BEYOND CONFIDENCE: RELIABLE MODELS SHOULD ALSO QUANTIFY ATYPICALITY

Mert Yuksekgonul[1*], **Linjun Zhang**[2], **James Zou**[3†], **Carlos Guestrin**[1†]
[1]Department of Computer Science, Stanford University
[2]Department of Statistics, Rutgers University
[3]Department of Biomedical Data Science, Stanford University

ABSTRACT

While most machine learning models can provide confidence in their predictions, confidence is insufficient to understand and use the model's uncertainty reliably. For instance, the model may have a low confidence prediction for a sample that is far from the training distribution or is inherently ambiguous. In this work, we investigate the relationship between how atypical (or rare) a sample is and the reliability of a model's confidence for this sample. First, we show that atypicality can predict miscalibration. In particular, we empirically show that predictions for atypical examples are more miscalibrated and overconfident, and support our findings with theoretical insights. Using these insights, we show how being atypicality-aware improves uncertainty quantification. Finally, we give a framework to improve decision-making and show that the atypicality framework improves selectively reporting uncertainty sets. Given these insights, *we propose that models should be equipped not only with confidence but also with an atypicality estimator for reliable uncertainty quantification*. Our results demonstrate that simple post-hoc atypicality estimators can provide significant value.

1 INTRODUCTION

Typicality is an item's resemblance to other category members (Rosch & Mervis, 1975). For example, while a dove and a sparrow are highly typical birds, a penguin is an atypical bird. A large body of cognitive science literature (e.g., Rips (1989); Rips et al. (1973); Mervis & Pani (1980)) suggests that typicality plays a crucial role in the human understanding of categories. For instance, humans have been shown to learn, remember, and refer to typical items a lot faster (Murphy, 2004). Similarly, the representativeness heuristic is the tendency of humans to use the typicality of an event as a basis for decision-making (Tversky & Kahneman, 1974). This cognitive bias is effective in allowing people to make swift decisions, but it can also result in poor judgments of uncertainty. For instance, the likelihood of typical events can be overestimated (Tversky & Kahneman, 1974), or uncertainty judgments can be very poor for *atypical* events (Tversky & Kahneman, 1992).

While it is hard to quantify the uncertainty of human judgments, machine learning models report confidence in their predictions. However, confidence alone can be insufficient to understand the reliability of a prediction. For instance, a low-confidence prediction could have an ambiguity that is easily communicated, or it could be a result of the sample being far from the training distribution. Similarly, a high-confidence prediction could be reliable or miscalibrated. Our main proposal is that *models should quantify not only the confidence but also the atypicality* to understand how reliable predictions are, or to understand the coverage of the training distribution. However, a large volume of machine learning practice works through downloading pretrained models that can report only confidence, without access to a notion of atypicality. To support our position, we use a simple formalization of atypicality estimation. With the following empirical and theoretical studies, we show how simple atypicality estimators improve a model's reliability.

*Corresponding author: merty@stanford.edu
† Equal senior authors.

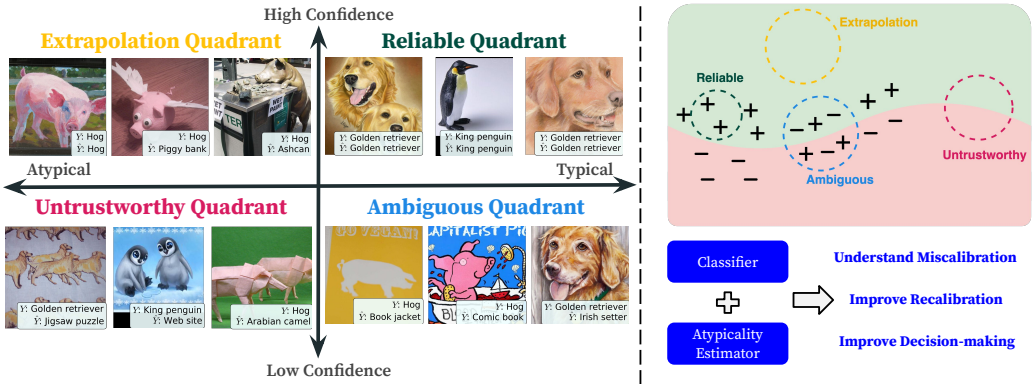

Figure 1: **Atypicality in Uncertainty.** On the left, we show examples from the ImageNet-R dataset with our atypicality framework. On the top right, we provide a conceptualization of each of the quadrants. On the bottom right, we demonstrate the use cases. By using atypicality, we can better understand miscalibration (Section 3), improve calibration (Section 4), and improve decision-making (Section 5).

Understanding Miscalibration: Calibration is a key metric for assessing the quality of probabilistic models (Gneiting & Raftery, 2007), measuring how well the predicted probabilities of a model match the true likelihoods of outcomes. Neural networks (Guo et al., 2017) or even logistic regression (Bai et al., 2021) can be miscalibrated out-of-the-box. Here, we argue that using atypicality can give insights into when to trust a model's confidence. Through extensive experiments, we demonstrate that atypicality is highly correlated with miscalibration. We show that predictions for atypical points are more miscalibrated and overconfident. We support our empirical study with theoretical insights.

Improving Calibration: *Recalibration* methods offer some mitigation to the problem of miscalibration (Platt et al., 1999; Guo et al., 2017). These methods involve adjusting the output of a potentially miscalibrated probabilistic model to improve its calibration. We show that models need different adjustments to the confidence according to the atypicality of individual points, and hence atypicality is a key factor in recalibration. Here, we propose a method called *Atypicality-Aware Recalibration*. Our algorithm takes into account the atypicality of a sample in the recalibration process and is simple to implement. We empirically show that complementing recalibration methods with atypicality improves uncertainty quantification, and support our findings with theoretical insights.

Improving Decision-making: Uncertainty quantification can improve transparency and increase user trust (Bhatt et al., 2021). One way to achieve this is through selective classification (Geifman & El-Yaniv, 2017), where the model can abstain from making a prediction. Another approach leverages conformal prediction (Vovk et al., 2005) which provides the user with uncertainty sets, i.e., a set of classes that contain the true label with high probability. However, these actions are often evaluated independently. We argue that atypicality offers a unifying perspective, and propose Selective Conformal Prediction. We show that by taking atypicality into account, we can better decide when to report uncertainty sets and when to abstain.

Summary of Contributions: Overall, atypicality offers a valuable framework for understanding and improving uncertainty quantification. First, we demonstrate that predictions for atypical examples are more miscalibrated and overconfident. Using our insights, we present how being atypicality aware improves and complements existing recalibration methods. Finally, we use atypicality to selectively report uncertainty sets. Our findings are supported by both experiments and theory. Thus, we propose that **models should also quantify atypicality, and we show simple- and cheap-to-implement atypicality estimators can provide significant value**.

2 Interpreting Uncertainty with Atypicality

Motivation: In many applications of machine learning, we have access to a model's confidence. It aims to quantify the likelihood that a prediction will be accurate. In classification, the model output is a probability distribution over the classes. In most practical scenarios, model confidence is the primary tool used to evaluate the uncertainty of a prediction. However, the uncertainty obtained from

confidence can contain multiple sources of uncertainty (Mukhoti et al., 2021), and distinguishing the source can lead to actionable outcomes such as abstention (Geifman & El-Yaniv, 2017), or data collection (Kirsch et al., 2019). Here, we first argue that model confidence can have different semantics in different cases. We present examples from the Imagenet-R (Hendrycks et al., 2021) dataset for illustration. In Appendix B we discuss how confidence is not enough to distinguish between **high-confidence and representative** and **high-confidence yet far from the support** examples; or having **low confidence due to being far from distribution** and **low confidence due to ambiguity**. In summary, relying solely on model confidence does not provide a complete understanding of the uncertainty in the predictions, and to better interpret the uncertainty we need a better vocabulary. Even when the model is well-calibrated, confidence does not help us distinguish between an ambiguous sample from a sample that is far from the distribution.

Atypicality Provides a Natural Decomposition We propose that *atypicality* provides a natural way to understand uncertainty when combined with confidence. In general, we consider a sample to be typical if it is representative of the previously observed examples. For instance, an image of a dog that closely resembles other dogs that were seen during training would be considered typical. However, if the image is unlike any other observed during training, it would be considered atypical. In the following sections, we will formally define and explain the concept of typicality. We discuss the relation to earlier works in Appendix C.1.

The **Reliable Quadrant** contains ***typical, high-confidence*** examples. Since the sample is typical and the model has high confidence in its prediction, these examples are likely to have higher-quality predictions and uncertainty estimates. The **Extrapolation Quadrant** contains ***atypical, high-confidence*** examples. These samples are farther from the support of the training data, but the model still has high confidence in its predictions. We note that in this case, the model may be poorly calibrated and the uncertainty estimates may not be reliable. The **Untrustworthy Quadrant** comprises ***atypical, low-confidence*** examples. These examples are highly atypical, and the model has low confidence in its predictions. This quadrant can include corrupted examples, such as images with heavy noise or blur, or samples from classes that the model has not been trained on. The **Ambiguous Quadrant** comprises ***typical, low-confidence*** examples. These examples are typical in the sense that they may belong to multiple classes, but due to the inherent ambiguity in the sample, the model has low confidence in its predictions. This quadrant includes samples with inherent ambiguity, such as an image that could be of either a ferret or a polecat, or multi-label examples that contain multiple objects that are from valid classes in an object recognition task.

Formalizing Atypicality We use *atypicality* with respect to a training distribution. Informally, a sample is *atypical* if it is *far* from the training distribution of a model, e.g., if there are no or a limited number of similar examples to a sample, then it can be called atypical. More formally, let $X \in \mathbb{R}^d$ be the random variable denoting features and $Y \in \mathcal{Y} = \{1, \ldots, C\}$ denote the label where we focus on C-class classification settings. We use $a(x)$ to denote the atypicality of a sample x.

Definition 2.1 (Atypicality)**.** For a sample x, we define *the atypicality of a sample as*[1]

$$a(x) = 1 - \max_y \mathbb{P}(X = x | Y = y) \tag{1}$$

For a dog image x, if $\mathbb{P}(X = x | Y = \text{dog})$ is low, then we call x an atypical dog image. If $a(x)$ is high, then we call x an atypical sample. In words, if a sample is not typical for any class, then it is atypical with respect to the training distribution. Similarly, we can also use a notion of distance[2] or marginals $\mathbb{P}(X = x)$ to quantify atypicality. Appendix E discusses *estimation of atypicality*.

3 Understanding Calibration with Atypicality

Here, we demonstrate how our framework can be applied to predict the miscalibration of samples. Let us denote a probabilistic predictor by $\hat{\mathbb{P}}$, where $\hat{\mathbb{P}}(Y = y | X = x)$ represents the predicted probability of an input x belonging to class y. Let us define $\hat{y} = \arg\max_{y \in \mathcal{Y}} \hat{\mathbb{P}}(Y = y | X = x)$ as the predicted class for input x. We describe the datasets and models we use in Appendix F.

[1]This notion of atypicality is different than *typical sets* in information theory (Thomas & Joy, 2006).

[2]For a sample x, if the nearest neighbor distance is large, x is called atypical as samples in the training set are far from x. The density and distance notions of atypicality are connected through non-parametric estimation.

3.1 Model Calibration and Atypicality

First, we use atypicality to understand calibration. Calibration is used to evaluate the quality of a probabilistic model (Gneiting & Raftery, 2007). Informally, a model is considered perfectly calibrated if all events that are predicted to occur $P\%$ of the time occur $P\%$ of the time for any $P \in [0, 100]$.

There are different notions of miscalibration. We employ the following definition: For the sake of simplicity, consider a binary classification problem where $Y \in \{0, 1\}$, and the predictor $\hat{\mathbb{P}} : \mathcal{X} \to [0, 1]$. Calibration Error (CE) is defined as: $\text{CE}[\hat{\mathbb{P}}] = \mathbb{E}[|\mathbb{P}(Y|\hat{\mathbb{P}}(X) = p) - p|]$. However, it is computationally infeasible to calculate the above expectation with the conditional probability $\mathbb{P}(Y|\hat{\mathbb{P}}(X) = p)$. In practice Expected Calibration Error (ECE) is used to estimate CE, see Appendix G.1.

