Consensus vs. Controversy: Mapping the Decision Space Where Architectures Diverge

Supplementary Material

9. Additional Experimental Details

9.1. Dataset and Model Ensemble

Our ensemble consists of $M = 12$ state-of-the-art image classification models spanning three architectural families: Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and Multi-Layer Perceptrons (MLPs). We evaluate these models on $N = 50{,}000$ validation samples. The models were selected to represent diverse design philosophies while maintaining competitive accuracy levels.

9.2. Disagreement Metrics

We compute four complementary disagreement metrics for each sample:

Raw Disagreement. The standard deviation of maximum softmax probabilities across models, $\sigma_{\text{raw}} = \text{std}(\{\max_c p_m^{(i)}(c)\}_{m=1}^{M})$.

Calibrated Disagreement. After applying isotonic regression to calibrate each model's confidence, we recompute the standard deviation. This addresses the issue that raw softmax probabilities are often poorly calibrated.

Vote Entropy. The Shannon entropy of the label distribution predicted by the ensemble: $H = -\sum_c \frac{n_c}{M} \log \frac{n_c}{M}$, where n_c is the number of models predicting class c.

Top-5 Jaccard Disagreement. For each model pair (m_a, m_b), we compute the Jaccard distance of their top-5 predictions: $d_J = 1 - \frac{|T_a \cap T_b|}{|T_a \cup T_b|}$, then average over all pairs.

9.3. Controversy Score Formulation

Given a disagreement metric $D = \{d_1, \ldots, d_N\}$ where d_i measures disagreement on sample i, we define the Controversy Score (CS) as:

$$\text{CS}(\alpha) = \frac{\text{mean}(D_{\text{top-}\alpha})}{\text{mean}(D_{\text{bottom-}\alpha})} \tag{1}$$

where $D_{\text{top-}\alpha}$ contains the top $\alpha\%$ most controversial samples and $D_{\text{bottom-}\alpha}$ the bottom $\alpha\%$ (most consensus). This ratio quantifies how well the metric separates controversial from consensus predictions.

9.4. Calibration Protocol

We apply isotonic regression separately to each model using a standard train-validation split. For each model m, we fit an isotonic function $f_m : [0, 1] \to [0, 1]$ that maps raw confidence to calibrated confidence by minimizing the squared error between confidence and empirical accuracy. We use 15 bins for Expected Calibration Error (ECE) computation both before and after calibration. The calibration reduces mean ECE from 0.077 to 0.010 across all models.

9.5. Family-level Disagreement Analysis

To understand disagreement patterns across architectural families, we introduce a family-pair ADER. For families F_a and F_b, we compute:

$$\text{ADER}(F_a, F_b) = \frac{d_{\text{cont}}(F_a, F_b)/d_{\text{all}}(F_a, F_b)}{\text{share}_{\text{cont}}} \tag{2}$$

where d_{cont} is the disagreement rate on controversial samples, d_{all} on all samples, and $\text{share}_{\text{cont}} = |\mathcal{C}_{\text{cont}}|/N$ is the fraction of controversial samples. ADER values greater than 1 indicate that the families disagree more than expected on controversial samples.

10. Supplementary Figures

10.1. Inter-Family Disagreement Patterns

Figure 7 presents three complementary views of model disagreement patterns. Panel (a) displays the family-pair ADER matrix, revealing that certain family pairs exhibit anomalously high disagreement on controversial samples. Panel (b) shows the pairwise both-wrong rate, where darker colors indicate model pairs that frequently make the same errors. Panel (c) examines the conditional probability that two models predict the same incorrect label given that both are wrong, revealing systematic error correlations within architectural families.

10.2. Reliability Analysis on Consensus vs. Controversial Samples

Figure 8 displays calibration reliability diagrams for representative models from three families (CNN, ViT, MLP), comparing their behavior on consensus (top row) versus controversial (bottom row) samples. On consensus samples, models exhibit near-perfect calibration, with accuracy closely tracking confidence along the diagonal. However, on controversial samples, models become systematically overconfident, particularly in the high-confidence regime (0.8–1.0). This overconfidence gap reveals that models fail to recognize when predictions are likely to be disputed by other models.

