 2. A high-level dashboard summarizing the average performance of each calibration method across all experiments. Error bars show standard deviation. Lower is better for ECE; higher is better for Accuracy and Consistency.

Quantifying these visual trends, Table 1 reveals that `dir_iso` and `iso` are the top performers in reducing ECE. Not only do they succeed, but they achieve the greatest success on average. An intriguing nuance in the table is that some methods such as `TS` perfectly preserve the original model's accuracy. This is by design, as Temperature Scaling applies a monotonic transformation that cannot change the model's top prediction. In contrast, more flexible methods such as `ISO` exhibit a slight, likely insignificant, decrease in accuracy. This is a known trade-off: their ability to correct more complex calibration errors comes with a slight risk of overfitting to the calibration set, resulting in these minor performance variations. This is a minimal price to pay for the substantial improvement in calibration quality.

Table 1. Mean performance metrics across all model and dataset configurations. Lower ECE is better; higher Accuracy is better.

Method	Mean ECE (↓)	ECE Std. Dev.	Mean Accuracy (↑)
Uncalibrated	0.076	0.080	**0.780**
TS	0.076	0.080	**0.780**
ISO	**0.026**	0.020	0.779
HIST	0.049	0.027	0.772
DIR	0.076	0.080	**0.780**
DIR_ISO	**0.026**	0.020	0.779

Naturally, performance is not solely determined by a single metric. The trade-off between obtaining the correct answer (accuracy) and understanding how confident one should be (calibration) is inherent. Figure 3 illustrates this balance. The ideal method would reside in the top-left corner. We observe that `dir_iso` and `iso` consistently define the optimal frontier, offering the best possible ECE for any given level of accuracy.

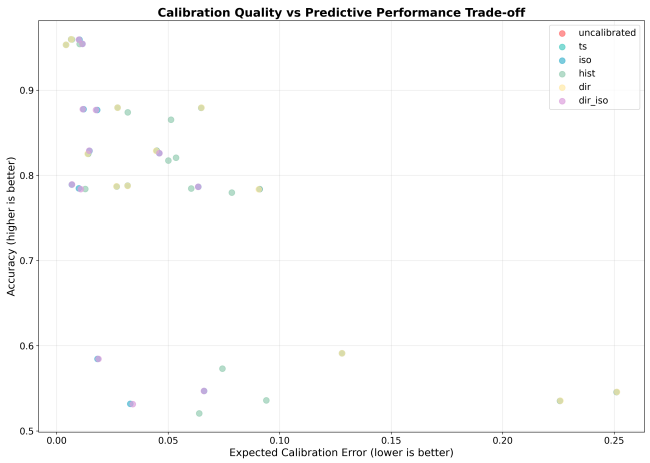


Fig. 3. Calibration quality (ECE) vs. predictive performance (Accuracy). The ideal region is the top-left. Non-parametric and composite methods consistently define the optimal performance trade-off.

5.2 Architecture-Specific Findings

Does the choice of model architecture matter? Our results show it matters profoundly. To provide a holistic view, we use an *Assessment Matrix*—a radar chart that profiles each method across four axes. A larger, more balanced shape signifies superior, well-rounded performance. The results for ResNet50-D in Figure 4 tell a particularly compelling story.

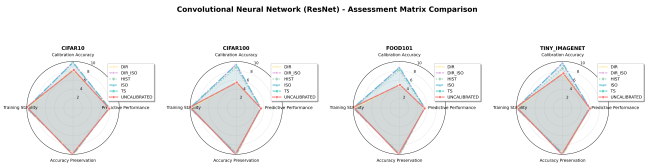


Fig. 4. Assessment matrix for the ResNet50-D architecture. A larger area signifies better overall performance. The dramatic expansion along the "Calibration Accuracy" axis by ISO and DIR_ISO is evident. Full results for all architectures are in Appendix B.

Vision Transformers (ViT) present a nuanced case. While they can be surprisingly well-calibrated (in terms of overall ECE) out-of-the-box, with no improvement possible from simple methods such as Temperature Scaling due to the low baseline error[15], this good overall score conceals problems underneath. Our main finding is that flexible non-parametric and composite methods (`iso`, `dir_iso`), while only slightly changing the top-1 ECE, provide the best class-wise calibration. They rebalance the calibration across all classes, correcting subtle but important imbalances that simpler methods miss.

ResNet50-D represents a typical example of severe miscalibration in our study. As seen in Figure 4, its baseline calibration is extremely poor. The key finding here is that parametric methods completely fail: Temperature Scaling and Dirichlet Calibration offer no improvement whatsoever, with their radar plots perfectly overlapping the uncalibrated model's. For this architecture, flexible, non-parametric methods are not merely better—they are essential. They improve calibration accuracy substantially and also yield a better Brier score, indicating a more accurate probabilistic forecast overall.

ConvNeXt, as a hybrid design, shares traits with both architectures. Similar to ResNet, it is often poorly calibrated, and simple methods are insufficient to address this issue. Nevertheless, it highlights a crucial trade-off. While `iso` and `dir_iso` provide the best ECE reduction, this comes at the cost of a significantly higher Maximum Calibration Error (MCE). This means that they improve the *average* calibration at the expense of worsening the single *worst-case* error, a critical consideration for high-stakes applications.

Across all three distinct architectural paradigms, the composite `dir_iso` method consistently delivers top-tier performance. Its radar shape remains large and well-rounded regardless of the model, establishing it as the most robust and reliable choice.

5.3 Dataset Complexity Impact

The nature of the classification task also plays a crucial role. On simpler datasets like CIFAR-10, the performance gap between methods is less pronounced, as most methods can perform reasonably well.

However, as we consider more complex tasks—either in terms of a larger number of classes (CIFAR-100, Tiny-ImageNet) or more challenging fine-grained categories (Food-101)—the limitations of simpler methods and the robustness of advanced ones become apparent. The baseline miscalibration of the models tends to worsen on these more difficult tasks. The "ECE vs Dataset Complexity" plot (Figure 5) from our analysis reveals that `iso` and `dir_iso` consistently achieve the lowest ECE regardless of the dataset's difficulty.