Here, we aim to examine the relationship between model calibration and atypicality. Given any $K > 1$, we consider the quantiles of $a(X)$, $a_1, a_2, \ldots, a_{K+1}$ such that $\mathbb{P}(a(X) \in (a_k, a_{k+1}]) = 1/K$ for $k \in [K]$. Specifically, we investigate the atypicality-conditional calibration error $\text{ECE}[\hat{\mathbb{P}} \mid a(X) \in (a_k, a_{k+1}]]$, i.e., the expected calibration error of a sample that falls within the atypicality quantile k.

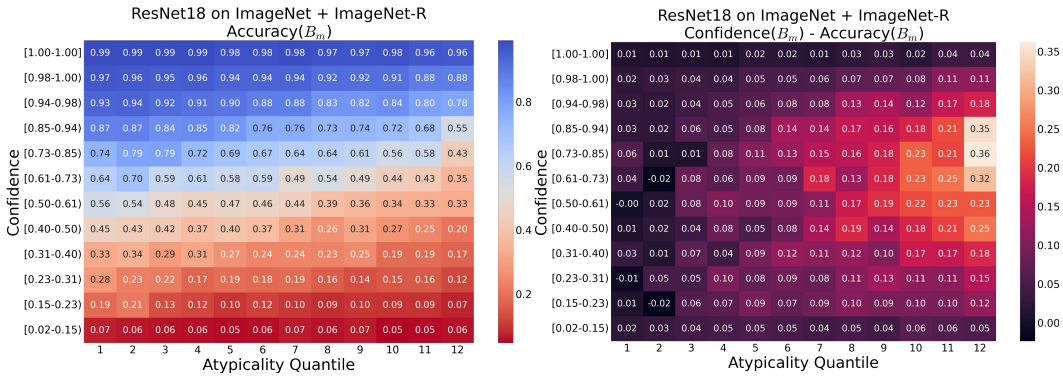

Figure 2: **Miscalibration Distribution Across Confidence and Atypicality.** Here, points are grouped according to the Atypicality Quantile (x-axis) and Confidence Quantiles (y-axis). On the left, values show the accuracy within a bin. On the right, values show the difference between the confidence and the accuracy for a bin, lighter color indicates higher overconfidence. We observe that within the same confidence range, atypical groups have larger miscalibration rates and are more overconfident.

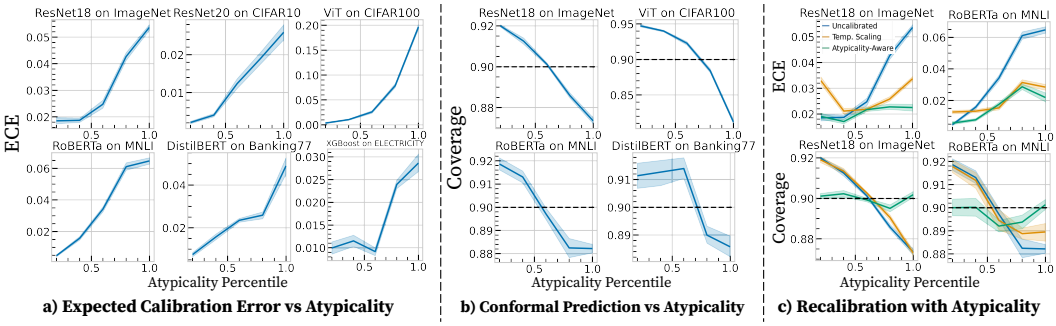

Figure 3: **Atypicality for Reliable Uncertainty. a) Expected Calibration Error vs Atypicality.** We observe that atypical examples are poorly calibrated compared to typical examples. **b) Conformal Prediction vs Atypicality** We observe that atypical examples have worse uncertainty sets where the coverage is not satisfied compared to typical examples. The dashed line shows the 90% marginal coverage guarantee. **c) Recalibration with Atypicality** We observe that with atypicality-aware recalibration, we can improve the calibration of models and obtain more uniform coverage rates.

Atypical Examples are Poorly Calibrated: In Figure 6, we show the distribution of miscalibration where each bin within the grid contains the intersection of the corresponding confidence and atypical-

ity quantiles. We observe that within the same confidence range, predictions for atypical points have lower accuracies and are more miscalibrated and overconfident.

In Figure 3a, we present the miscalibration (ECE) analyses. In these experiments, we split samples into 5 quantiles according to atypicality and compute the expected calibration error for samples within each group. The results show a monotonically increasing relationship between atypicality and ECE. Specifically, we see that predictions for atypical examples are more miscalibrated compared to typical samples. In Appendix Figure 7, we present the reliability diagrams decomposed by atypicality. *Predictions are more overconfident for atypical points; and our Theorem 3.1 supports this finding*.

Overall, these results demonstrate that atypicality can predict calibration. This finding has several practical implications. First, having an atypicality estimator can help us identify samples where probabilistic predictors are less reliable. Furthermore, it suggests that predictions for minority groups (groups with lower probability mass) are likely to be more miscalibrated, and this violates the sufficiency criterion in the fairness literature (Barocas et al., 2017).

3.2 Conformal Prediction: Atypical Samples Have Worse Uncertainty Sets

Conformal Prediction (Angelopoulos & Bates, 2021) is a framework that assigns a calibrated uncertainty set to each instance. Given an instance $X \in \mathcal{X}$ and the true label $Y \in \mathcal{Y}$, the goal is to find a function $\mathcal{C} : \mathcal{X} \to 2^{\mathcal{Y}}$ that returns a subset of the label space such that $Y \in \mathcal{C}(X)$ with high probability. The framework guarantees *marginal coverage*, i.e. $\mathbb{P}(Y \in \mathcal{C}(X)) \geq 1 - \alpha$, for a choice of α. See Appendix H.2 for further details.

Conditional Coverage quantifies the coverage for a group, i.e., $\mathbb{P}(Y \in \mathcal{C}(X) \mid X \in G)$. While conformal prediction provides marginal coverage guarantees, it may not guarantee the same for specific groups, lacking conditional coverage guarantees. Lu et al. (2022) showed that conformal prediction for skin lesion classifiers does not satisfy coverage when conditioned on groups of skin color. Here, we investigate how conformal prediction performs with respect to atypicality, and focus on analyzing *atypicality-conditional coverage*: $\mathbb{P}(Y \in \mathcal{C}(X) \mid a(X) \in (a_{k-1}, a_k]) \geq 1 - \alpha$.

Results: In Figure 3b, we present the results. For each experiment, we fit a threshold using the calibration set and produced prediction sets for each of the samples in the test set. Next, similar to previous experiments, we split samples into atypicality quantiles and compute the coverage for that group. The results show a monotonically decreasing relationship between atypicality and coverage. Specifically, we observe that while typical examples satisfy the conditional coverage criterion, atypical examples do not perform as well. See Tables 3,4 for the results in tabular format.

3.3 Theoretical Analysis: Characterizing Calibration Error with Atypicality

Here, we aim to characterize the calibration error with atypicality. We build on the results from Bai et al. (2021) and extend the analyses for atypicality-conditional calibration. We analyze Logistic Regression with the Gaussian data model and describe the setting more formally in Appendix K.1.

Theorem 3.1. *Consider the data generative model and the learning setting in Appendix K.1. For any $K > 1$, suppose we consider the quantiles of $a(X)$, $a_1, a_2, ..., a_K, a_{K+1}$ such that $\mathbb{P}(a(X) \in (a_k, a_{k+1}]) = 1/K$ for $k \in [K]$. We assume $\|\beta^*\| \leq c_0$, and $d/n = \kappa$, for some sufficiently small $c_0, \kappa > 0$. Then, for sufficiently large n, for $k = 2, \ldots, K$, we have*

$$\mathbb{E}[u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u) \mid a(X) \in [a_{k-1}, a_k]] > \mathbb{E}[u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u) \mid a(X) \in (a_k, a_{k+1}]]$$

That is, the resulting classification model is over-confident, and the level of over-confidence becomes larger when the data is more atypical (with smaller $a(X)$). In addition, the gap becomes larger for smaller sample sizes n. Proof of the theorem can be found in Appendix K.2.

4 Improving Recalibration with Atypicality

In the following section, we delve into how atypicality can complement and improve post-hoc recalibration methods. Specifically, we will examine Temperature Scaling (TS) (Guo et al., 2017). TS adjusts the probability estimates of a model to improve its calibration. By utilizing our understanding of atypicality, we will show how atypicality awareness helps improve TS and conformal prediction. We empirically and theoretically show that taking atypicality into account improves recalibration.

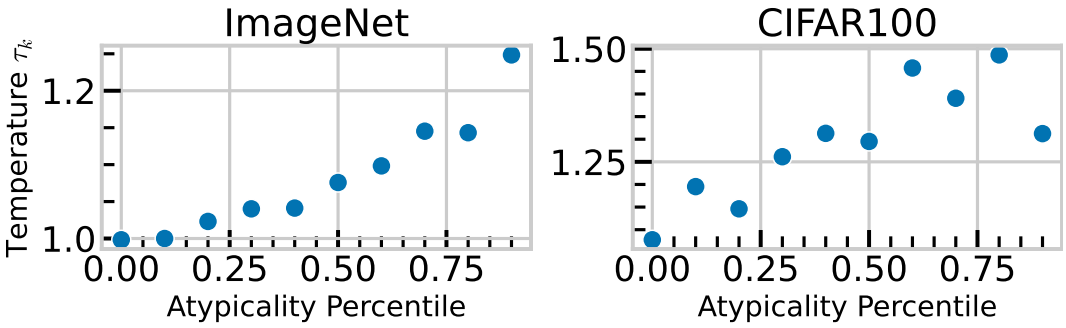

Figure 4: **Fitted Temperature vs Atypicality.** We observe a monotonically increasing relationship between the atypicality of a group and the temperature parameter fitted to that group with TS.

Parametric Recalibration: Different Groups need Different Temperatures. Temperature scaling, a single parameter variant of Platt Scaling (Platt et al., 1999), is a simple recalibration algorithm that calibrates the model using a single temperature parameter. Namely, the predictor is of the form

$$\hat{\mathbb{P}}_{\text{TS}}(X) = \text{Softmax}(f(\boldsymbol{X})/\tau), \tag{2}$$

where $f : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}, \tau \in \mathbb{R}$. f is the model that takes an input and outputs scores/logits, and τ is the recalibration parameter. In practice, τ is optimized using a calibration set to minimize a proper scoring rule such as the cross-entropy loss. We refer the reader to Guo et al. (2017) for more details on TS. Next, we look at the relationship between ECE and atypicality after recalibration.

Results: We separately perform Temperature Scaling on points grouped according to the atypicality quantiles. Namely, we fit a separate temperature parameter to each of the atypicality quantiles, let us denote it by τ_{a_k} for the quantile covering $a(X) \in (a_{k-1}, a_k]$. In Figure 4, we observe an increasing relationship between the values a_k and τ_{a_k}. We observe that different atypicality groups need different adjustments to become calibrated, and more atypical groups need larger temperatures. *This suggests that being atypicality aware can potentially improve calibration. Furthermore, this can mean that while a single temperature value improves average calibration, it may hurt other groups.*

Atypicality-Aware Recalibration. Our findings point us to a simple proposal: Recalibration algorithms should be parameterized not only by the model confidence but also by the atypicality. Generally, we suggest that a recalibration method should take the following form:

$$\hat{\mathbb{P}}_{\text{ATS}}(X) = \mathcal{R}(\hat{\mathbb{P}}(X), a(X)), \tag{3}$$

where the function is monotonic in both of its arguments, which are confidence and atypicality. Concretely, we define a simple implementation with a quadratic function $\tau(a(X)) = c_2 a(X)^2 + c_1 a(X) + c_0$, where $\hat{\mathbb{P}}_{\text{ATS}}(X) = \text{Softmax}(f(\boldsymbol{X})/\tau(a(X)))$. We optimize the parameters c_0, c_1, c_2 over a calibration set to perform post-hoc recalibration. See Appendix H.1 for implementation details.

