When Does Calibration Matter for Safe Model Routing? Conformal Risk Control Under Imperfect Gate Calibration

Iqtedar Uddin
Illinois Institute of Technology

Mazin Khider
Illinois Institute of Technology

André Bauer
Illinois Institute of Technology

Abstract

In safety-critical or regulated deployments, practitioners often wish to use simpler or interpretable surrogate models that approximate a more accurate but expensive black-box model, yet lack formal guarantees of when they are sufficiently accurate. Model routing addresses this by deciding, for each input, whether a surrogate can safely replace the expensive model without degrading predictions beyond a tolerance τ. A lightweight classifier (the *gate*) predicts from input features alone whether the surrogate is safe. A conformal procedure based on Clopper–Pearson bounds then selects a routing threshold on held-out data, guaranteeing that the violation rate among routed inputs is at most α with probability $1 - \delta$. We establish a formal separation between probabilistic calibration and distribution-free safety. While conformal thresholding guarantees validity regardless of the gate's Expected Calibration Error (ECE), we empirically demonstrate that applying post-hoc calibration (e.g., Beta scaling) stabilizes the conformal procedure. Calibration selectively filters borderline instances, reducing empirical violation rates closer to the target nominal level across 35 OpenML datasets, and yields a highly interpretable routing threshold.

1 Introduction

In high-stakes applications, practitioners deploy accurate black-box models but prefer simpler or interpretable surrogates for transparency or compliance. The key question is not whether the surrogate is globally accurate, but on which inputs it can be used without exceeding a degradation tolerance τ. Model routing addresses this by deciding, for each input, whether the surrogate is safe. A lightweight gate predicts whether the surrogate is safe; however, its predicted probabilities are often miscalibrated [Guo et al., 2017].

We show that probabilistic calibration is neither necessary nor sufficient for safe routing. Instead, we use conformal risk control [Bates et al., 2021] with Clopper–Pearson binomial bounds [Clopper and Pearson, 1934] to pick thresholds with exact finite-sample guarantees: $\Pr(V(t^*) > \alpha) \leq \delta$, no matter the gate's ECE. This separates two ideas that are easy to conflate:

- **Probabilistic calibration** (ECE $\to 0$): predicted probabilities match true frequencies. This controls routing precision, namely which inputs are routed and how clean the routed set is.
- **Conformal calibration** (Clopper–Pearson thresholding): distribution-free control of the violation rate. This controls validity: the safety guarantee holds at any ECE.

We test this separation by applying beta calibration [Kull et al., 2017] and temperature scaling [Guo et al., 2017] to our gate. The results show that recalibration does not increase coverage but does improve the quality of routing decisions: it selectively removes borderline-unsafe inputs from the routed set, reducing the violation rate and pushing the guarantee tighter toward the target δ.

We also derive a feasibility condition (Proposition 1) linking the gate's likelihood ratio to the base safe rate π and risk budget α. Our work builds on split conformal prediction [Vovk et al., 2005] and conformal risk control [Bates et al., 2021, Angelopoulos et al., 2022]. Unlike selective prediction [Geifman and El-Yaniv, 2017], which abstains on hard inputs, we route to a cheaper model. Dynamic input-based routing methods such as SkipNet [Wang et al., 2018] and BlockDrop [Wu et al., 2018] also make routing decisions before full computation, but they optimize accuracy–efficiency trade-offs without

Workshop "Towards Trustworthy Predictions: Theory and Applications of Calibration for Modern AI" at AISTATS 2026, Tangier, Morocco.

providing distribution-free guarantees on the conditional error of the routed subset. Unlike cascading [Chen et al., 2023], which requires running the cheap model on every input, our gate decides from input features alone before either model runs and provides explicit risk control. Unlike agreement-based routing (e.g., $|f(x) - g(x)| \leq \tau$), which measures consistency with f rather than ground truth Ribeiro et al. [2016], Lundberg and Lee [2017], our degradation criterion controls surrogate error relative to true labels. Conformal thresholding then yields a finite-sample bound on the violation rate over routed inputs.

2 Setup and Method

Problem. Let f be an accurate black-box model and g a simpler surrogate. For input x with label y, define degradation $d(x, y) = |y - g(x)| - |y - f(x)|$. An input is *safe* if $d(x, y) \leq \tau$. Let $Y = \mathbf{1}[d \leq \tau]$ and $\pi = \Pr(Y{=}1)$. A routing policy $\hat{\pi}(x) = \mathbf{1}[s(x) \geq t]$ sends input x to the surrogate when gate score $s(x)$ exceeds threshold t. The violation rate among routed inputs is $V(t) = \Pr(Y{=}0 \mid s(X) \geq t)$. We want the threshold t that maximizes coverage $\Pr(s(X) \geq t)$ subject to $V(t) \leq \alpha$.