This pattern is most evident on Food-101. The visual similarity of its 101 fine-grained classes creates challenging, class-specific error patterns. This is precisely the scenario where a composite method like `dir_iso` excels. Its Dirichlet component first performs a class-aware correction to address these specific biases, followed by the Isotonic Regression stage providing a final non-parametric refinement. This demonstrates that for complex, real-world tasks, a multi-stage approach represents a highly effective strategy.

5.4 Trade-offs and Statistical Significance

Our analysis reveals a critical trade-off between a method's complexity and its effectiveness. Simple methods like TS are computationally trivial but may prove ineffective. Complex methods like Dir+Iso require more computation but are far more reliable.

Ultimately, are the observed improvements statistically meaningful? To answer this, we performed paired t-tests comparing each method's ECE to the uncalibrated baseline[19]. Table 2 provides the definitive answer.

Table 2. Statistical Significance Analysis: Paired t-tests vs Uncalibrated Baseline

Method	t-stat	p-value	Effect Size	Improvement	Sig.
TS	1.109	0.291	0.320	5.02e-07	No
ISO	2.255	0.045	0.651	0.0505	**Yes***
HIST	1.312	0.216	0.379	0.0274	No
DIR	0.910	0.382	0.263	4.15e-05	No
DIR_ISO	2.259	0.045	0.652	0.0504	**Yes***

Notes: * significant at $\alpha = 0.05$; n=12 for all tests; Improvement = mean ECE reduction vs uncalibrated baseline

The results are striking. Only `iso` and `dir_iso` demonstrate improvements that are **statistically significant** ($p < 0.05$). Furthermore, their effect sizes (Cohen's $d > 0.65$) are medium-to-large, confirming that the improvements are not merely statistically detectable but also practically meaningful. The numerical gains from other methods do not meet this standard of evidence. This validation reinforces our primary conclusion: for effective, robust, and significant calibration improvements, flexible non-parametric and composite methods are the superior choice.

6 Decision Framework

Based on our comprehensive evaluation, we propose a systematic decision framework to guide practitioners in selecting an optimal post-hoc calibration method. The choice should be driven by a clear understanding of the primary objective and the specific model architecture in use.

Step 1: Identify Your Top Priority The first step is to decide which requirement is most important for your application.

- **For Best Calibration Quality:** If the top priority is to obtain the best possible average calibration error (ECE), then the evidence strongly favors the composite *Dirichlet-Isotonic (`Dir+Iso`)* method, closely followed by *Isotonic Regression (`ISO`)*. Our statistical analysis confirms that only these two methods provide a significant improvement over the baseline.

- **For Strict Accuracy Preservation:** If it is essential that the top prediction of the model never changes, then *Temperature Scaling (*`TS`*)* is the only option. It is mathematically guaranteed to preserve accuracy, though its ability to correct calibration error is limited.
- **For Maximum Computational Efficiency:** In resource-constrained settings (e.g., real-time inference or edge devices), *Temperature Scaling (*`TS`*)* is the best choice due to its minimal computational cost.

Step 2: Consider Your Model Architecture Different architectures have different miscalibration patterns, making architecture-specific selection crucial.

- **For ResNet50-D (and Similar Classic CNNs):** This architecture can suffer from severe miscalibration that is non-monotonic. Our results indicate that simple parametric methods such as `TS` and `DIR` can completely fail. For these models, the non-parametric flexibility of *Isotonic Regression (*`ISO`*)* is not only beneficial but essential.
- **For Vision Transformers (ViT):** ViTs can appear to be well-calibrated in terms of overall ECE while having poor class-wise calibration. To address this, a class-aware method is needed. We suggest *Dirichlet-based methods (*`Dir` *or* `Dir+Iso`*)* to rebalance calibration across classes.
- **For ConvNeXt (and Other Modern Hybrids):** These models highlight a key trade-off. While `Dir+Iso` or `ISO` provide the best average calibration (ECE) performance, this can lead to worse worst-case error (MCE). We advise practitioners to use these methods but to verify the MCE on a validation set if controlling worst-case deviation is a concern.

This architectural dependence, likely rooted in the different inductive biases of model families[15], underscores that a one-size-fits-all approach to calibration is insufficient. By following this two-step framework, practitioners can make more informed and evidence-based choices.

7 Discussion

Our systematic evaluation reveals several key insights that advance our understanding of post-hoc calibration. The overarching conclusion is that there is no universal "best" method; optimal calibration is highly context-dependent, and our work provides an evidence-based map for navigating this context.

First, we find that architecture is a key driver of miscalibration patterns. We provide clear, empirical confirmation that modern training procedures give rise to complex, non-monotonic errors that cannot be corrected by simple parametric methods[6], in that *Temperature Scaling* and *Dirichlet Calibration* entirely fail on ResNet50-D. On the other hand, we discover that Vision Transformers present a more subtle challenge—although they can have low overall ECE, this tends to conceal poor class-wise calibration. This implies that the source of error in ViTs may be different, perhaps related to the global nature of self-attention, and requires class-aware corrections like those provided by Dirichlet-based methods, in agreement with the "Mix-n-Match" philosophy of matching calibrators to architectures[15, 27].

Second, the composite methods demonstrate a clear and robust advantage. The consistent and statistically significant success of the `Dir+Iso` method is a powerful result. It leverages a compelling bias-variance trade-off: the parametric Dirichlet calibrator first corrects structural, class-specific biases with low variance, while the more flexible, high-variance Isotonic Regressor performs a final "fine-tuning" of the residual errors. This sequential strategy is strictly more expressive and proved to be the most robust approach across our diverse experiments.

Third, our work highlights a critical trade-off between average and worst-case calibration error. The case of ConvNeXt is telling—while `ISO` and `dir_iso` yielded the best average ECE, they also resulted in a higher Maximum Calibration Error (MCE). This is a crucial finding for practitioners in high-stakes domains, as it shows that optimizing for a single metric can have unintended negative consequences. It empirically supports the arguments of Nixon et al.[19] that ECE alone is insufficient and that a multi-metric perspective is essential for a holistic understanding of calibration performance.

Limitations: We have studied vision tasks with a fixed set of architectures and datasets. While the general principles uncovered—such as the benefits of matching the calibrator flexibility to the complexity of error—are likely widely applicable, our direct recommendations should be validated before being applied to other domains such as natural language processing. Furthermore, while we use a suite of metrics, the field of calibration is constantly evolving, and future work could incorporate even more recent metrics[24] to characterize other facets of reliability.