In Figure 3c, with the top 2 plots we show the ECE analyses, and with the bottom 2 plots we show the results of performing conformal prediction after Atypicality-Aware Recalibration. Overall, we observe that we improve the calibration and coverage rates obtained with existing methods. For instance, we observe that coverage rates are more uniform across typical and atypical points, without explicitly performing conformal prediction across these groups. Similarly, we obtain lower calibration errors across all groups. In Table 1, we present the results in the tabular format. Our results show that atypicality awareness can improve the existing recalibration methods, and in this sense, it complements existing post-hoc approaches. **Theory:** In Appendix Theorem K.2, we analyze atypicality-aware recalibration and show that it achieves lower calibration errors compared to TS.

5 Improving Decision-Making with Atypicality

Reporting uncertainty can improve transparency and increase user trust (Bhatt et al., 2021; Chua et al., 2022). Babbar et al. (2022) reports that the use of uncertainty sets improves user trust and Human+AI performance. In the Selective Classification setting (Geifman & El-Yaniv, 2017), the model can choose when to report a prediction or when to defer to a user. We argue that the atypicality framework gives us a way to unify these actions. Our argument follows from the difference between the

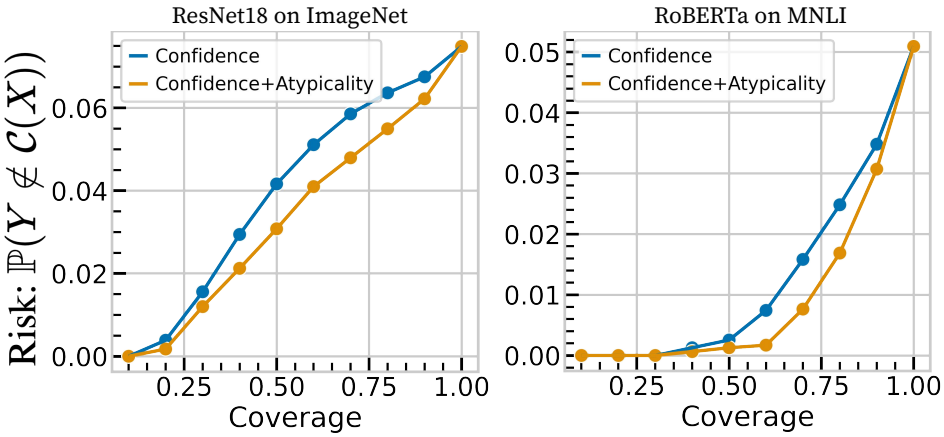

Figure 5: **Selective Conformal Prediction with Atypicality.** Here, coverage is the fraction of samples that were not rejected, and the risk is the average number of samples where the true label was not in the uncertainty set. We observe that using atypicality provides better Risk-Coverage curves.

Untrustworthy and Ambiguous quadrants. The Ambiguous quadrant has higher-quality uncertainty sets, while the Untrustworthy quadrant has lower coverage, as we showed that atypical examples have poor uncertainty sets. We give a simple classification setup for *Selective Conformal Prediction*:

$$O(X, Y, \hat{\mathbb{P}}) = \begin{cases} Y \in \mathcal{C}(X) & \text{\#Use uncertainty set} \\ Y \notin \mathcal{C}(X) & \text{\#Abstain} \end{cases} \tag{4}$$

If we can reliably predict $O(X, Y)$, then we can pick the action based on this predicted outcome. To do so, we can train a simple classifier $\hat{O}(X, \hat{\mathbb{P}})$ to predict this outcome. To demonstrate how atypicality provides useful information over using only confidence, we train two such classifiers: 1-Using only the confidence: $\hat{O}(X, \hat{\mathbb{P}}) = g(\hat{\mathbb{P}}(X))$; and 2-Using atypicality and the confidence: $\hat{O}(X, \hat{\mathbb{P}}) = g(\hat{\mathbb{P}}(X), a(X))$. If the latter performs better than the former, we understand that using atypicality can improve the decision-making process. The exact form of g can be flexible. In the experiments we use XGBoost to make the decisions; we refer to Appendix J for details.

To evaluate our approach, we employ the Risk-Coverage curves used in the Selective Classification literature (Geifman & El-Yaniv, 2017). [3] For a fixed coverage rate, the goal is to achieve lower risk by knowing when to abstain. Here, we define the notion of risk to be the fraction of samples where the uncertainty set does not contain the true label, namely $\mathbf{1}[Y \notin \mathcal{C}(X)]$, and the abstention is needed.

Results: In Figure 10, we present the Risk-Coverage curves. Using atypicality in the decision function provides strictly better Risk-Coverage curves, providing an easy way to improve the decision-making process. Namely, we can better prioritize when it is better to reject the sample, and when the uncertainty set is more likely to contain the true label.

6 Conclusion

In conclusion, our research demonstrates that atypicality offers a valuable framework to better understand model reliability and generate actionable insights. The atypicality framework can help us better predict uncertainty, better recalibrate models, and make better decisions. We propose that pretrained models should be released not only with confidence but also with an atypicality estimator. While there are many other relevant notions in the literature, our main goal is to show that atypicality can provide a unifying perspective to discuss uncertainty, understand individual data points, and improve fairness. While we analyzed only classification problems, our analyses can be naturally extended to regression or generation settings. Furthermore, we would like to extend the theoretical analysis to more general settings, as our empirical results demonstrate that the observed phenomena hold more broadly. In light of our findings, we further encourage the community to mitigate the adverse effects that could be caused by using models on the atypical data.

[3] Please note that in this context, the term Coverage means the fraction of samples where we did not abstain.

REFERENCES

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Varun Babbar, Umang Bhatt, and Adrian Weller. On the utility of prediction sets in human-ai teams. *arXiv preprint arXiv:2205.01411*, 2022.

Yu Bai, Song Mei, Huan Wang, and Caiming Xiong. Don't just blame over-parametrization for over-confidence: Theoretical analysis of calibration in binary classification. In *International Conference on Machine Learning*, pp. 566–576. PMLR, 2021.

Noam Barda, Gal Yona, Guy N Rothblum, Philip Greenland, Morton Leibowitz, Ran Balicer, Eitan Bachmat, and Noa Dagan. Addressing bias in prediction models by improving subpopulation calibration. *Journal of the American Medical Informatics Association*, 28(3):549–558, 2021.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Nips tutorial*, 1: 2, 2017.

Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Practical adversarial multivalid conformal prediction. *arXiv preprint arXiv:2206.01067*, 2022.

Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 401–413, 2021.

Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on NLP for ConvAI - ACL 2020*, mar 2020. Data available at https://github.com/PolyAI-LDN/task-specific-datasets.

Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *Advances in neural information processing systems*, 31, 2018.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Michelle Chua, Doyun Kim, Jongmun Choi, Nahyoung G Lee, Vikram Deshpande, Joseph Schwab, Michael H Lev, Ramon G Gonzalez, Michael S Gee, and Synho Do. Tackling prediction uncertainty in machine learning for healthcare. *Nature Biomedical Engineering*, pp. 1–8, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Adam Fisch, Tommi Jaakkola, and Regina Barzilay. Calibrated selective classification. *arXiv preprint arXiv:2208.12084*, 2022.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

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

Hila Gonen, Srini Iyer, Terra Blevins, Noah A Smith, and Luke Zettlemoyer. Demystifying prompts in language models via perplexity estimation. *arXiv preprint arXiv:2212.04037*, 2022.

Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*, 2020.

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

Guy Hacohen, Avihu Dekel, and Daphna Weinshall. Active learning on a budget: Opposite strategies suit high and low budgets. *arXiv preprint arXiv:2202.02794*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pp. 1939–1948. PMLR, 2018.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ArXiv*, abs/1610.02136, 2016.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021.

Tom Joy, Francesco Pinto, Ser-Nam Lim, Philip HS Torr, and Puneet K Dokania. Sample-dependent adaptive temperature scaling for improved calibration. *arXiv preprint arXiv:2207.06211*, 2022.

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

Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32, 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2021.

Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in Neural Information Processing Systems*, 33:7498–7512, 2020.

Lydia T Liu, Max Simchowitz, and Moritz Hardt. The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*, pp. 4051–4060. PMLR, 2019a.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019b.

Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 12008–12016, 2022.

Pratyush Maini, Saurabh Garg, Zachary C Lipton, and J Zico Kolter. Characterizing datapoints via second-split forgetting. *arXiv preprint arXiv:2210.15031*, 2022.

Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1485–1488, 2010.

Carolyn B Mervis and John R Pani. Acquisition of basic object categories. *Cognitive Psychology*, 12 (4):496–522, 1980.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty: A simple baseline. *arXiv e-prints*, pp. arXiv–2102, 2021.

Gregory Murphy. *The big book of concepts*. MIT press, 2004.

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.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Janis Postels, Hermann Blum, Cesar Cadena, Roland Siegwart, Luc Van Gool, and Federico Tombari. Quantifying aleatoric and epistemic uncertainty using density estimation in latent space. *arXiv preprint arXiv:2012.03082*, 2020.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pp. 63–71. Springer, 2004.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

Lance J . Rips. *Similarity, typicality, and categorization*, pp. 21–59. Cambridge University Press, 1989. doi: 10.1017/CBO9780511529863.004.

Lance J Rips, Edward J Shoben, and Edward E Smith. Semantic distance and the verification of semantic relations. *Journal of verbal learning and verbal behavior*, 12(1):1–20, 1973.

Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel J Candès. With malice towards none: Assessing uncertainty via equalized coverage. *arXiv preprint arXiv:1908.05428*, 2019.

Eleanor Rosch and Carolyn B Mervis. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4):573–605, 1975.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

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.

Nabeel Seedat, Jonathan Crabbé, Ioana Bica, and Mihaela van der Schaar. Data-iq: Characterizing subgroups with heterogeneous outcomes in tabular data. *arXiv preprint arXiv:2210.13043*, 2022.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

MTCAJ Thomas and A Thomas Joy. *Elements of information theory*. Wiley-Interscience, 2006.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131, 1974.

Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4):297–323, 1992.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. Openml: networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013. doi: 10.1145/2641190.2641198.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2018.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2020.

A Code

You can find the code to generate the analyses in this paper at the anonymized repository `https://anonymous.4open.science/status/beyond-confidence-iclr-trustworthyml-02EF`.

B Atypicality is Needed to Understand Uncertainty

High-confidence and representative: A sample and prediction are considered reliable if the model makes a high-confidence prediction that is well-calibrated. For instance, the first image on the top right quadrant of Figure 1 is highly representative of the golden retriever class, and the model makes a high-confidence prediction for this sample.

High-confidence yet far from the support: Even though the model has high confidence in a prediction, this confidence may not be reliable. In case the sample does not have support in the training distribution yet the model has a confident prediction (extrapolation), the prediction may not be reliable. For instance, the second image in the top left quadrant of Figure 1 is a toy hog, and the model may not have seen similar objects during training.

Low confidence and far from distribution: A model can have low confidence in far-from-the-distribution samples. For example, the third image in Figure 1 bottom left quadrant is an origami, which can be out-of-distribution. We expect low-quality predictions in such cases.

Low confidence due to ambiguity: A low confidence prediction may be the result of ambiguity in a sample. For instance, the second image in the bottom left quadrant of Figure 1 is both a hog and can be a comic book. This is an inherent ambiguity, which is tied to the concept of aleatoric uncertainty (Der Kiureghian & Ditlevsen, 2009).