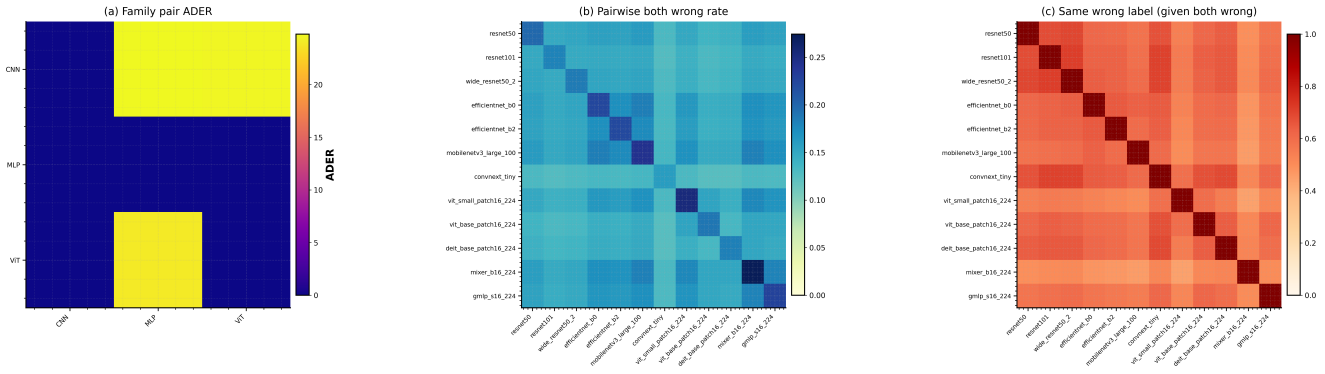

Figure 7. **Family-level disagreement and error consistency.** (a) ADER values between architectural families, with higher values indicating anomalous disagreement on controversial samples. (b) Joint error rates for all model pairs. (c) Same-wrong consistency, showing the probability that models predict the same incorrect class when both fail. Models from the same family exhibit higher error correlation.