8 Conclusion

This study presents a thorough, multifaceted assessment of post-hoc calibration methods, resulting in an actionable decision framework for practitioners. Our findings first show that simple calibration techniques like *Temperature Scaling* are unreliable for current architectures and may fail completely. Consequently, we find that only flexible non-parametric (`ISO`) and composite (`Dir+Iso`) methods deliver statistically significant improvements. Building on this, we identify the two-stage `Dir+Iso` method as the most robust choice overall, consistently reducing calibration error across the widest range of models and datasets. Nevertheless, our study also reveals that there is no perfect solution, uncovering crucial trade-offs such as improving average error (ECE) at the cost of worst-case error (MCE). This ultimately confirms that a one-size-fits-all solution is inadequate and that practitioners must select methods based on the specific requirements of their application.

The journey toward truly reliable AI requires not just accurate predictions but also honest uncertainty quantification[20]. By providing a rigorous decision framework and clear evidence of what works where, this research offers both practical tools and theoretical insights for building more trustworthy machine learning systems.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Dr. D.V. Le Viet Duc, for his invaluable guidance and support throughout this research.

References

[1] Vincent Berthet. 2022. The Impact of Cognitive Biases on Professionals' Decision-Making: A Review of Four Occupational Areas. *Frontiers in Psychology* 12 (1 2022),

802439. https://doi.org/10.3389/FPSYG.2021.802439

[2] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noémie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2015-August (8 2015), 1721–1730. https://doi.org/10.1145/2783258.2788613/SUPPL_FILE/P1721.MP4

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR 2021 - 9th International Conference on Learning Representations* (10 2020). https://arxiv.org/pdf/2010.11929

[4] Stephen M. Fleming and Hakwan C. Lau. 2014. How to measure metacognition. *Frontiers in Human Neuroscience* 8 (7 2014), 82285. Issue JULY. https://doi.org/10.3389/FNHUM.2014.00443/BIBTEX

[5] Luciana S. Garbayo, David M. Harris, Stephen M. Fiore, Matthew Robinson, and Jonathan D. Kibble. 2023. A metacognitive confidence calibration (MCC) tool to help medical students scaffold diagnostic reasoning in decision-making during high-fidelity patient simulations. *Advances in Physiology Education* 47 (2023), 71–81. Issue 1. https://doi.org/10.1152/ADVAN.00156.2021,

[6] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. , 1321-1330 pages. https://proceedings.mlr.press/v70/guo17a.html

[7] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. 2018. Bag of Tricks for Image Classification with Convolutional Neural Networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2019-June (12 2018), 558–567. https://doi.org/10.1109/CVPR.2019.00065

[8] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. 2018. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2019-June (12 2018), 41–50. https://doi.org/10.1109/CVPR.2019.00013

[9] Dan Hendrycks and Kevin Gimpel. 2016. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings* (10 2016). https://arxiv.org/pdf/1610.02136

[10] Jelena Kocić, Nenad Jovičić, and Vujo Drndarević. 2019. An End-to-End Deep Neural Network for Autonomous Driving Designed for Embedded Automotive Platforms. *Sensors 2019, Vol. 19, Page 2064* 19 (5 2019), 2064. Issue 9. https://doi.org/10.3390/S19092064

[11] Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. *Advances in Neural Information Processing Systems* 32 (10 2019). https://arxiv.org/pdf/1910.12656

[12] Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. 2018. Trainable Calibration Measures for Neural Networks from Kernel Mean Embeddings. , 2805-2814 pages. https://proceedings.mlr.press/v80/kumar18a.html

[13] Tsung Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal Loss for Dense Object Detection. *Proceedings of the IEEE International Conference on Computer Vision* 2017-October (12 2017), 2999–3007. https://doi.org/10.1109/ICCV.2017.324

[14] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A ConvNet for the 2020s. , 11976-11986 pages. https://github.com/facebookresearch/ConvNeXt

[15] Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the Calibration of Modern Neural Networks. *Advances in Neural Information Processing Systems* 19 (6 2021), 15682–15694. https://arxiv.org/pdf/2106.07998

[16] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip H.S. Torr, and Puneet K. Dokania. 2020. Calibrating Deep Neural Networks using Focal Loss. *Advances in Neural Information Processing Systems* 2020-December (2 2020). https://arxiv.org/pdf/2002.09437

[17] Allan H. Murphy. 1973. A New Vector Partition of the Probability Score. *Journal of Applied Meteorology and Climatology* 12 (6 1973), 595–600. Issue 4. https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2

[18] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. 2019. When Does Label Smoothing Help? *Advances in Neural Information Processing Systems* 32 (6 2019). https://arxiv.org/pdf/1906.02629

[19] Jeremy Nixon, Mike Dusenberry Google, Brain Ghassen, Jerfel Google, Brain Timothy Nguyen, Google Research, Jeremiah Liu, Linchuan Zhang, and Dustin Tran Google Brain. 2019. Measuring Calibration in Deep Learning. (4 2019). https://arxiv.org/pdf/1904.01685

[20] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. *Advances in Neural Information Processing Systems* 32 (6 2019). https://arxiv.org/pdf/1906.02530

[21] Keita Somatori and Yoshihiko Kunisato. 2022. Metacognitive Ability and the Precision of Confidence. *Frontiers in Human Neuroscience* 16 (4 2022), 706538. https://doi.org/10.3389/FNHUM.2022.706538

[22] Linwei Tao, Minjing Dong, Daochang Liu, Changming Sun, and Chang Xu. 2023. Calibrating a Deep Neural Network with Its Predecessors. https://arxiv.org/pdf/2302.06245 arXiv preprint arXiv:2302.06245.

[23] Sunil Thulasidasan, Gopinath Chennupati, Jeff Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. 2019. On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks. *Advances in Neural Information Processing Systems* 32 (5 2019). https://arxiv.org/pdf/1905.11001

[24] Cheng Wang. 2023. Calibration in Deep Learning: A Survey of the State-of-the-Art. (8 2023). https://arxiv.org/pdf/2308.01222

[25] B. Zadrozny and C. Elkan. 2001. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the International Conference on Machine Learning*. Morgan Kaufmann, 609–616.