C Related Works

C.1 Relation to earlier works

The notion of atypicality has been shared in earlier works. Our distinct goal is to show that the atypicality perspective is necessary and complementary to understanding and improving various aspects of model uncertainty. We propose that if all models are equipped with simple-to-use atypicality estimators, we can improve the reliability of the predictions. Methodologically, Mukhoti et al. (2021) take a very similar approach to quantify density to distinguish epistemic and aleatoric uncertainty. However, the source of uncertainty for an atypical example could be a combination of both epistemic and aleatoric uncertainty, as noted by Mukhoti et al. (2021), and achieving this decomposition is not our primary goal. An image of an unusual-colored dog would be atypical with high epistemic uncertainty, yet a corrupted or highly noisy sample would be atypical with high aleatoric uncertainty. Liu et al. (2020) propose a related notion of 'distance awareness'. The underlying principles are highly relevant, however, the findings and the methodologies proposed by both works are different. The notion of 'out-of-distribution' (Hendrycks & Gimpel, 2016) is closely related to atypicality, however, we do not aim to make the binary decision between 'in-distribution' and 'out-of-distribution'. For instance, we emphasize that samples could be in-distribution and atypical, e.g. rare subgroups (Sagawa et al., 2019), and our goal is to perform reliably in the entire spectrum.

D Additional Related Work

Uncertainty and Atypicality: Closest in spirit and methodology to our work is Mukhoti et al. (2021); Postels et al. (2020) where they use density estimation in the feature space to quantify and disentangle the epistemic uncertainty and aleatoric uncertainty. Using this decomposition, they show significant improvements in active learning and OOD detection. Similarly, Seedat et al. (2022) use epistemic-aleatoric decomposition to characterize tabular data points. Lee et al. (2018) use Mahalanobis distance to detect OOD examples, utilizing class-conditional gaussian likelihoods. In concurrent work, Gonen et al. (2022) reports that perplexity, a measure of atypicality, is correlated with the performance of language models in zero-shot classification. Hacohen et al. (2022) discuss active learning, showing that model performance depends on atypicality in different regimes of active learning. Liu et al. (2020) propose the relevant notion of distance-awareness in uncertainty estimation, showing that accounting for distance leads to better uncertainty quantification. Namely, they propose architecture and training modifications to improve uncertainty quantification whereas here, we analyze the uncertainty of existing models with respect to our framework, and propose very simple, post-hoc approaches to mitigate the found issues. Liu et al. (2019a) relate the excess risk of a binary classifier to calibration, showing that the calibration gap for discrete groups is bounded by the excess risk for the group. Here, we extend these results and provide the characterization with respect to atypicality. Multicalibration (Hébert-Johnson et al., 2018; Jung et al., 2021) is a relevant notion where the calibration criterion holds for a collection of discrete groups. The notion of 'out-of-distribution' (Hendrycks & Gimpel, 2016) is highly relevant to atypicality, yet our goal is not to make the binary distinction between 'in-distribution' and 'out-of-distribution'. We aim to perform reliably in the entire spectrum from atypical (e.g., rare groups) to typical (e.g., ambiguous samples).

Related findings include poor calibration under distribution shifts (Ovadia et al., 2019), uncertainty in Gaussian Processes (Rasmussen, 2004), forgetting time for rare examples (Maini et al., 2022), the poor performance of groups with lower sample sizes (Chen et al., 2018), using energy-based models improving calibration (Grathwohl et al., 2020) showing the relationship between some notion of atypicality versus uncertainty-related quantities. Our new findings include showing that predictions for atypical samples are more miscalibrated and overconfident, and atypicality awareness improves recalibration. *More broadly, while there are many other relevant notions in the literature, our distinct goal is to show that post-hoc atypicality estimation is a simple yet useful framework to understand and improve uncertainty quantification, and complements existing methods. Thus, we take the position that released models should have atypicality estimators.*

Recalibration: There is a rich literature on recalibration methods: Temperature scaling (Guo et al., 2017), Platt Scaling (Platt et al., 1999), conformal calibration (Shafer & Vovk, 2008; Angelopoulos et al., 2020) among many. Lu et al. (2022); Romano et al. (2019); Barda et al. (2021); Bastani et al. (2022) make a relevant observation in classification and regression respectively, showing that the coverage of conformal prediction is not equal across all groups. They propose group conformal calibration, which requires group labels whereas our proposal is unsupervised and does not depend on any attribute information. Concurrent work (Joy et al., 2022) explores sample-dependent TS, where they modify the training pipeline and train a separate network. However, our parameterization of temperature is based only on atypicality, and our approach is post-hoc and cheap.

Uncertainty and Decision-Making: Uncertainty of a model has been used in various ways. Selective classification (Geifman & El-Yaniv, 2017; Fisch et al., 2022) proposes abstaining as an option in downstream decisions. Conformal prediction (Shafer & Vovk, 2008; Angelopoulos et al., 2020) proposes to provide uncertainty sets instead of a single class prediction. However, it is still an open research direction to assess the utility of these approaches. Bhatt et al. (2021) gives a broad overview of the topic. Babbar et al. (2022) studies the utility of prediction sets, showing that when used carefully they can improve user trust and Human-AI collaboration performance. Overall, our approach demonstrates that post-hoc atypicality estimators help unify these settings and achieve improvements.

E Atypicality Estimation

Quantifying atypicality requires access to the class-conditional / marginal distributions. However, in practice, we do not have access to these, and hence, we need to compute the estimates. This estimation can be challenging if the dimensionality is large, or the data is unstructured as in language or vision modalities, requiring assumptions about the distributions. Prior work such as Mukhoti et al. (2021); Lee et al. (2018) have shown that Gaussian Mixture Models in the embedding space of neural networks can be used to model these distributions. In our experiments we use class-conditional Gaussian distributions with shared covariance matrices, i.e. $\hat{\mathbb{P}}(X = x|Y = c) \sim N(\hat{\mu}_c, \hat{\Sigma})$, to estimate atypicality. We perform these in the penultimate layer of the neural networks used to make predictions. The parameters are estimated using maximum-likelihood estimation with samples from the training dataset. We explore other methods to compute atypicality, such as k-Nearest Neighbors distance as an atypicality metric. We give implementation details of these methods in Appendix E.1 and report results with different atypicality metrics. Further, we refer to the probability of the top-class, i.e. $\max_y \hat{\mathbb{P}}(Y = y|X = x)$ as the *Confidence*.

E.1 Density Estimation

To estimate atypicality, we use two ways to estimate the likelihood of a point under the training distribution.

Fitting individual Gaussians to Class Conditionals Here, we follow a similar approach to Mukhoti et al. (2021). Namely, we model the clas-conditionals with a gaussian, where the covariance matrix is tied across classes:

$$\hat{\mathbb{P}}(X|Y = y) \sim N(X; \mu_y, \Sigma) \tag{5}$$

We fit the parameters μ_y and Σ with maximum likelihood estimation. The reason to tie the covariance matrix is due to the number of samples required to fit the density. Namely, for a d-dimensional

problem, the total number of parameters to fit individual matrices become $O(yd^2)$, which results in low-quality estimates. Then, the atypicality becomes

$$a(X) = 1 - \max_{y \in \mathcal{Y}} \hat{\mathbb{P}}(X|Y = y) \tag{6}$$

E.2 Computing distance with k-Nearest Neighbors

k-Nearest Neighbors: Similarly, we can use the nearest neighbor distance. Concretely, we use the nearest neighbor distance, $a(x) = d_{\min}(x, \mathcal{D}_{\text{train}}) = \min_{x' \in \mathcal{D}_{\text{train}}} |x' - x|$, as the atypicality metric. Alternatively, we can use different notions such as the average of k-nearest neighbors, or the distance to the kth neighbor. Below, we report the results by using the average distance to 10-nearest neighbors.

E.3 Fitting the estimators

For all of the atypicality estimators, we fit the estimators using samples from the training sets and make inference for the calibration and test sets. For instance, we use the training split of ImageNet to fit the corresponding density estimator and compute the atypicality for the samples from the validation/test split of ImageNet. All of our results using atypicality are reported for the test splits of the below datasets.

E.4 Atypicality and Confidence

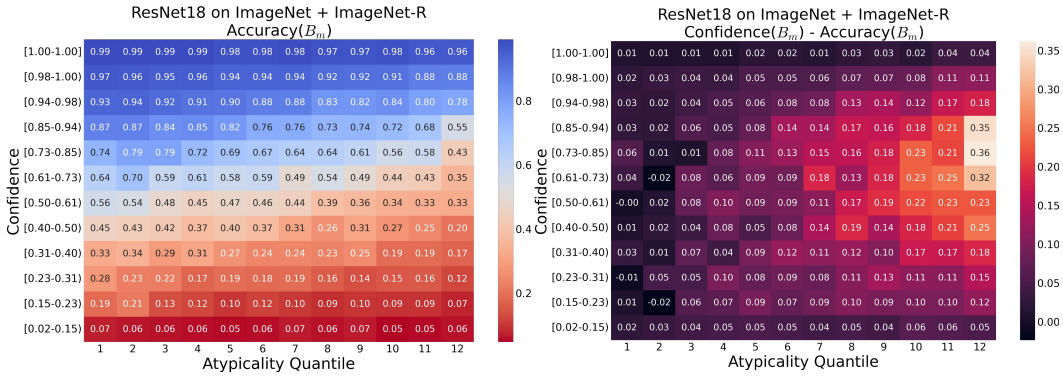

Figure 6: **Atypicality and Confidence.** Here, x-axis reflects the typicality quantile, and y-axis indicates the confidence. The coloring for the figure on the left indicates the accuracy within a bin, and the figure on the right has the difference betwee the confidence and the accuracy within a bin. We observe that even within the same confidence range, atypical examples tend to be more miscalibrated compared to typical examples.

Are atypicality and confidence equally informative? Beyond the data perspective given in Figure 1, here we provide quantitative results to demonstrate the difference. In Figure 6, we provide a grid plot where the x-axis indicates the typicality quantile of a point, and the y-axis indicates the confidence of a point. The coloring on the left indicates the accuracy within a bin split according to accuracy, and the right has the difference between average confidence and the accuracy. Observe that for a certain confidence interval, larger values of typicality mean better quality probabilistic estimates, and larger atypicality means more miscalibration.

F Experimental Details

Throughout this study, we focus on classification problems across a range of datasets and architectures. We utilize publicly available datasets and models, which are detailed in Appendix F. Specifically, we evaluate image classifiers ResNet (He et al., 2016) and ViT (Dosovitskiy et al., 2020), text

classifiers RoBERTa (Liu et al., 2019b) and DistilBERT (Sanh et al., 2019), and XGBoost (Chen & Guestrin, 2016) for tabular classification. We employ CIFAR10, CIFAR100 (Krizhevsky, 2009), and ImageNet (Deng et al., 2009) for object recognition, MNLI (Williams et al., 2018) for natural language inference, Banking77 (Casanueva et al., 2020) for banking intent classification, and the Electricity dataset (Vanschoren et al., 2013), which is a tabular classification dataset for predicting electricity price shifts.

F.1 DATASETS

Below is a full list of datasets:

1. ImageNet (Deng et al., 2009) from Torchvision (Marcel & Rodriguez, 2010) is an object recognition dataset with 1000 classes.
2. CIFAR10, CIFAR100 (Krizhevsky, 2009) from Torchvision (Marcel & Rodriguez, 2010) are object recognition datasets with 10/100 classes.
3. MNLI (Williams et al., 2018) from Huggingface Datasets (Lhoest et al., 2021) is a natural language inference dataset with 3 classes, indicating entailment, neutral, and contradiction outcomes.
4. Banking77 (Casanueva et al., 2020) from Huggingface Datasets (Lhoest et al., 2021) is a banking intent classification dataset with 77 classes.
5. Electricity from OpenML (Vanschoren et al., 2013) obtained through Scikit-Learn (Pedregosa et al., 2011)

For the electricity dataset, we use the average distance to 10-nearest-neighbors as the atypicality metric. All of our experiments were run on a single Nvidia GeForce RTX 2080Ti GPU.