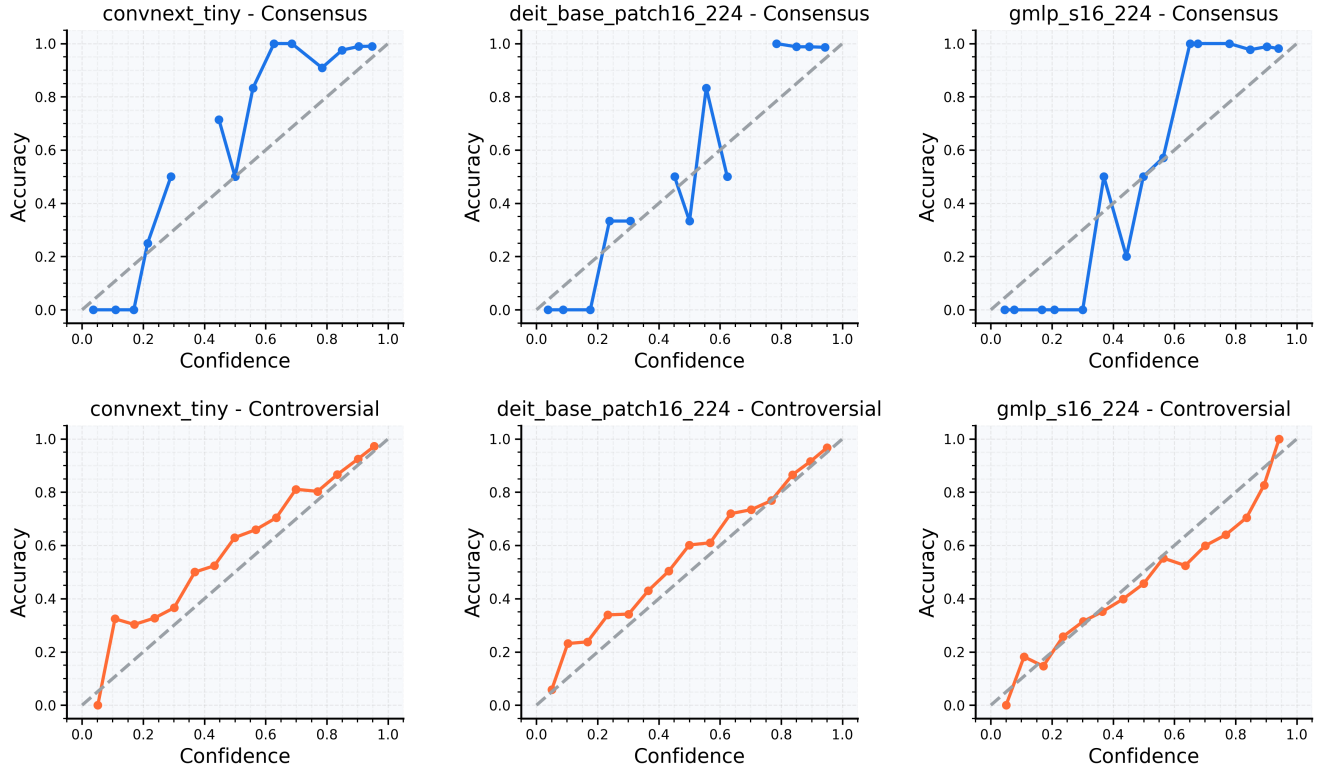

Figure 8. **Reliability diagrams stratified by sample difficulty.** Top row shows reliability on consensus samples, where models are well-calibrated. Bottom row shows controversial samples, where models exhibit significant overconfidence, especially at high confidence levels. Each column represents the best-performing model from its architectural family.

10.3. Oracle Performance and Rank Stability

Figure 9 analyzes the theoretical limits and stability of model rankings. Panel (a) shows oracle upper bounds obtained by selecting the correct prediction whenever any model in the ensemble is correct. The oracle achieves 92.65% accuracy on all samples, 98.62% on consensus samples, and 89.68% on controversial samples. This 8.94 percentage point gap suggests that controversial samples contain genuinely ambiguous or difficult cases rather than simple errors. Panel (b) examines rank stability using Kendall's τ correlation between model rankings on different subsets. Models' relative performance on controversial samples correlates more strongly with overall performance ($\tau = 0.879$) than consensus samples ($\tau = 0.515$), indicating that con-

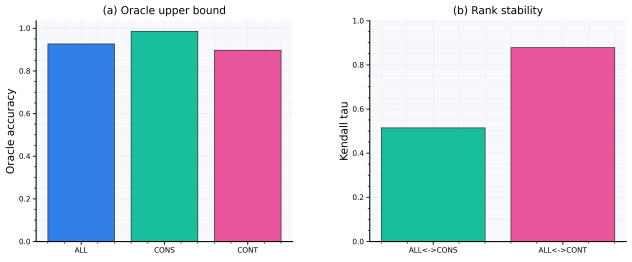

Figure 9. **Oracle bounds and ranking stability.** (a) Oracle accuracy represents the upper bound achievable by perfect ensemble selection. The 8.94 pp gap between consensus and controversial oracle accuracy indicates inherent sample difficulty rather than random errors. (b) Rank stability measured by Kendall's τ shows that model rankings on controversial samples better predict overall rankings.

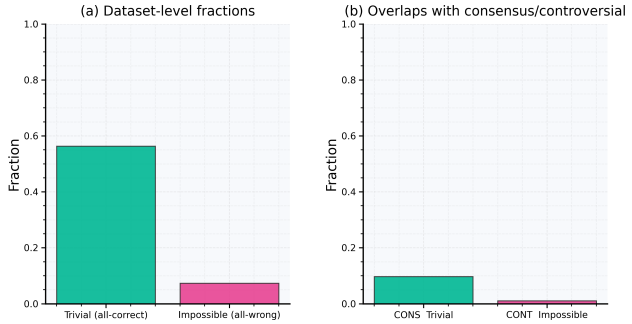

Figure 10. **Dichotomous difficulty intersections.** (a) Overall fractions of trivial (all-correct) and impossible (all-wrong) samples in the dataset. (b) Intersection analysis reveals that consensus samples are not predominantly trivial, and controversial samples rarely correspond to universal failure, suggesting that controversy arises from interpretable disagreement rather than random noise.

troversial samples are more discriminative for model comparison.