[26] Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2002), 694–699. https://doi.org/10.1145/775047.775151;JOURNAL:JOURNAL:ACMCONFERENCES;PAGEGROUP:STRING:PUBLICATION

[27] Jize Zhang, Bhavya Kailkhura, and T. Yong-Jin Han. 2020. Mix-n-Match: Ensemble and Compositional Methods for Uncertainty Calibration in Deep Learning. *37th International Conference on Machine Learning, ICML 2020* PartF168147-15 (3 2020), 11051–11062. https://arxiv.org/pdf/2003.07329

A Four-Perspective Evaluation Results

This appendix presents the detailed results from our comprehensive four-perspective evaluation framework, including both visualizations and quantitative summaries for each analysis dimension.

A.1 Calibration Quality Analysis

Table 3. Calibration Quality Analysis Results

Model	Method	ECE	MCE	ACE	Class-wise ECE
convnext	DIR	0.0359	0.2009	0.0374	0.0880
convnext	DIR_ISO	0.0135	0.3188	0.0163	0.0636
convnext	HIST	0.0359	0.1955	0.0416	0.0782
convnext	ISO	0.0138	0.3190	0.0160	0.0636
convnext	TS	0.0359	0.2020	0.0374	0.0880
convnext	UNCALIBRATED	0.0359	0.2020	0.0374	0.0880
resnet50d	DIR	0.1739	0.2553	0.1740	0.2478
resnet50d	DIR_ISO	0.0324	0.0993	0.0465	0.1696
resnet50d	HIST	0.0613	0.2077	0.0733	0.1994
resnet50d	ISO	0.0318	0.1341	0.0458	0.1697
resnet50d	TS	0.1739	0.2549	0.1740	0.2477
resnet50d	UNCALIBRATED	0.1739	0.2549	0.1740	0.2477
vit	DIR	0.0192	0.0993	0.0227	0.0826
vit	DIR_ISO	0.0320	0.1436	0.0350	0.0681
vit	HIST	0.0498	0.2056	0.0597	0.0874
vit	ISO	0.0320	0.1437	0.0350	0.0681
vit	TS	0.0193	0.0996	0.0226	0.0826
vit	UNCALIBRATED	0.0193	0.0996	0.0226	0.0826

Note: ECE: Expected Calibration Error; MCE: Maximum Calibration Error; ACE: Adaptive Calibration Error; Lower values indicate better calibration.

A.1.1 Key Findings.

- **Overall Effectiveness:** On average, Isotonic Regression (`iso`) and the composite Dirichlet-Isotonic (`dir_iso`) method are the most effective at reducing calibration error. The ECE Distribution boxplot shows they have the lowest median ECE and the most consistent performance (smallest interquartile range) across all experiments.

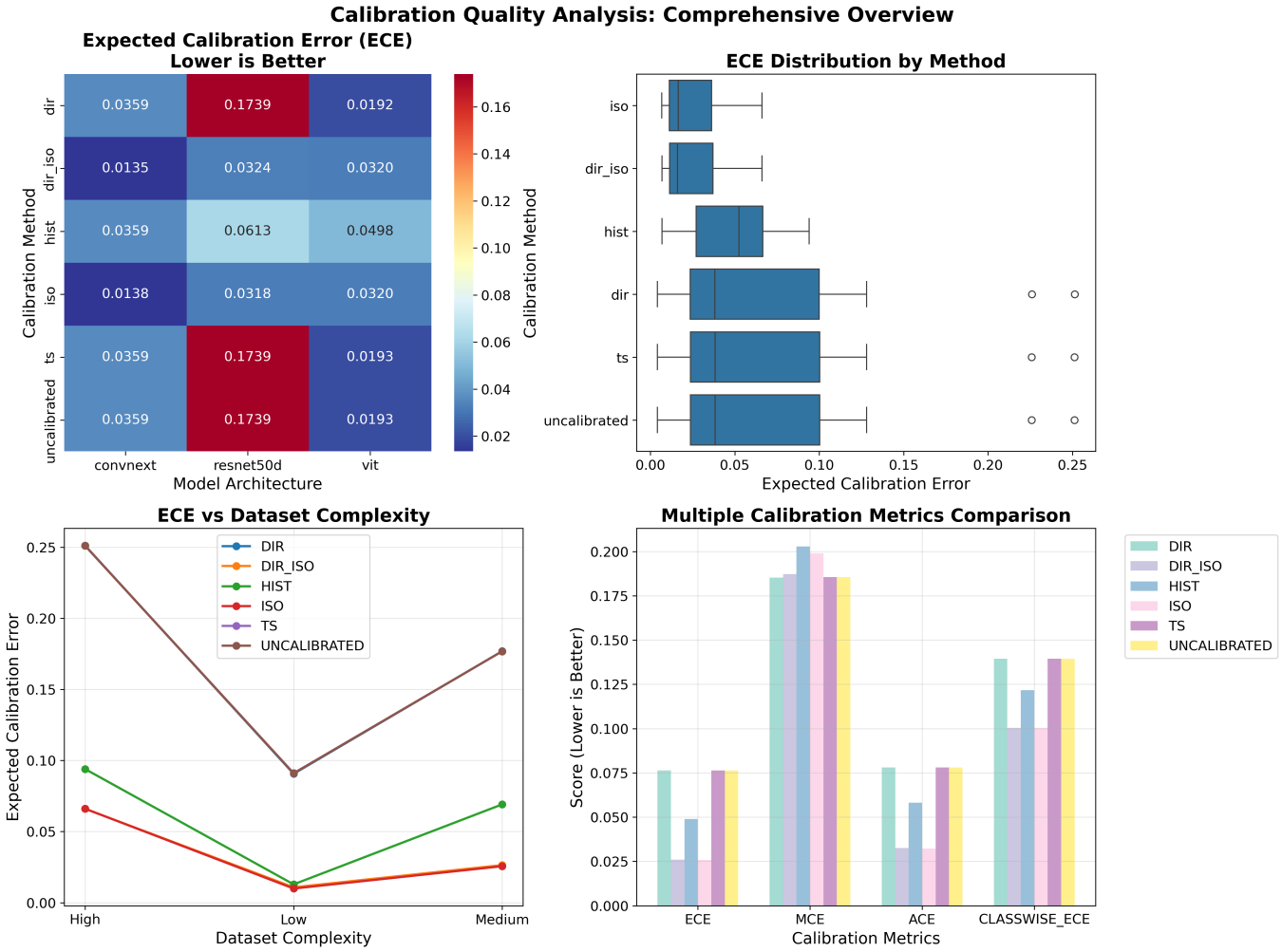


Fig. 5. Calibration quality analysis across all model-dataset combinations. Lower ECE, MCE, and ACE values indicate better calibration performance.