F.2 MODELS

Most of the models are public models, e.g., obtained from the Transformers Library (Wolf et al., 2020) or Torchvision (Marcel & Rodriguez, 2010). Below we give the full model details and how one can access them:

1. **ViT**(`HuggingFace Ahmed9275/Vit-Cifar100`), pretrained on Imagenet-21k (Ridnik et al., 2021) then finetuned on CIFAR100. One can use the id given here on HuggingFace to download the model.
2. **RoBERTa**(`HuggingFace roberta-large-mnli`) trained on the MNLI dataset. One can use the id given here on HuggingFace to download the model.
3. **ResNet18** from (`Torchvision` (Marcel & Rodriguez, 2010)) trained on ImageNet.
4. **ResNet20** trained on CIFAR10, which can be downloaded from (`PyTorchCV`[4])
5. **DistilBERT** trained on Banking77(HuggingFace `optimum/distilbert-base-uncased-finetuned-bank`
which can be downloaded from HuggingFace
6. **XGBoost**: We train our own XGBoost model on the tabular dataset where we use the XGBoost library, `https://github.com/dmlc/xgboost`.

For all BERT (Devlin et al., 2019) style models we use the `[CLS]` token embeddings in the final layer, and for all vision models, we use the penultimate layer embeddings to fit the density estimators and perform the analyses.

G CALIBRATION

We run all our experiments with 10 different random seeds, where the seeds are $\{0, 1, 2, \ldots, 9\}$. Randomness is over fitting the atypicality estimators, and calibration-test splits (we use the same splits with the recalibration experiments for the sake of consistency).

[4]https://github.com/osmr/imgclsmob

G.1 Expected Calibration Error

To compute ECE, we generate $\mathbb{B} = \{B_1, B_2, ..., B_M\}$, M equally-spaced bins where samples are sorted and grouped according to their confidence, to compute

$$\text{ECE}[\hat{\mathbb{P}}] = \sum_{m=1}^{M} \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)| \tag{7}$$

where $\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i=1}^{|B_m|} \mathbf{1}[\hat{y}_i = y_i]$ is the accuracy for the bin m, and $\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i=1}^{|B_m|} \hat{\mathbb{P}}(Y = \hat{y}_i | X = x_i)$ gives the average confidence within the bin. $|B_m|$ is the size of the bin m, N is the total number of samples, and $\mathbf{1}[\cdot]$ is the indicator function.

Throughout our experiments, we let the number of bins $|\mathbb{B}| = 10$ by default.

G.2 Reliability Diagrams

Here, we present the reliability diagrams decomposed by atypicality for more fine-grained analysis. For instance, one can observe that for ImageNet and ResNet18, the direction of miscalibration is towards overconfidence. Particularly, predictions for more atypical points are more overconfident compared to the typical points, see Figure 7.

Furthermore, if we look at the miscalibration rates after recalibration, we observe that miscalibration rates get more flat with Atypicality-Aware recalibration, improving calibration across the board, see Figure 8.

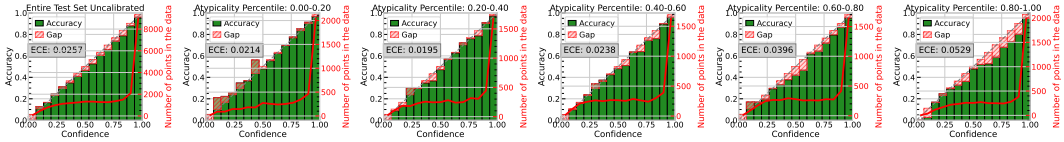

Figure 7: **Reliability diagrams for ResNet18 on ImageNet.** Here, we share the reliability diagrams for all recalibration methods and atypicality quantiles. We observe that a) the model is more overconfident for atypical points, b) Atypicality-Aware Recalibration improves ECE for all groups.

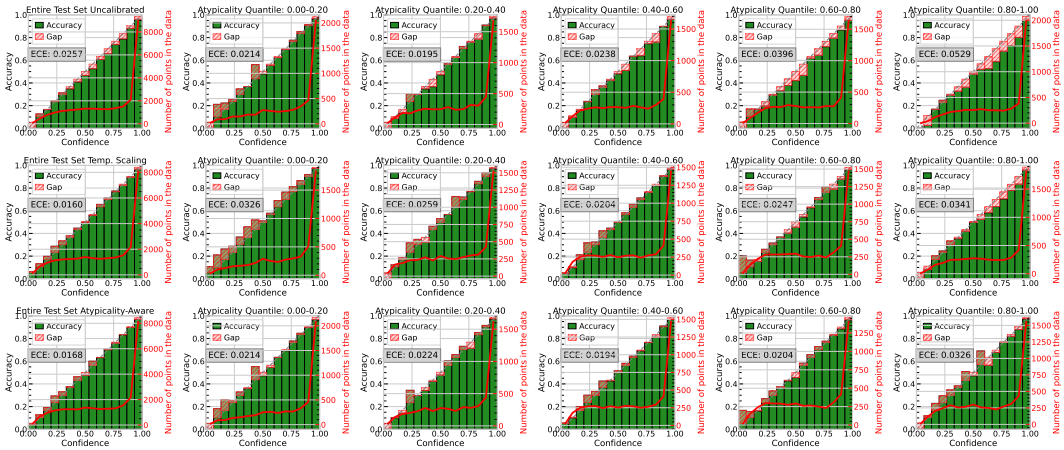

Figure 8: **Reliability diagrams for recalibration with ResNet18 on ImageNet.** Here, we share the reliability diagrams for all recalibration methods and atypicality quantiles. We observe that a) the model is more overconfident for atypical points, b) Atypicality-Aware Recalibration improves ECE for all groups.

G.3 ECE and Atypicality Results with Different Atypicality Metrics

We further experiment with different atypicality metrics, such as the average distance to the 10-nearest neighbors (Figure 9). We broadly observe that while there are slight differences in the quantitative results between different atypicality metrics, the qualitative phenomena remain intact.

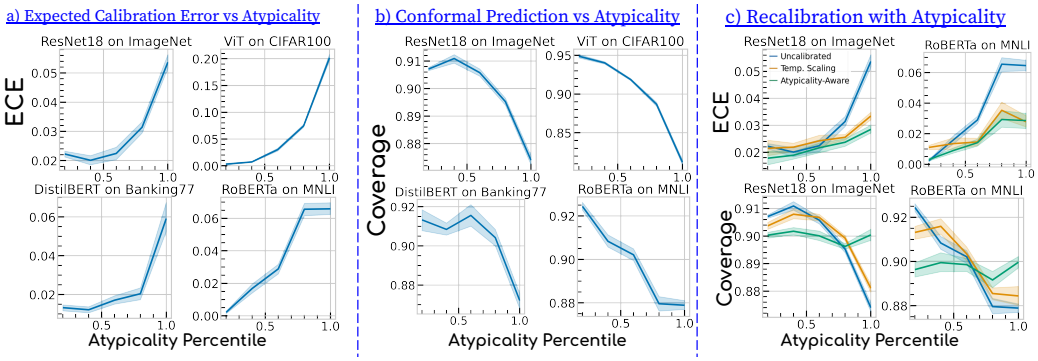

Figure 9: **Atypicality with 10-nearest neighbors and Uncertainty.** Here, we report the results of the same experiments as Figure 3 with the average of the distance to the 10-nearest neighbors as the atypicality metric. See the Tables 2 and 4 for the results in tabular format.

H Recalibration

Through all our recalibration results, we first split the test set into two equally sized calibration and test splits. Then, we fit the recalibration method using the calibration split and compute the performance on the test split. We run all our experiments with 10 different random seeds.

H.1 Temperature Scaling

To perform temperature scaling (Guo et al., 2017), we use the calibration set to fit the temperature parameter. To perform the optimization, we use the LBFGS (Liu & Nocedal, 1989) algorithm from PyTorch with strong Wolfe line search, following Guo et al. (2017). Namely, we optimize the parameter τ with

$$\hat{\mathbb{P}}_{\text{TS}}(X) = \text{Softmax}(f(\boldsymbol{X})/\tau) \tag{8}$$

and then use it during inference to rescale the logits produced by f.

H.2 Conformal Prediction

We follow the presentation in Angelopoulos & Bates (2021); Angelopoulos et al. (2020). Let $\pi(X)$ be the permutation of $\mathcal{Y} = \{1, \ldots, C\}$ that sorts $\hat{\mathbb{P}}(Y = c|X)$, i.e. the predicted probabilities for each class c. We define a score function

$$s(x, y) = \sum_{j=1}^{c} \hat{\mathbb{P}}(Y = j|X), \text{where } y = \pi_c. \tag{9}$$

This means greedily including classes until the set contains the true label, and using the cumulative sum of the probabilities as the score function. We compute all of the scores for the calibration set, $S_{\text{calib}} = \{s(x_1, y_1), ..., s(x_N, y_N)\}$, we the $\frac{\lceil (N+1)(1-\alpha) \rceil}{N}$th quantile of the scores, $\hat{q}$. Then, the uncertainty set is defined as

$$\mathcal{C}(x) = \{y : s(x, y) \leq \hat{q}\} \tag{10}$$

We can further add randomization to the procedure where we have the uncertainty set function to be $\mathcal{C}(x, u) : \mathcal{X} \times [0, 1]$ for randomization purposes to satisfy exact coverage. We refer to Vovk et al. (2005); Angelopoulos et al. (2020); Angelopoulos & Bates (2021) for a more thorough presentation.

H.3 ATYPICALITY-AWARE RECALIBRATION

To perform the Atypicality-Aware Recalibration, we parameterize the temperature with

$$\tau(X) = c_2 a(X)^2 + c_1 a(X) + c_0 \tag{11}$$

which leads to

$$\hat{\mathbb{P}}_{\text{ATS}}(X) = \text{Softmax}(f(\boldsymbol{X})/\tau(X))$$

Similar to vanilla temperature scaling, we use LBFGS with strong wolfe search to optimize the three parameters, with the same splits as temperature scaling.

I TABLES FOR RESULTS

Here, we present the table version of the results in Figure 3. Tables 1 and 2 contain the ECE analysis, and Tables 3 and 2 contain the coverage analyses for the two atypicality metrics.

Table 1: Analyzing Expected Calibration Error versus Atypicality. Here, the atypicality metric is through fitting Gaussians as class-conditionals. All experiments were run with 10 random seeds and means along with standard errors are reported.

Atypicality ($-\max_c \hat{\mathbb{P}}(X\|Y=c)$) **vs Expected Calibration Error**			
Atypicality Quantile	Atypicality-Aware Recalibration	Temperature Scaling	Before Recalibration
	ResNet18 on ImageNet		
0.2	0.019 ± 0.001	0.033 ± 0.001	0.019 ± 0.001
0.4	0.017 ± 0.001	0.021 ± 0.001	0.019 ± 0.001
0.6	0.022 ± 0.001	0.022 ± 0.001	0.025 ± 0.002
0.8	0.023 ± 0.001	0.026 ± 0.001	0.043 ± 0.002
1.0	0.023 ± 0.001	0.034 ± 0.001	0.054 ± 0.001
	ResNet20 on CIFAR10		
0.2	0.001 ± 0.000	0.004 ± 0.000	0.002 ± 0.000
0.4	0.005 ± 0.001	0.004 ± 0.001	0.004 ± 0.000
0.6	0.011 ± 0.001	0.010 ± 0.002	0.012 ± 0.001
0.8	0.014 ± 0.001	0.016 ± 0.001	0.019 ± 0.002
1.0	0.020 ± 0.002	0.022 ± 0.002	0.026 ± 0.002
	ViT on CIFAR100		
0.2	0.004 ± 0.001	0.009 ± 0.001	0.004 ± 0.001
0.4	0.004 ± 0.001	0.009 ± 0.001	0.011 ± 0.001
0.6	0.011 ± 0.001	0.010 ± 0.001	0.026 ± 0.001
0.8	0.042 ± 0.002	0.045 ± 0.002	0.079 ± 0.002
1.0	0.085 ± 0.003	0.100 ± 0.003	0.196 ± 0.002
	RoBERTa on MNLI		
0.2	0.006 ± 0.001	0.013 ± 0.001	0.005 ± 0.001
0.4	0.008 ± 0.001	0.013 ± 0.001	0.016 ± 0.001
0.6	0.018 ± 0.001	0.015 ± 0.001	0.034 ± 0.002
0.8	0.029 ± 0.002	0.031 ± 0.002	0.061 ± 0.002
1.0	0.022 ± 0.003	0.028 ± 0.002	0.065 ± 0.002
	DistilBERT on Banking77		
0.2	0.004 ± 0.001	0.003 ± 0.001	0.007 ± 0.001
0.4	0.005 ± 0.001	0.009 ± 0.001	0.016 ± 0.001
0.6	0.011 ± 0.002	0.015 ± 0.001	0.023 ± 0.001
0.8	0.040 ± 0.004	0.036 ± 0.003	0.026 ± 0.002
1.0	0.059 ± 0.004	0.068 ± 0.005	0.049 ± 0.004

Table 2: Analyzing Expected Calibration Error versus Atypicality. Here, the atypicality metric is the average distance to 10-nearest neighbors. All experiments were run with 10 random seeds and means along with standard errors are reported.