Gate. We train a logistic regression gate to predict Y from input features x, with Platt scaling via cross-validated sigmoid calibration [Platt, 1999]. We use logistic regression deliberately, as it represents a relatively weak gate, to demonstrate that the framework does not rely on strong classification performance.

Conformal threshold selection. Following conformal risk control [Bates et al., 2021, Angelopoulos et al., 2022], we use a held-out calibration set of n exchangeable points $\{(s_i, Y_i)\}_{i=1}^{n}$ to select the routing threshold. We pick the lowest threshold where the Clopper–Pearson [Clopper and Pearson, 1934] upper confidence bound on the violation rate is at most α:

$$t^* = \min\big\{t : \; B_{1-\delta}^{-1}(k(t) + 1, \; n(t) - k(t)) \leq \alpha\big\}, \quad (1)$$

where $n(t) = |\{i : s_i \geq t\}|$, $k(t) = |\{i : s_i \geq t, Y_i = 0\}|$, and $B_{1-\delta}^{-1}$ is the $1 - \delta$ quantile of the Beta distribution. If no valid t exists, we set $t^* = \infty$ and route nothing.

This yields an exact finite-sample guarantee: $\Pr(V(t^*) > \alpha) \leq \delta$ for any gate and any data distribution, requiring only exchangeability between calibration and test data.

Data splits. We use a strict 55/15/15/15 split: train (f, g, gate), validation (surrogate depth and recalibration fitting), calibration (t^*), test (evaluate). No information leaks between splits.

3 Routing Feasibility

Proposition 1 (Feasibility condition)**.** *The constraint $V(t) \leq \alpha$ can be met with positive coverage if and only if there exists t with:*

$$\frac{\mathrm{TPR}(t)}{\mathrm{FPR}(t)} \geq \frac{(1-\pi)(1-\alpha)}{\pi\alpha} \triangleq C(\pi, \alpha). \quad (2)$$

Proof. By Bayes' rule, $V(t) = \frac{(1-\pi)\mathrm{FPR}(t)}{(1-\pi)\mathrm{FPR}(t) + \pi\,\mathrm{TPR}(t)}$. Setting $V(t) \leq \alpha$ and rearranging gives (2). □

Three factors interact here: (i) π, set by τ and the model pair, (ii) α, the risk budget, and (iii) the gate's likelihood ratio, which depends on its ranking quality. The condition has the form of a likelihood-ratio constraint in the Neyman–Pearson sense [Neyman and Pearson, 1933]. When π is small (strict τ), $C(\pi, \alpha)$ is large and only strong gates can route. In practice, π matters more than gate quality. When most inputs are safe, even a weak gate works.

3.1 Critical AUC

Theorem 1 (Sufficient AUC for feasibility)**.** *Define the critical AUC:*

$$\Phi_c(\pi, \alpha) = \min\left(1, \; \frac{C(\pi, \alpha)}{2}\right) = \min\left(1, \; \frac{(1-\pi)(1-\alpha)}{2\pi\alpha}\right). \quad (3)$$

If $\mathrm{AUC} \geq \Phi_c(\pi, \alpha)$*, then there exists a routing threshold t satisfying $V(t) \leq \alpha$ with positive coverage.*

Proof Sketch. If no threshold satisfies the feasibility ratio, then $\mathrm{TPR}(u) < C(\pi, \alpha)\, u$ for all $u \in (0, 1]$. Integrating over the ROC curve yields

$$\mathrm{AUC} < \frac{C(\pi, \alpha)}{2}.$$

Taking the contrapositive gives the sufficient AUC condition. Full details are provided in Appendix A. □

4 Experimental Results

Setup. We evaluate on 35 datasets from LCBench [Zimmer et al., 2021, Vanschoren et al., 2014], each with 2000 configurations. We treat random forest [Breiman, 2001] with 1500 trees as a black box. The surrogate is a decision tree, with depth selected on the validation set. We test six tolerance levels $\tau \in \{-1.5, -1.0, 0, 0.5, 1.0, 2.0\}$ and ten risk levels $\alpha \in \{0.05, \ldots, 0.80\}$, with $\delta = 0.10$.