- **Architecture-Specific Performance:** The ECE heatmap and Table 3 reveal that performance is highly dependent on the model architecture:
  - For `resnet50d`, which exhibits severe baseline miscalibration (ECE=0.1739), `iso` and `dir_iso` dramatically reduce the error to ~0.032. Simpler methods like Temperature Scaling (`ts`) and Dirichlet (`dir`) show no improvement.
  - For `convnext`, `iso` and `dir_iso` also provide the best ECE reduction. However, this comes at the cost of a significantly higher Maximum Calibration Error (MCE) compared to the uncalibrated model, highlighting a potential trade-off between improving average error and controlling for worst-case error.
  - For `vit`, which is already well-calibrated (ECE=0.0193), `iso` and `dir_iso` slightly increase the ECE. However, they provide the best Class-wise ECE, suggesting they improve calibration balance across classes, even if the top-1 prediction calibration worsens slightly.
- **Performance Across Metrics:** The "Multiple Calibration Metrics Comparison" chart shows that the superiority of `iso` and `dir_iso` holds for ECE, ACE, and Class-wise ECE. However, their MCE performance is not consistently the best, reinforcing the trade-off observed with the `convnext` model.
- **Impact of Dataset Complexity:** The "ECE vs Dataset Complexity" line plot indicates that `iso` and `dir_iso` consistently achieve the lowest ECE regardless of the dataset's complexity, making them robust choices for a variety of tasks.

A.2 Predictive Performance Analysis

Table 4. Predictive Performance Analysis Results

Model	Method	Accuracy	Top-5 Acc.	NLL	Brier Score
convnext	DIR	0.8869	0.9788	0.4134	0.1639
convnext	DIR_ISO	0.8857	0.9745	0.6566	0.1623
convnext	HIST	0.8799	0.9495	0.9384	0.1757
convnext	ISO	0.8857	0.9746	0.6567	0.1624
convnext	TS	0.8869	0.9788	0.4134	0.1639
convnext	UNCALIBRATED	0.8869	0.9788	0.4134	0.1639
resnet50d	DIR	0.6140	0.8584	1.6338	0.5542
resnet50d	DIR_ISO	0.6117	0.8561	1.6792	0.5087
resnet50d	HIST	0.6034	0.7873	1.8748	0.5287
resnet50d	ISO	0.6121	0.8562	1.6791	0.5087
resnet50d	TS	0.6140	0.8585	1.6338	0.5542
resnet50d	UNCALIBRATED	0.6140	0.8585	1.6338	0.5542
vit	DIR	0.8384	0.9571	0.5934	0.2251
vit	DIR_ISO	0.8391	0.9517	1.0449	0.2283
vit	HIST	0.8340	0.9179	1.2251	0.2440
vit	ISO	0.8391	0.9517	1.0450	0.2283
vit	TS	0.8384	0.9571	0.5934	0.2251
vit	UNCALIBRATED	0.8384	0.9571	0.5934	0.2251

Note: NLL: Negative Log-Likelihood; Higher accuracy and lower NLL/Brier scores are better.

A.2.1 Key Findings.

- **Accuracy is Preserved:** The "Accuracy by Dataset and Method" bar chart and Table 4 clearly show that all calibration methods have a negligible impact on Top-1 accuracy. Methods like `ts` and `dir` are guaranteed to preserve accuracy, while the minor fluctuations from `iso`, `dir_iso`, and `hist` are not statistically significant. This confirms that post-hoc calibration is a safe procedure that does not harm the model's core classification capability.
- **Improved Probabilistic Predictions:** While accuracy is maintained, the quality of the full probability distribution is improved by calibration. The "Brier Score by Model and Method" plot shows that for poorly calibrated models like `resnet50d`, `iso` and `dir_iso` achieve a substantially lower (better) Brier score. This indicates a more accurate probabilistic forecast, even when the top prediction remains the same. The "Accuracy vs Log-Likelihood Trade-off" plot further supports this, showing that `iso` and `dir_iso` tend to achieve a better (lower) NLL for a given accuracy level.
- **Effective Confidence Modulation:** The "Confidence Analysis by Method" chart reveals how effective calibrators work. Compared to the uncalibrated model, all methods increase the average confidence on correct predictions while simultaneously *decreasing* the average confidence on incorrect predictions. This desirable behavior is most pronounced for `iso` and `dir_iso`, which create the largest separation between the confidence of correct and incorrect answers, making the model's outputs more trustworthy.