Atypicality (average distance to 10-nearest neighbors) vs Expected Calibration Error			
Atypicality Quantile	Atypicality-Aware Recalibration	Temperature Scaling	Before Recalibration
	ResNet18 on ImageNet		
0.2	0.018 ± 0.001	0.021 ± 0.002	0.022 ± 0.001
0.4	0.019 ± 0.001	0.022 ± 0.001	0.020 ± 0.001
0.6	0.022 ± 0.001	0.024 ± 0.001	0.022 ± 0.001
0.8	0.024 ± 0.001	0.026 ± 0.001	0.031 ± 0.001
1.0	0.029 ± 0.001	0.033 ± 0.001	0.054 ± 0.001
	ResNet20 on CIFAR10		
0.2	0.001 ± 0.000	0.006 ± 0.000	0.004 ± 0.000
0.4	0.001 ± 0.000	0.006 ± 0.000	0.004 ± 0.000
0.6	0.007 ± 0.001	0.005 ± 0.001	0.006 ± 0.001
0.8	0.016 ± 0.002	0.019 ± 0.002	0.023 ± 0.002
1.0	0.018 ± 0.002	0.031 ± 0.002	0.035 ± 0.002
	ViT on CIFAR100		
0.2	0.002 ± 0.000	0.009 ± 0.001	0.003 ± 0.001
0.4	0.003 ± 0.001	0.012 ± 0.001	0.007 ± 0.001
0.6	0.014 ± 0.002	0.011 ± 0.001	0.030 ± 0.001
0.8	0.033 ± 0.002	0.042 ± 0.002	0.074 ± 0.001
1.0	0.080 ± 0.004	0.103 ± 0.003	0.202 ± 0.003
	RoBERTa on MNLI		
0.2	0.003 ± 0.000	0.011 ± 0.001	0.002 ± 0.000
0.4	0.009 ± 0.001	0.013 ± 0.002	0.016 ± 0.001
0.6	0.014 ± 0.001	0.015 ± 0.001	0.029 ± 0.001
0.8	0.029 ± 0.003	0.035 ± 0.003	0.066 ± 0.002
1.0	0.029 ± 0.003	0.028 ± 0.002	0.065 ± 0.002
	DistilBERT on Banking77		
0.2	0.002 ± 0.001	0.007 ± 0.001	0.013 ± 0.001
0.4	0.003 ± 0.001	0.004 ± 0.001	0.012 ± 0.001
0.6	0.003 ± 0.001	0.007 ± 0.001	0.017 ± 0.001
0.8	0.022 ± 0.003	0.020 ± 0.002	0.020 ± 0.002
1.0	0.065 ± 0.006	0.084 ± 0.005	0.059 ± 0.004

J DECISION-MAKING

For the decision-making setup, we train 2 sets of classifiers:

1. Using only the confidence: $\hat{O}(X, \hat{\mathbb{P}}) = g(\hat{\mathbb{P}}(X))$
2. Using atypicality and the confidence: $\hat{O}(X, \hat{\mathbb{P}}) = g(\hat{\mathbb{P}}(X), a(X))$

We XGBoost to predict the outcome. We run hyperparameter search with 5-fold cross-validation using the calibration set, over the following parameters: `min_child_weight` $\in \{1, 5, 10\}$, `gamma` $\in \{0.5, 1, 1.5, 2, 5\}$, `subsample` $\in \{0.6, 0.8, 1.0\}$, `colsample_bytree` $\in \{0.6, 0.8, 1.0\}$, `max_depth` $\in \{2, 3, 4, 5\}$, `n_estimators` $\in \{10, 20, 40\}$. We refer the reader to the XGBoost documentation[5] page for the detailed semantics of each hyperparameter.

[5]https://xgboost.readthedocs.io/en/stable/

Table 3: Analyzing Coverage versus Atypicality. Here, the atypicality metric is through fitting Gaussians as class-conditionals. All experiments were run with 10 random seeds and means along with standard errors are reported.

Atypicality ($-\max_c \hat{\mathbb{P}}(X\|Y=c)$) **vs Coverage**			
Atypicality Quantile	Atypicality-Aware Recalibration	Temperature Scaling	Before Recalibration
	ResNet18 on ImageNet		
0.2	0.901 ± 0.001	0.919 ± 0.001	0.920 ± 0.001
0.4	0.902 ± 0.002	0.913 ± 0.001	0.913 ± 0.002
0.6	0.899 ± 0.001	0.903 ± 0.001	0.901 ± 0.001
0.8	0.895 ± 0.002	0.891 ± 0.001	0.886 ± 0.001
1.0	0.902 ± 0.002	0.873 ± 0.001	0.873 ± 0.002
	ResNet20 on CIFAR10		
0.2	0.901 ± 0.003	0.904 ± 0.004	0.910 ± 0.004
0.4	0.897 ± 0.003	0.903 ± 0.004	0.906 ± 0.003
0.6	0.899 ± 0.004	0.898 ± 0.003	0.902 ± 0.003
0.8	0.901 ± 0.003	0.894 ± 0.004	0.898 ± 0.005
1.0	0.900 ± 0.003	0.893 ± 0.004	0.893 ± 0.002
	ViT on CIFAR100		
0.2	0.902 ± 0.004	0.913 ± 0.002	0.947 ± 0.002
0.4	0.902 ± 0.003	0.910 ± 0.004	0.940 ± 0.001
0.6	0.890 ± 0.004	0.905 ± 0.003	0.923 ± 0.003
0.8	0.878 ± 0.002	0.877 ± 0.002	0.884 ± 0.003
1.0	0.909 ± 0.002	0.886 ± 0.003	0.812 ± 0.003
	RoBERTa on MNLI		
0.2	0.900 ± 0.004	0.918 ± 0.003	0.919 ± 0.003
0.4	0.900 ± 0.004	0.912 ± 0.003	0.913 ± 0.003
0.6	0.892 ± 0.003	0.895 ± 0.005	0.897 ± 0.003
0.8	0.894 ± 0.003	0.889 ± 0.003	0.882 ± 0.004
1.0	0.901 ± 0.002	0.890 ± 0.004	0.882 ± 0.002
	DistilBERT on Banking77		
0.2	0.904 ± 0.008	0.906 ± 0.005	0.911 ± 0.005
0.4	0.912 ± 0.005	0.913 ± 0.006	0.913 ± 0.005
0.6	0.910 ± 0.004	0.918 ± 0.004	0.914 ± 0.004
0.8	0.885 ± 0.006	0.886 ± 0.006	0.890 ± 0.003
1.0	0.895 ± 0.005	0.872 ± 0.005	0.886 ± 0.004

K THEORY

K.1 SETTING

Data Generative Model: We consider the well-specified logistic model for binary classification with Gaussian data, where $Y \in \{-1, 1\}$ and the probability of $Y = 1$ given X is defined by the sigmoid function:

$$\mathbb{P}(Y = 1 \mid X) = \sigma(\langle \beta^*, X \rangle), \quad X \sim N(0, I_d).$$

Where I_d denotes the d-dimensional identity matrix, β^* is the ground truth coefficient vector, $\sigma(x) = 1/(1 + e^{-x})$, and we have $i.i.d.$ observations $\{(x_i, y_i)\}_{i=1}^n$ sampled from the above distribution.

The estimator: We focus on studying logistic regression, which produces a solution that minimizes the following:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} [\log(1 + \exp(\beta^\top x_i)) - y_i \cdot \beta^\top x_i].$$

Table 4: Analyzing Coverage versus Atypicality. Here, the atypicality metric is the average distance to the 10 nearest neighbors. All experiments were run with 10 random seeds and means along with standard errors are reported.

Atypicality (average distance to 10-nearest neighbors) vs Coverage			
Atypicality Quantile	Atypicality-Aware Recalibration	Temperature Scaling	Before Recalibration
	ResNet18 on ImageNet		
0.2	0.900 ± 0.001	0.904 ± 0.001	0.907 ± 0.001
0.4	0.902 ± 0.001	0.908 ± 0.002	0.911 ± 0.002
0.6	0.900 ± 0.002	0.907 ± 0.001	0.906 ± 0.001
0.8	0.896 ± 0.001	0.899 ± 0.001	0.895 ± 0.001
1.0	0.900 ± 0.002	0.881 ± 0.002	0.874 ± 0.002
	ResNet20 on CIFAR10		
0.2	0.905 ± 0.002	0.908 ± 0.002	0.915 ± 0.003
0.4	0.901 ± 0.003	0.905 ± 0.003	0.909 ± 0.004
0.6	0.896 ± 0.003	0.902 ± 0.004	0.903 ± 0.004
0.8	0.889 ± 0.004	0.885 ± 0.005	0.893 ± 0.003
1.0	0.906 ± 0.003	0.891 ± 0.002	0.888 ± 0.002
	ViT on CIFAR100		
0.2	0.896 ± 0.003	0.914 ± 0.004	0.949 ± 0.002
0.4	0.899 ± 0.003	0.911 ± 0.003	0.940 ± 0.002
0.6	0.887 ± 0.004	0.902 ± 0.004	0.919 ± 0.001
0.8	0.882 ± 0.003	0.880 ± 0.002	0.887 ± 0.003
1.0	0.916 ± 0.003	0.885 ± 0.003	0.812 ± 0.003
	RoBERTa on MNLI		
0.2	0.896 ± 0.003	0.913 ± 0.003	0.924 ± 0.002
0.4	0.899 ± 0.004	0.916 ± 0.004	0.908 ± 0.003
0.6	0.899 ± 0.003	0.903 ± 0.003	0.902 ± 0.003
0.8	0.892 ± 0.004	0.886 ± 0.003	0.880 ± 0.004
1.0	0.900 ± 0.003	0.884 ± 0.004	0.879 ± 0.002
	DistilBERT on Banking77		
0.2	0.908 ± 0.007	0.915 ± 0.005	0.913 ± 0.005
0.4	0.900 ± 0.004	0.909 ± 0.006	0.908 ± 0.003
0.6	0.913 ± 0.006	0.918 ± 0.003	0.916 ± 0.006
0.8	0.901 ± 0.004	0.897 ± 0.007	0.904 ± 0.004
1.0	0.886 ± 0.005	0.856 ± 0.006	0.872 ± 0.004

For $k \in \{-1, 1\}$, the confidence $\hat{\mathbb{P}}_k(x)$ is an estimator of $\hat{\mathbb{P}}(y = k|x)$, and it takes the form $\hat{\mathbb{P}}_k(x) = \frac{1}{e^{-k \cdot \hat{\beta}^\top x}+1}$.