Baselines. We compare against several baselines: **Naive** routing at a fixed threshold ($t = 0.5$, no formal guarantee), an **Oracle** that routes iff $d(x) \leq \tau$ (upper bound on achievable coverage), **Always-BB / Always-CM** (never/always route), and **Random** routing at conformal's coverage level. We also compare against **Regression conformal**, which trains Ridge regression to predict $d(x)$ from input features, computes a split conformal prediction interval, and routes when the upper bound $\leq \tau$. We use Ridge (a linear model) to match the linear gate, ensuring any performance difference reflects the classification-vs-regression formulation rather than model capacity.

Architectural robustness. Repeating the conformal guarantee experiments with XGBoost and MLP as the expensive model yields similar behavior: violation rates remain near $\delta = 0.10$ across α values, and naive thresholding continues to fail substantially more often. This indicates that the validity of input-based conformal routing is not specific to tree ensembles.

Table 1: Gate properties across tolerance levels. AUC is moderate (∼0.58) and ECE is high (0.07–0.25), yet conformal routing maintains guarantees. Coverage and violation at $\alpha = 0.2$, means over 35 datasets.

τ	π	AUC	ECE	Conformal α=0.2	
				Cov.	Viol.
−1.5	0.13	0.71	0.07	0.000	—
−1.0	0.16	0.70	0.09	0.000	—
0.0	0.43	0.54	0.25	0.003	0.03
0.5	0.66	0.57	0.20	0.243	0.15
1.0	0.74	0.58	0.17	0.446	0.14
2.0	0.82	0.59	0.12	0.661	0.12

Table 2: Fraction of (dataset, τ) pairs where the violation rate exceeds α. Conformal remains near $\delta = 0.10$. Naive thresholding fails much more often.

α	Gate Conf.	Reg. Conf.	Naive
0.10	7/34 (21%)	31/50 (62%)	151/172 (88%)
0.20	8/66 (12%)	57/102 (56%)	128/172 (74%)
0.30	6/82 (7%)	80/144 (56%)	106/172 (62%)
0.50	8/112 (7%)	64/181 (35%)	62/172 (36%)

Gate properties (Table 1): At strict tolerances ($\tau \leq 0$), π is low and the conformal procedure abstains, yielding zero coverage because (2) cannot be satisfied. At practical tolerances ($\tau \geq 0.5$), π exceeds 0.66 and the gate routes 24–66% of inputs at $\alpha = 0.2$, despite moderate AUC and high ECE.

Guarantee reliability (Table 2): Conformal routing typically violates α on 7–12% of datasets, matching our $\delta = 0.10$ confidence level. While violations reach 21% at the strictest risk level ($\alpha = 0.10$), likely due to finite-sample noise and calibration-test mismatch, the method remains highly reliable compared to naive thresholding (36–88%) and conformal regression baselines (35–62%).

Calibration–efficiency separation (Figure 1). The left panel shows gate ECE peaking at $\tau = 0$ where class balance is worst. The right panel shows that conformal violation stays well below $\alpha = 0.1$ at every τ, while naive violations are 4–8× higher.

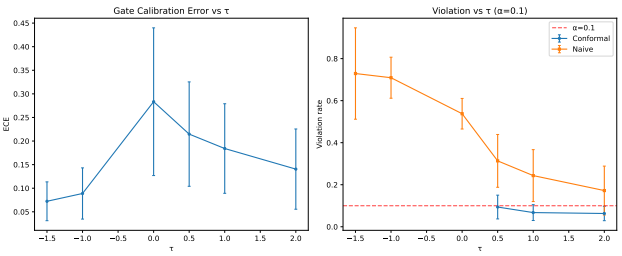

Figure 1: Left: gate ECE varies from 0.07 to 0.28 across τ. Right: despite high ECE, conformal violation remains below $\alpha = 0.1$. Naive violations are 4–8× higher.

4.1 Coverage–Violation Pareto Frontier

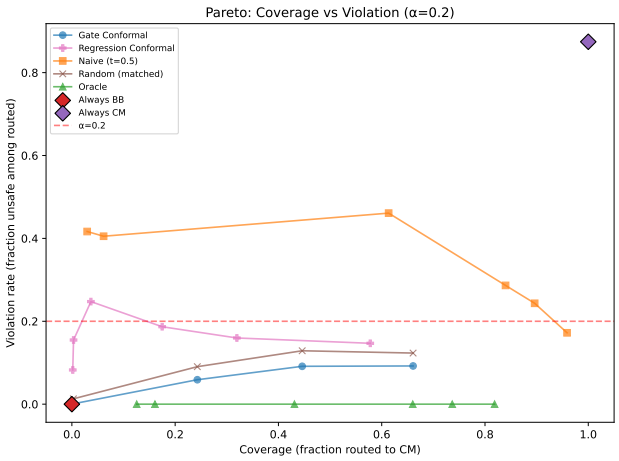

Figure 2: Coverage vs. violation rate at $\alpha = 0.2$. Each point is a τ value, averaged over 35 datasets. Gate conformal stays below α; regression conformal and naive operate above it.