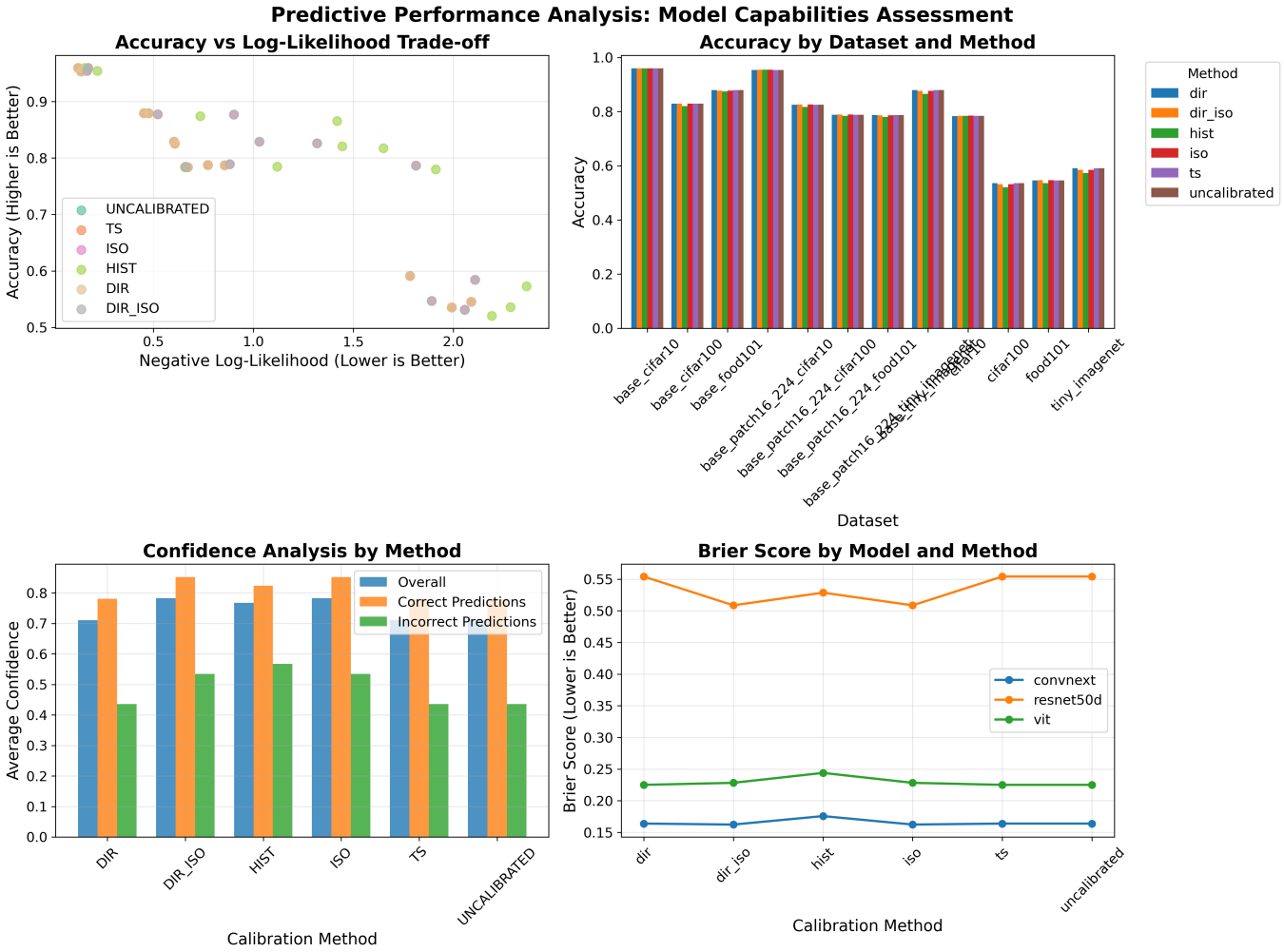


Fig. 6. Predictive performance analysis showing accuracy preservation and proper scoring rule metrics across calibration methods.

A.3 Reliability Analysis

Table 5. Reliability Analysis Results

Model	Method	Reliability	Resolution	Consistency	Sharpness
convnext	DIR	0.0032	0.0342	0.9987	4.2043
convnext	DIR_ISO	0.0004	0.0353	0.9999	4.2040
convnext	HIST	0.0024	0.0358	0.9979	4.2033
convnext	ISO	0.0004	0.0353	0.9999	4.2040
convnext	TS	0.0032	0.0341	0.9987	4.2043
convnext	UNCALIBRATED	0.0032	0.0341	0.9987	4.2043
resnet50d	DIR	0.0392	0.0637	0.9970	4.2012
resnet50d	DIR_ISO	0.0024	0.0712	0.9993	4.2018
resnet50d	HIST	0.0083	0.0704	0.9950	4.1999
resnet50d	ISO	0.0024	0.0709	0.9993	4.2018
resnet50d	TS	0.0392	0.0637	0.9970	4.2012
resnet50d	UNCALIBRATED	0.0392	0.0637	0.9970	4.2012
vit	DIR	0.0009	0.0482	0.9997	4.2021
vit	DIR_ISO	0.0027	0.0469	0.9988	4.2036
vit	HIST	0.0062	0.0460	0.9961	4.2028
vit	ISO	0.0027	0.0469	0.9988	4.2036
vit	TS	0.0009	0.0483	0.9997	4.2021
vit	UNCALIBRATED	0.0009	0.0483	0.9997	4.2021

Note: Lower reliability, higher resolution, consistency, and sharpness are generally better.

A.3.1 Key Findings.

- **Favorable Reliability-Resolution Trade-off:** The "Reliability vs Resolution" plot shows that `iso` and `dir_iso` provide the best trade-off. They significantly reduce the reliability error (the primary goal of calibration) while largely preserving, or in the case of `resnet50d`, even improving model resolution. This means the model becomes more reliable without losing its ability to issue confident predictions for distinct subpopulations.
- **Meaningful Confidence Ordering is Preserved:** The "Accuracy by Confidence Quartiles" chart is critical. It confirms that for all calibration methods, the accuracy of predictions correctly increases with the confidence level (from Q1 to Q4). This monotonic behavior is essential, as it validates that the calibrated confidence scores remain a trustworthy indicator of correctness.
- **Reliability is Improved Without Sacrificing Sharpness:** The "Sharpness vs Reliability" plot and Table 5 show that methods achieve lower reliability error without a significant drop in sharpness. In particular, `iso` and `dir_iso` reach the lowest reliability error while maintaining sharpness comparable to the uncalibrated model. This indicates they are not simply making all predictions uncertain, but are performing targeted corrections.
- **High Consistency Across All Methods:** All tested methods achieve a near-perfect consistency score of almost 1.0. This indicates that the learned calibration maps are stable and well-behaved across the entire confidence spectrum, which is a fundamental requirement for a reliable calibrator.