Calibration: Here we consider the case where $\mathbb{P}_1(X) > 1/2$, as the case where $\mathbb{P}_1(X) \leq 1/2$ can be analyzed similarly by symmetry. For $u \in (1/2, 1)$, the signed calibration error at a confidence level u is

$$u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u).$$

We want to show that when X is atypical, i.e., when $a(X) := \exp(-\|X\|^2/2)$ is smaller[6], the accuracy $\mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)$ would be generally smaller than the confidence u (over-confidence).

[6]The definition of atypicality follows from the data model: density for the Gaussian with zero mean and identity covariance.

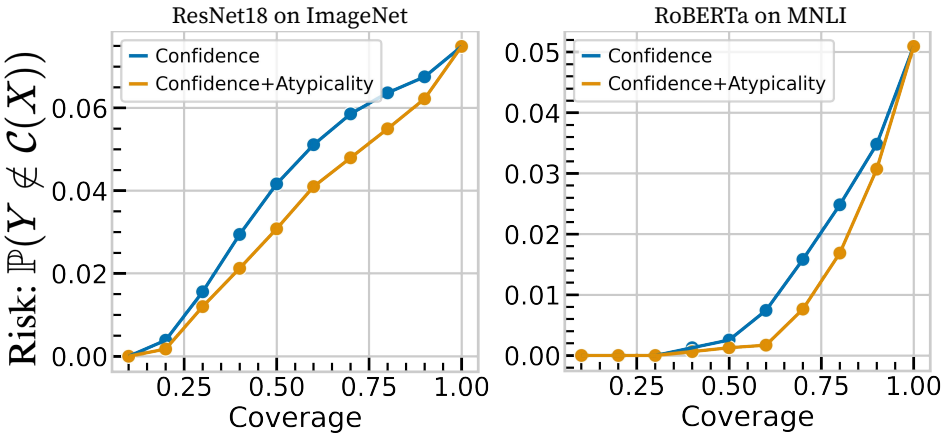

Figure 10: **Selective Conformal Prediction with Atypicality.** Here, coverage is the fraction of samples that were not rejected, and the risk is the average number of samples where the true label was not in the uncertainty set. We observe that using atypicality provides better Risk-Coverage curves.

K.2 Proof of Theorem 3.1

Theorem K.1 (Restatement of Theorem 3.1)**.** *Consider the data generative model with the algorithm described in Section 3.3. For any $K > 1$, suppose we consider the quantiles of $a(X)$, $a_1, a_2, ..., a_K, a_{K+1}$ such that $\mathbb{P}(a(X) \in (a_k, a_{k+1}]) = 1/K$ for $k \in [K]$. In addition, we assume $\|\beta^*\| \leq c_0$, and $d/n = \kappa$ for some sufficiently small $c_0, \kappa > 0$. Then for sufficiently large n, we have*

$$\mathbb{E}_u[u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u) \mid a(X) \in [a_{k-1}, a_k]] >$$
$$\mathbb{E}_u[u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u) \mid a(X) \in (a_k, a_{k+1}]],$$

for $k = 2, .., K$.

Proof. Following Bai et al. (2021), we have

$$u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u) = u - \mathbb{E}_Z[\sigma(\frac{\|\beta^*\|}{\|\hat{\beta}\|} \cos\hat{\theta} \cdot \sigma^{-1}(u)) + \sin\hat{\theta} \cdot \|\beta^*\| Z],$$

where $\cos\hat{\theta} = \frac{\hat{\beta}^\top \beta^*}{\|\hat{\beta}\| \cdot \|\beta^*\|}$ and $Z \sim N(0, 1)$.

According to Sur & Candès (2019), we have $\|\hat{\beta}\| \to R^* = R^*(\kappa, \beta^*)$ and $\cos\hat{\theta} \to c^* = c^*(\kappa, \beta^*)$, for two quantities R^* and c^* that depend on κ and β^*. We then have

$$u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u) = u - \mathbb{E}_Z[\sigma(\frac{\|\beta^*\|}{R^*} c^* \cdot \sigma^{-1}(u)) + \sqrt{1 - c^{*2}} \cdot \|\beta^*\| Z].$$

Using the proof of Theorem 3 in Bai et al. (2021), we have that

$$u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u) = C_\kappa(u) \cdot \kappa + o(\kappa),$$

where

$$C_\kappa(u) = c_1 \sigma'(\sigma^{-1}(u)) \cdot \sigma^{-1}(u) - c_2 \sigma''(\sigma^{-1}(u)),$$

for two positive constants c_1, c_2.

Since when $z \in [-1, 1]$, $z \cdot \sigma'(z)$ and $-\sigma''(z)$ are both increasing, we then have $C_\kappa(u)$ increasing for $\hat{\beta}^\top x = \sigma^{-1}(u) \in (-1, 1)$.

Proving the result for $\{k = 2, \ldots, K-1\}$ In addition, by our model assumption $x \sim N(0, I_d)$, we have that $\|x\|$ and $\frac{x}{\|x\|}$ are independent, and $\frac{x}{\|x\|} \sim S$ where S is a uniform distribution on the sphere in the d-dimensional space. As the monotonic transformations will not change the events defined by quantiles, and $\exp(-\|x\|^2/2)$ is a monotonic function in $\|x\|$, for the simplicity of presentation we use $a(X) = \|X\|$ in the rest of this proof. As a result, given $\|x\| = a$, we have

$$\hat{\beta}^\top x \mid \|x\| = a \stackrel{d}{=} a \cdot \hat{\beta}^\top S = a \cdot \|\hat{\beta}\| \cdot S_1,$$

where S_1 is the first coordinate of S.

Consequently, if we further condition on the event where $\hat{\beta}^\top x > 0$ (as we assume $u > 0$ throughout Section 3.3), we have

$$\hat{\beta}^\top x \overset{d}{=} a \cdot \|\hat{\beta}\| \cdot S_1 \mid S_1 > 0 \overset{d}{=} a \cdot \|\hat{\beta}\| \cdot \frac{Z_1}{\sqrt{Z_1^2 + Q}} \to a \cdot R^* \cdot \frac{Z_1}{\sqrt{Z_1^2 + Q}},$$

where $Q \sim \chi^2_{p-1}$, $Z_1 \sim N(0,1)$ and they are independent.

Due to the monotonicity of $C_\kappa(u)$ on u, we have that for any $a_1 > a_2$,

$$C_\kappa(u) \mid \|x\| = a_1 \overset{d}{>} C_\kappa(u) \mid \|x\| = a_2,$$

where the notation $\overset{d}{>}$ denotes stochastic dominance.

Consequently, we have

$$\mathbb{E}_u[u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u) \mid a(X) \in [a_{k-1}, a_k]] < \mathbb{E}_u[u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u) \mid a(X) \in (a_k, a_{k+1}]],$$

for $k = 2, .., K-1$.

Proving the result for $k = K$ To complete the proof, it suffices to show that the inequality is also true for Kth quantile:

$$\mathbb{E}_u[u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u) \mid a(X) \in [a_{K-1}, a_K]] < \mathbb{E}_u[u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u) \mid a(X) \in (a_k, a_{k+1}]],$$

which is equivalent to

$$\mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \cdot \mathbf{1}\{a(X) \in [a_{K-1}, a_K]\}] < \mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \cdot \mathbf{1}\{a(X) \in (a_k, a_{k+1}]\}].$$

In the above inequality, the right hand side can be decomposed into

$$\begin{aligned}
&\mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \cdot \mathbf{1}\{a(X) \in (a_k, a_{k+1}]\}] \\
=&\mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \cdot \mathbf{1}\{a(X) \in [a_K, 2p]\}] \\
&+ \mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \cdot \mathbf{1}\{a(X) \in [2p, a_{K+1}]\}].
\end{aligned}$$

Denote the α quantile of χ^2_p by $\chi^2_{\alpha,p}$. We then have $a_k = \chi^2_{\frac{k}{K+1},p}$. We further decompose the equation into

$$\begin{aligned}
&\mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \cdot \mathbf{1}\{a(X) \in [a_K, 2p]\}] \\
=&\mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \cdot \mathbf{1}\{a(X) \in [a_K, \chi^2_{\frac{K+\delta}{K+1},p}]\}] \\
&+ \mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \cdot \mathbf{1}\{a(X) \in [\chi^2_{\frac{K+\delta}{K+1},p}, 2p]\}].
\end{aligned}$$

In the following, we proceed to prove

$$\mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \cdot \mathbf{1}\{a(X) \in [\chi^2_{\frac{K+\delta}{K+1},p}, 2p]\}] > \mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \cdot \mathbf{1}\{a(X) \in [a_{K-1}, a_K\}]. \tag{12}$$

We now use the approximation of the chi-square quantile: when $p \to \infty$, we have

$$a_K = \frac{1}{2}(z_{\frac{K}{K+1}} + \sqrt{2p})^2 + o(1), \text{ and } \chi^2_{\frac{K+\delta}{K+1},p} = \frac{1}{2}(z_{\frac{K+\delta}{K+1}} + \sqrt{2p})^2 + o(1),$$

where z_α denotes the α-quantile of a standard normal random variable.

Then

$$\chi^2_{\frac{K+\delta}{K+1},p} - a_K = \frac{1}{2}(z_{\frac{K+\delta}{K+1}} - z_{\frac{K}{K+1}})(z_{\frac{K+\delta}{K+1}} + z_{\frac{K}{K+1}} + 2\sqrt{2p}).$$

Using the fact that $z_{1-\frac{1}{K}} = \sqrt{2\log K} + o(1)$ for $K \to \infty$, then we have

$$z_{\frac{K+\delta}{K+1}} - z_{\frac{K}{K+1}} = \frac{-\log(1-\delta)}{\sqrt{2\log K}} + o(1).$$

In addition, for any $a \in [\chi^2_{\frac{K+\delta}{K+1},p}, 2p]$ and $a' \in [a_{K-1}, a_K]$, we have

$$\begin{aligned}&\mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \mid a(X) = a] - \mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \mid a(X) = a']\\ \geq& C(z_{\frac{K+\delta}{K+1}} - z_{\frac{K}{K+1}}),\end{aligned}$$

for some universal constant C.

Therefore

$$\begin{aligned}&\mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \mid a(X) \in [\chi^2_{\frac{K+\delta}{K+1},p}, 2p]] - \mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \mid a(X) \in [a_{K-1}, a_K]]\\ \geq& C(z_{\frac{K+\delta}{K+1}} - z_{\frac{K}{K+1}}).\end{aligned}$$

Then

$$\begin{aligned}&\mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \cdot \mathbf{1}\{a(X) \in [\chi^2_{\frac{K+\delta}{K+1},p}, 2p]\}]\\ =&\mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \mid a(X) \in [\chi^2_{\frac{K+\delta}{K+1},p}, 2p]] \cdot \mathbb{P}(a(X) \in [\chi^2_{\frac{K+\delta}{K+1},p}, 2p])\\ \geq& \Big(\mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \mid a(X) \in [a_{K-1}, a_K]] + C(z_{\frac{K+\delta}{K+1}} - z_{\frac{K}{K+1}})\Big) \cdot (\frac{1}{K} - \frac{\delta}{K+1} + o(\frac{\delta}{K+1}))\\ =&\mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \cdot \mathbf{1}\{a(X) \in [a_{K-1}, a_K\}]\\ &+ C(z_{\frac{K+\delta}{K+1}} - z_{\frac{K}{K+1}}) - (1 + o(1))\frac{\delta}{K+1} \cdot \mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \mid a(X) \in [a_{K-1}, a_K]].\end{aligned}$$

The last equality uses the fact that $\mathbb{P}(a(X) \in [a_{K-1}, a_K]) = 1/K$, and therefore

$$\mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \mid a(X) \in [a_{K-1}, a_K] \cdot \frac{1}{K} = \mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \cdot \mathbf{1}\{a(X) \in [a_{K-1}, a_K\}]$$

Then use the fact that $|\mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \mid a(X) \in [a_{K-1}, a_K]]| = O(1)$ and we choose $\delta = o(1/\log K)$ so

$$\frac{\delta}{K} = o(|\frac{\log(1-\delta)}{\sqrt{\log K}}|).$$

Consequently,

$$C(z_{\frac{K+\delta}{K+1}} - z_{\frac{K}{K+1}}) - (1 + o(1))\frac{\delta}{K+1} \cdot \mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \mid a(X) \in [a_{K-1}, a_K]] > 0,$$

which implies

$$\mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \cdot \mathbf{1}\{a(X) \in [\chi^2_{\frac{K+\delta}{K+1},p}, 2p]\}] \geq \mathbb{E}_u[(u - \mathbb{P}(Y = 1 \mid \hat{\mathbb{P}}_1(X) = u)) \cdot \mathbf{1}\{a(X) \in [a_{K-1}, a_K\}].$$

We, therefore, prove equation 12 and complete the proof. □

K.3 Theory: Atypicality-Aware Recalibration

We analyze atypicality-aware recalibration under the same learning setting and generative model as Theorem 3.1. For a predictor $\hat{\mathbb{P}}$, let us denote its conditional calibration error at an atypicality level γ by $\text{CE}_\gamma(\hat{\mathbb{P}}) = \mathbb{E}[(\hat{\mathbb{P}}(Y = 1|X) - \mathbb{E}[Y|\hat{\mathbb{P}}(Y = 1|X)])^2|a(X) = \gamma]$.