Figure 2 shows the Pareto frontier at $\alpha = 0.2$. Gate conformal traces a frontier consistently below the α

line, reaching 66% coverage at $\tau = 2.0$ with only 12% mean violation. Regression conformal achieves lower coverage (58%) at *higher* violation (15%), and naive thresholding violates α at every operating point. Gate conformal dominates on both axes at all practical τ values (≥ 0.5). At low τ, regression conformal routes a tiny fraction (0.1–3%) with massive violation (8–51%), while gate conformal correctly abstains. This illustrates the value of the conformal abstention mechanism: routing nothing is preferable to routing with violated guarantees.

The Pareto results confirm that gate conformal controls violations while maximizing coverage. We next ask: does improving the gate's calibration further improve these routing decisions?

5 The Calibration–Efficiency Separation

5.1 Calibration, Validity, and Efficiency

The conformal guarantee holds for any gate, regardless of calibration. This yields an exact finite-sample guarantee: $\Pr(V(t^*) > \alpha) \leq \delta$ for any gate and any data distribution, requiring only exchangeability between calibration and test data. A gate with ECE = 0.25 gets the same safety guarantee as one with ECE = 0. The procedure compensates automatically: if scores are shifted or compressed, it pushes t^* upward until the bound is satisfied.

Calibration does affect which inputs get routed. A perfectly calibrated gate has $s(x) = \Pr(Y{=}1 \mid x)$, so the conformal procedure picks $t^* \approx 1 - \alpha$ and routes all inputs scoring above this. A miscalibrated gate distorts this mapping. An overconfident gate assigns unsafe inputs scores that are too high, allowing them into the routed set. An underconfident gate assigns safe inputs scores that are too low, excluding them from the routed set. In both cases the guarantee holds, but the routing decisions differ.

5.2 Empirical Test: Post-Hoc Recalibration

To test this, we applied three post-hoc recalibration methods to the gate outputs: beta calibration [Kull et al., 2017], temperature scaling [Guo et al., 2017], and isotonic regression [Zadrozny and Elkan, 2002]. Each recalibrator was fit on a held-out validation set and then applied to the calibration and test scores before conformal threshold selection. The calibration set itself was used only for conformal thresholding and was not used to fit the recalibration models.

Table 3 shows that beta calibration reduces ECE from 0.184 to 0.060 at $\tau = 1.0$, a 3× improvement. Coverage barely changes (−2.5pp for beta/temperature), but violation drops from 9.1% to 8.5%. This means the inputs removed by recalibration were disproportionately unsafe. At $\tau = 1.0$ with approximately 300 test points, the Platt gate routes about 134 inputs, of which 12 are unsafe. Beta calibration routes about 126 inputs, of which 11 are unsafe. Roughly 8 inputs left the routed set, of which 1–2 were unsafe, corresponding to an unsafe rate of ∼15–25% among the removed inputs, much higher than the 9% overall. Recalibration is selectively filtering out the worst routing decisions.

Table 3: Effect of post-hoc recalibration at $\alpha = 0.2$ (means over 35 datasets). Recalibration reduces ECE by 3×. Coverage changes minimally, while violation decreases, particularly for isotonic regression.

τ	Recalibration	ECE	Cov.	Viol.
0.5	None (Platt)	.215	.243	.059
0.5	Beta	.066	.243	.059
0.5	Temperature	.077	.243	.059
0.5	Isotonic	.050	.216	.043
1.0	None (Platt)	.184	.446	.091
1.0	Beta	.060	.421	.085
1.0	Temperature	.071	.421	.085
1.0	Isotonic	.048	.398	.072
2.0	None (Platt)	.141	.661	.092
2.0	Beta	.052	.651	.095
2.0	Temperature	.061	.644	.091
2.0	Isotonic	.036	.622	.077

Table 4: Fraction of active (dataset, τ) pairs violating α by recalibration method. Beta and temperature remain near $\delta = 0.10$. Isotonic is more conservative.