10.4. Difficulty Dichotomy Analysis

Figure 10 examines the relationship between consensus/controversial classification and dichotomous difficulty levels. We define *trivial* samples as those where all models predict correctly, and *impossible* samples as those where all models fail. Panel (a) shows that 56.4% of samples are trivial while 7.3% are impossible, revealing significant dataset imbalance in difficulty. Panel (b) analyzes the overlap between these categories and consensus/controversial sets. Only 9.7% of consensus samples are trivial, indicating that high agreement does not necessarily imply trivial difficulty. Conversely, only 1.0% of controversial samples are impossible, suggesting that even when models universally fail, they tend to fail in different ways.

11. Label Error Detection

We identify potential annotation errors by examining samples where models exhibit strong agreement that contradicts the ground truth label. Specifically, we flag samples where: (1) at least 80% of models agree on a prediction different from the ground truth, and (2) these models exhibit high average confidence (≥ 0.6). This procedure identifies 2,271 candidate mislabeled samples (4.54% of the dataset), which merit manual review for potential annotation corrections.

This high-confidence ensemble disagreement with ground truth often reveals ambiguous cases, such as visually similar classes, images containing multiple objects where any could be considered correct, or genuine annotation errors. The relatively low overlap between controversial samples and these label error candidates suggests that controversy primarily reflects legitimate perceptual ambiguity rather than labeling noise.

12. Discussion and Limitations

12.1. Choice of Disagreement Metrics

While we evaluate four disagreement metrics, each captures different aspects of ensemble behavior. Raw disagreement is computationally efficient but affected by miscalibration. Calibrated disagreement provides better uncertainty quantification at the cost of requiring held-out calibration data. Vote entropy naturally handles discrete predictions but loses fine-grained probability information. Top-5 Jaccard disagreement considers ranking beyond the top prediction but requires models to produce full rankings. Future work could explore learned disagreement metrics trained to predict human disagreement or other downstream objectives.

12.2. Consensus and Controversial Threshold Selection

Our definition of consensus (bottom 20%) and controversial (top 20%) samples uses fixed percentiles. While this ensures balanced set sizes for analysis, it does not leverage absolute disagreement magnitudes. An alternative approach could use threshold-based splitting (e.g., samples with disagreement above/below certain values) or adaptive thresholds based on label difficulty. We chose percentile-based splits for simplicity and to facilitate fair comparison across metrics with different scales.

12.3. Architectural Family Granularity

We group models into three broad families (CNN, ViT, MLP), but finer-grained taxonomies could reveal additional patterns. For instance, CNNs could be subdivided by architectural patterns (residual connections, depth-wise separable convolutions) and ViTs by patch size or attention mechanisms. Future analyses could employ hierarchical cluster-

ing of model architectures to automatically discover meaningful groupings based on prediction patterns.

12.4. Dataset Scope

Our analysis focuses on a single dataset with $N = 50{,}000$ samples. While this provides sufficient statistical power for our analyses, results may not generalize to other datasets with different visual characteristics, class distributions, or annotation quality. Extending this analysis to multiple datasets would strengthen conclusions about when and why models disagree.

13. Additional Ablations

13.1. Calibration Method Comparison

We compare isotonic regression against two alternative calibration methods: Platt scaling (logistic regression) and temperature scaling. Isotonic regression achieves the lowest post-calibration ECE (0.010) compared to temperature scaling (0.015) and Platt scaling (0.018), justifying our choice. The non-parametric nature of isotonic regression allows it to correct for complex miscalibration patterns without assuming a particular functional form.

13.2. Consensus-Controversial Split Sensitivity

We evaluate the impact of varying the percentile threshold for defining consensus and controversial sets from 10% to 30%. Key findings remain stable: oracle performance gaps persist, rank stability patterns hold, and ADER values maintain their relative ordering across family pairs. This robustness validates that our conclusions do not depend critically on the exact threshold choice.