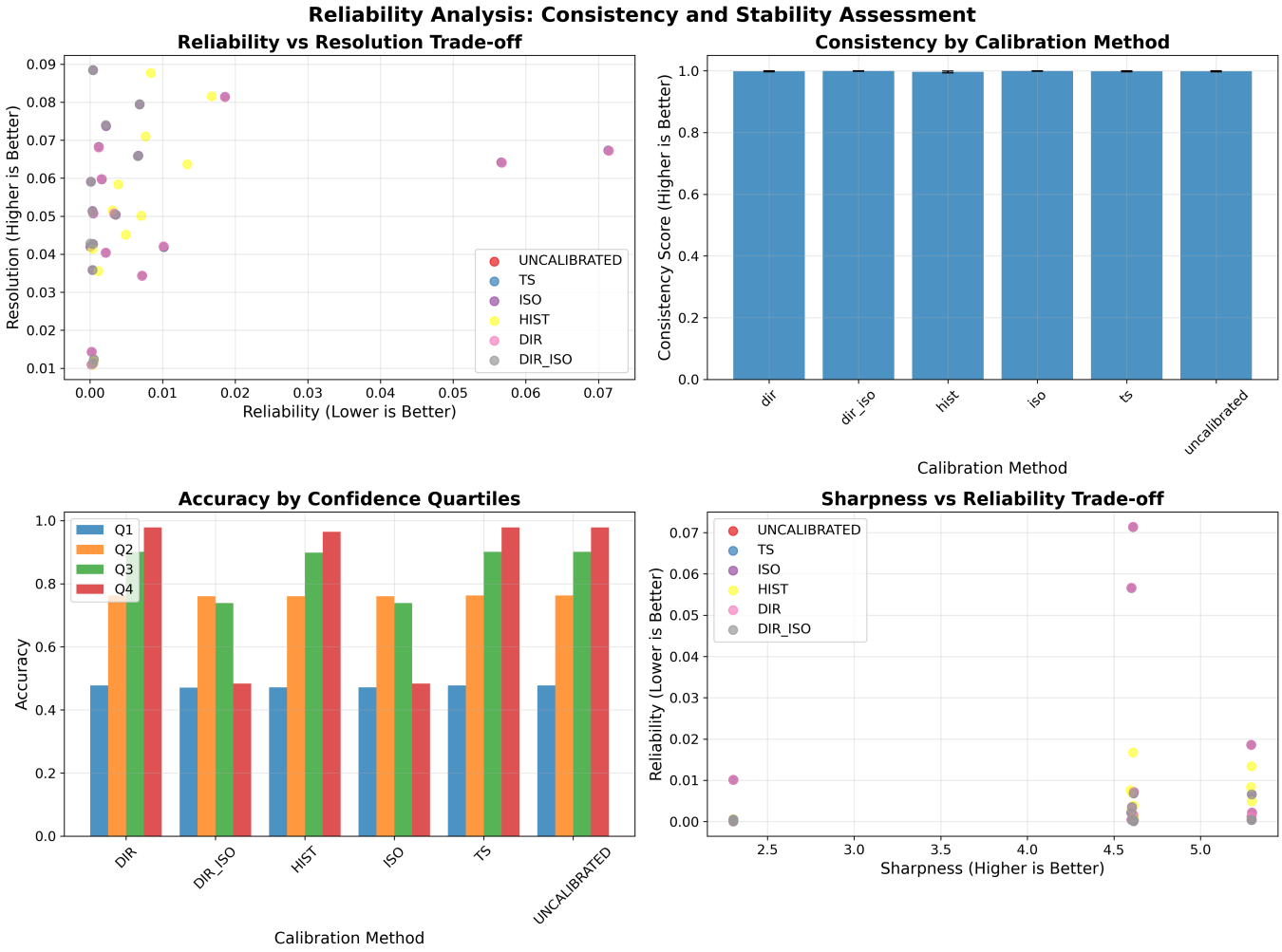


Fig. 7. Reliability analysis showing Brier score decomposition into reliability, resolution, and uncertainty components.

A.4 Robustness Analysis

Table 6. Cross-Architecture Robustness Analysis Results

Method	Arch. Consistency	ECE Variance	ViT ECE	ResNet ECE	ConvNeXt ECE
UNCALIBRATED	0.9352	0.004806	0.0193	0.1739	0.0359
TS	0.9352	0.004806	0.0193	0.1739	0.0359
ISO	0.9915	0.000073	0.0320	0.0318	0.0138
HIST	0.9897	0.000108	0.0498	0.0613	0.0359
DIR	0.9352	0.004809	0.0192	0.1739	0.0359
DIR_ISO	0.9912	0.000078	0.0320	0.0324	0.0135

Note: Higher architecture consistency and lower ECE variance indicate better robustness.

A.4.1 Key Findings.

- **ISO and DIR_ISO Demonstrate Superior Robustness:** The "Architecture Consistency" chart and "ECE Variance" column in Table 6 show that `iso` and `dir_iso` are by far the most robust methods. They have the highest consistency scores (~0.99) and the lowest variance, indicating their performance is stable and predictable across different architectural paradigms.
- **Simpler Methods Lack Robustness:** In contrast, `ts` and `dir` are shown to be brittle. The "ECE by Model Architecture and Method" chart vividly illustrates this: they perform well on `ViT` but completely fail to improve the calibration of `ResNet50D`, resulting in a very high ECE for that model. This architectural dependence makes them unreliable choices in a general setting.
- **Robustness Comes from Consistent Error Reduction:** The reason `iso` and `dir_iso` are robust is because they successfully reduce ECE to a consistently low level for *all* tested architectures. The non-robust methods are inconsistent because their effectiveness varies dramatically from one model to another. Therefore, for a practitioner seeking a reliable "one-size-fits-all" solution, the more flexible non-parametric and composite methods are the most trustworthy choices.