Theorem K.2. *Consider the setting in Theorem 3.1. Let the temperature function* $\hat{\tau}(a(X)) = \arg\min_\tau \mathbb{E}[l(Y, \text{Softmax}(f(X)/\tau(a(X))))]$ *where* l *is the cross-entropy loss, and* $\hat{\mathbb{P}}_{\text{ATS}}(X) = \text{Softmax}(f(X)/\hat{\tau}(a(X)))$. *Then*

$$\text{CE}_\gamma(\hat{\mathbb{P}}_{\text{ATS}}) \leq \min\{\text{CE}_\gamma(\hat{\mathbb{P}}_{\text{TS}}), \text{CE}_\gamma(\hat{\mathbb{P}})\}. \tag{13}$$

That is, for all atypicality values γ, The proof can be found in Appendix K.4.

K.4 PROOF OF THEOREM K.2

Theorem K.3 (Restatement of Theorem K.2)**.** *Consider the same setting as Theorem 3.1. Suppose the temperature function* $\hat{\tau}(a(X)) = \arg\min_\tau \mathbb{E}[l(Y, \text{Softmax}(f(X)/\tau(a(X))))]$ *with* l *being the cross entropy loss, and let* $\hat{\mathbb{P}}_{\text{ATS}}(X) = \text{Softmax}(f(X)/\hat{\tau}(a(X)))$*. Then*

$$\text{CE}_\gamma(\hat{\mathbb{P}}_{\text{ATS}}) \leq \min\{\text{CE}_\gamma(\hat{\mathbb{P}}_{\text{TS}}), \text{CE}_\gamma(\hat{\mathbb{P}})\}. \tag{14}$$

For a prediction function f, we first define the conditional mean squared error of f at an atypicality level γ by $\text{MSE}_\gamma(f) = \mathbb{E}[(f(X) - Y)^2 \mid a(X) = \gamma]$, then we have

$$\begin{aligned}
\text{MSE}_\gamma(f) - CE_\gamma(f) =& \mathbb{E}[(f(X) - Y)^2 \mid a(X) = \gamma] - \mathbb{E}[(f(X) - \mathbb{E}[Y \mid f(X), a(X) = \gamma])^2 \mid a(X) = \gamma] \\
=& \mathbb{E}[(\mathbb{E}[Y \mid f(X), a(X) = \gamma] - Y) \cdot (2f(X) - \mathbb{E}[Y \mid f(X), a(X) = \gamma] - Y) \mid a(X) = \gamma] \\
=& \mathbb{E}[(\mathbb{E}[Y \mid f(X), a(X) = \gamma] - Y) \cdot (\mathbb{E}[Y \mid f(X), a(X) = \gamma] - Y) \mid a(X) = \gamma] \\
&+ 2\mathbb{E}[(\mathbb{E}[Y \mid f(X), a(X) = \gamma] - Y) \cdot (f(X) - \mathbb{E}[Y \mid f(X), a(X) = \gamma])) \mid a(X) = \gamma]
\end{aligned}$$

Since

$$\begin{aligned}
&\mathbb{E}[Y\mathbb{E}[Y \mid f(X), a(X) = \gamma] \mid a(X) = \gamma] \\
=& \mathbb{E}_{f(X)|a(X)=\gamma}\mathbb{E}[Y\mathbb{E}[Y \mid f(X), a(X) = \gamma] \mid f(X), a(X) = \gamma]] \\
=& \mathbb{E}[(\mathbb{E}[Y \mid f(X), a(X) = \gamma])^2 \mid a(X) = \gamma],
\end{aligned}$$

we have

$$\mathbb{E}[(\mathbb{E}[Y \mid f(X), a(X) = \gamma] - Y) \cdot (f(X) - \mathbb{E}[Y \mid f(X), a(X) = \gamma])) \mid a(X) = \gamma] = 0,$$

and therefore

$$\text{MSE}_\gamma(f) - CE_\gamma(f) = \mathbb{E}[(\mathbb{E}[Y \mid f(X), a(X) = \gamma] - Y)^2 \mid a(X) = \gamma]$$

Now that $\hat{\mathbb{P}}_{ATS}(\hat{\mathbb{P}}(x), a(x))$ is monotonic on the $\hat{\mathbb{P}}(x)$, we have

$$\mathbb{E}[Y \mid \hat{\mathbb{P}}(x), a(X) = \gamma] = \mathbb{E}[Y \mid \hat{\mathbb{P}}_{ATS}(\hat{\mathbb{P}}(x), a(X)), a(X) = \gamma],$$

implying

$$\text{MSE}_\gamma(\hat{\mathbb{P}}_{ATS}) - CE_\gamma(\hat{\mathbb{P}}_{ATS}) = \text{MSE}_\gamma(\hat{\mathbb{P}}) - CE_\gamma(\hat{\mathbb{P}}). \tag{15}$$

Similarly, we have

$$\text{MSE}_\gamma(\hat{\mathbb{P}}_{TS}) - CE_\gamma(\hat{\mathbb{P}}_{TS}) = \text{MSE}_\gamma(\hat{\mathbb{P}}) - CE_\gamma(\hat{\mathbb{P}}). \tag{16}$$

In the following, we will show that

$$\text{MSE}_\gamma(\hat{\mathbb{P}}_{ATS}) < \min\{\text{MSE}_\gamma(\hat{\mathbb{P}}_{TS}), \text{MSE}_\gamma(\hat{\mathbb{P}})\}. \tag{17}$$

First, as we consider the binary classification setting, with l being the cross entropy loss, we have

$l(Y, \text{Softmax}(f(X)/\tau(a(X)))) = Y\log(\sigma(f_1(X)/\tau(a(X))) + (1-Y)\log(1-\sigma(f_1(X)/\tau(a(X))),$

where $\sigma(x) = 1/(1 + e^x)$.

Then, by the definition of $\hat{\tau}(a(X))$, we have that

$$\begin{aligned}
\hat{\tau}(a(X)) =& \arg\min_\tau \mathbb{E}[Y\log(\sigma(f_1(X)/\tau(a(X))) + (1-Y)\log(1-\sigma(f_1(X)/\tau(a(X)))] \\
=& \arg\min_\tau \mathbb{E}[\mathbb{E}[Y\log(\sigma(f_1(X)/\tau(a(X))) + (1-Y)\log(1-\sigma(f_1(X)/\tau(a(X))) \mid a(X)]].
\end{aligned}$$

Taking the derivative on the last line and setting it to zero, we have

$$\mathbb{E}[\frac{Y}{\sigma(f_1(X)/\hat{\tau}(a(X))} - \frac{1-Y}{1-\sigma(f_1(X)/\hat{\tau}(a(X))} \mid a(X)] = 0,$$

implying

$$\mathbb{E}[\sigma(f_1(X)/\hat{\tau}(a(X)) \mid a(X)] = \mathbb{E}[Y \mid a(X)].$$

This makes the derivative of $\mathbb{E}[(Y - \sigma(f_1(X)/\tau(a(X))))^2 \mid a(X)]$ zero and therefore $\hat{\tau}(a(X))$ is also a minimizer of $\mathbb{E}[(Y - \sigma(f_1(X)/\tau(a(X))))^2 \mid a(X)]$:

$$\hat{\tau}(a(X)) = \arg\min_{\tau} \mathbb{E}[Y \log(\sigma(f_1(X)/\tau(a(X)))+(1-Y)\log(1-\sigma(f_1(X)/\tau(a(X)))] = \arg\min_{\tau} \mathbb{E}[(Y-\sigma(f_1(X)/\tau(a(X)))$$

Letting $g(\gamma) = \arg\min_c \mathbb{E}[(Y - \sigma(f_1(X)/c))^2 \mid a(X) = \gamma]$, we have that

$$g(a(X)) = \arg\min_{\tau} \mathbb{E}[(Y - \sigma(f_1(X)/\tau(a(X))))^2 \mid a(X)],$$

and therefore

$$g(a(X)) = \arg\min_{\tau} \mathbb{E}[\mathbb{E}[(Y - \sigma(f_1(X)/\tau(a(X))))^2 \mid a(X)] = \hat{\tau}(a(X)).$$

As a result,

$$\begin{aligned}
\text{MSE}_\gamma(\hat{\mathbb{P}}_{ATS}) =& \mathbb{E}[(\hat{\mathbb{P}}_{ATS}(X) - Y)^2 \mid a(X) = \gamma] \\
=& \mathbb{E}[(\hat{\mathbb{P}}_{ATS}(X) - Y)^2 \mid a(X) = \gamma] \\
=& \mathbb{E}[(\sigma(f_1(X)/\hat{\tau}(a(X)) - Y)^2 \mid a(X) = \gamma] \\
=& \mathbb{E}[(\text{Softmax}(\hat{\mathbb{P}}(X)/g(a(X)) - Y)^2 \mid a(X) = \gamma] \\
=& \arg\min_c \mathbb{E}[(\sigma(f_1(X)/c - Y)^2 \mid a(X) = \gamma] \\
\leq& \mathbb{E}[(\sigma(f_1(X) - Y)^2 \mid a(X) = \gamma] \\
=& \text{MSE}_\gamma(\hat{\mathbb{P}}).
\end{aligned}$$

Similarly, we have $\text{MSE}_\gamma(\hat{\mathbb{P}}_{ATS}) \leq \text{MSE}_\gamma(\hat{\mathbb{P}}_{TS})$, and therefore equation 17 holds.

Combining with equation 15 and equation 16, we have

$$CE_\gamma(\hat{\mathbb{P}}_{ATS}) \leq \min\{CE_\gamma(\hat{\mathbb{P}}_{TS}), CE_\gamma(\hat{\mathbb{P}})\}.$$

L Limitations

L.1 Quantifying Atypicality

Since we do not have access to the true distribution of $\mathbb{P}(X)$, we estimate it through the model, e.g. using the embeddings. This means we are capturing the atypicality not solely with respect to the training distribution but also the model. It is possible that a model that does not fit the data well and produces low-quality atypicality estimates. In general, we observe that our findings hold for large datasets and widely used models, and atypicality gives a semantically meaningful way to group data points qualitatively. However, initial explorations with smaller datasets resulted in cases with noisy estimates. Our findings suggest that we can unify the understanding and improve uncertainty quantification and recalibration methods with atypicality, however, practitioners should be careful about incorporating atypicality, as poor atypicality estimates can lead to worse performance.

L.2 Theoretical Analysis

Following the earlier work (Bai et al., 2021; Sur & Candès, 2019), we analyzed the calibration behavior of well-specified logistic regression. However, our empirical findings suggest that the phenomena are much more broadly applicable. We suggest that future work can analyze the behavior in more general settings to better understand the dynamics.