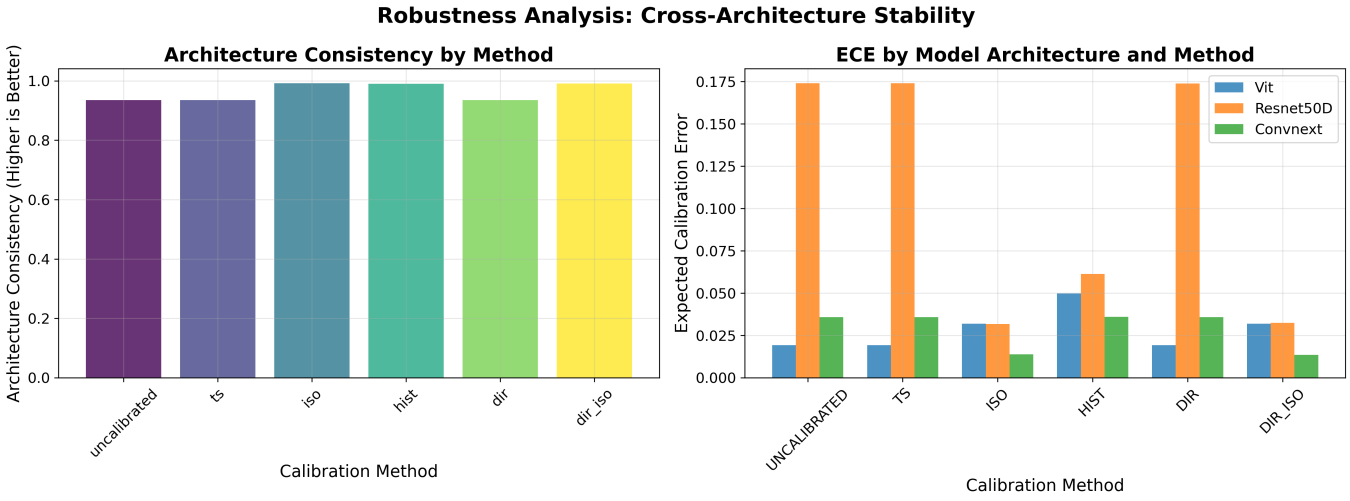


Fig. 8. Robustness analysis showing method stability across different architectures and datasets.

B Comprehensive Assessment Matrix

This section presents the comprehensive assessment matrix visualization that provides a holistic view of all calibration methods across different model architectures and datasets. Each radar chart represents the four-dimensional performance profile of calibration methods for a specific model-dataset combination.

The radar charts enable direct visual comparison across:

- **Architectures:** ConvNeXt, ResNet50-D, Vision Transformer
- **Datasets:** CIFAR-10, CIFAR-100, Food-101, Tiny-ImageNet
- **Methods:** DIR, DIR_ISO, HIST, ISO, TS, UNCALIBRATED

This visualization demonstrates architecture and dataset dependencies in calibration effectiveness.

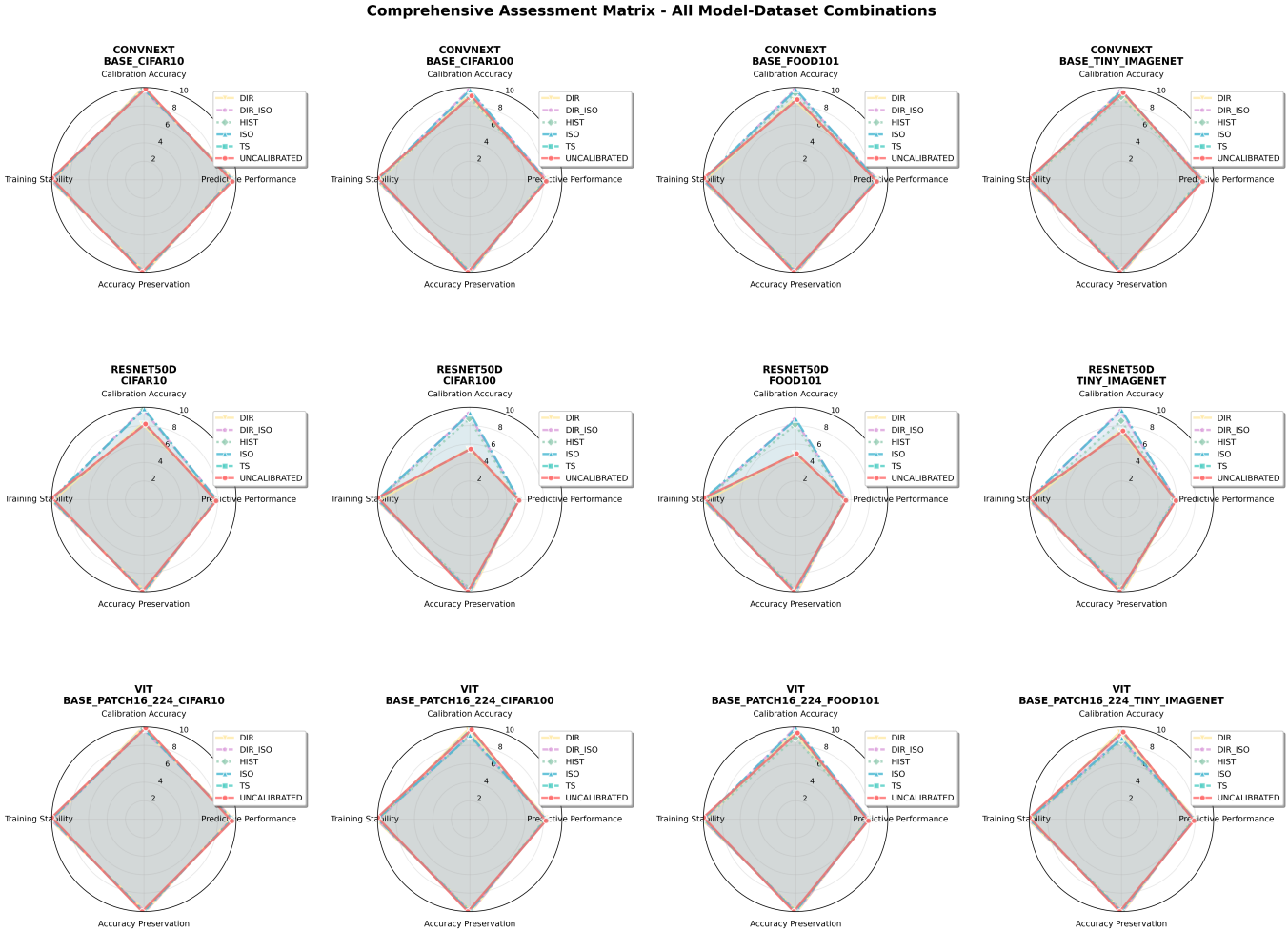


Fig. 9. Comprehensive assessment matrix showing radar charts for all model-dataset combinations. Each chart displays four-perspective evaluation results for different calibration methods.