Recalibration	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$
None (Platt)	20.6%	12.1%	7.3%
Beta	21.2%	10.9%	8.5%
Temperature	21.2%	11.1%	7.4%
Isotonic	4.0%	1.6%	2.6%

Table 4 shows this at the dataset level. The conformal guarantee is a ceiling: $V(t^*) \leq \alpha$ with probability $1 - \delta$. Better calibration pushes the realized violation rate further below that ceiling. At $\alpha = 0.2$, the fraction of problem instances violating α drops from 12.1% (Platt) to 10.9% (beta), bringing it closer to the target $\delta = 0.10$. Isotonic is more aggressive: 1.6% violation, well below δ, meaning it is being overly conservative and wasting coverage. Beta calibration hits the right balance: tighter guarantees without unnecessary conservatism.

After recalibration, the threshold t^* also becomes more interpretable. For a perfectly calibrated gate, $t^* \approx 1 - \alpha = 0.8$ carries direct meaning: "route inputs where the estimated safety probability exceeds 80%." With the Platt-scaled gate, the threshold lives on a

distorted scale. Beta recalibration moves t^* closer to this interpretable regime, which is useful for auditing and debugging.

5.3 Decomposing the Efficiency Gap

At $\tau = 1.0$ and $\alpha = 0.2$, the oracle routes 74% of inputs, corresponding to all safe ones. Our Platt-scaled gate routes 45%. After beta calibration, it routes 42%. The 29pp gap between gate and oracle is almost entirely from ranking limitations: with AUC = 0.58, the gate cannot reliably separate safe from unsafe inputs regardless of calibration. Recalibration does not recover this gap. Instead, it refines the decisions being made within the gate's ranking capacity.

6 Discussion

The value of calibration for routing. Our results give a precise answer to "when does calibration matter?" Calibration does not affect whether the safety guarantee holds. Conformal procedures ensure validity regardless. But calibration improves the precision of routing: better-calibrated scores selectively filter out the inputs most likely to be unsafe, reducing the actual violation rate below the α ceiling without meaningful coverage loss. For safety-critical applications, this matters even when the formal guarantee is satisfied. Calibration improves interpretability of gate scores. After beta recalibration, $s(x) = 0.8$ roughly corresponds to an 80% chance of safety, which is useful for auditing and combining with other decision criteria.

In this sense, the routed subset defines a formally controlled deployment region for the surrogate model under tolerance τ and risk budget (α, δ). Practitioners may interpret this as a finite-sample guarantee that, with probability at least $1 - \delta$, the surrogate will not exceed the allowed degradation rate on routed inputs. This connects calibration analysis with deployment-level reliability.

Where calibration falls short. The efficiency gap between our gate and the oracle is almost entirely from ranking quality, not calibration error. Reducing ECE from 0.18 to 0.06 recovers at most 2.5pp of coverage. For decision tasks like routing, calibration methods that also improve discrimination would be most valuable. Standard post-hoc methods (temperature scaling, beta calibration, isotonic regression) are monotone transforms that fix calibration while leaving ranking unchanged. An interesting direction is to test whether a deliberately underconfident gate, such as raw logistic regression without Platt scaling, responds differently. In that case, recalibration could push scores upward and increase coverage, rather than decrease it as in our overconfident setting.

Limitations. The guarantee is marginal over the calibration/test split and does not ensure conditional validity on specific input subregions. In particular, subpopulations with distinct feature distributions may exhibit higher violation rates, even when the overall guarantee holds. Extending the framework to conditional or subgroup-aware conformal guarantees is an important direction for future work.

We evaluate on tabular regression tasks using a random forest as the primary black box model and a decision tree surrogate, with additional robustness experiments using XGBoost [Chen and Guestrin, 2016] and MLPs [Goodfellow et al., 2016]. While this controlled setting isolates the routing mechanism, broader evaluation on vision, NLP, and large-scale models would be needed to assess generality.

Finally, our experiments assume exchangeability between calibration and test data. Performance under distribution shift, where exchangeability is violated, remains an open question.

References

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control, 2022. URL `https://arxiv.org/abs/2110.01052`.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6), 2021. doi: 10.1145/3478535.

Leo Breiman. Random forests. *Machine Learning*, 45 (1):5–32, 2001. doi: 10.1023/A:1010933404324.

Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance, 2023. URL `https://arxiv.org/abs/2305.05176`.

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*, pages 785–794, 2016. doi: 10.1145/2939672.2939785.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330, 2017.

Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 623–631, 2017.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*, 231:289–337, 1933. doi: 10.1098/rsta.1933.0009.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74, 1999.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016. doi: 10.1145/2939672.2939778.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2014. doi: 10.1145/2641190.2641198.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. doi: 10.1007/b106715.

Xin Wang, Fisher Yu, Zi-Yi Dou, Trevor Darrell, and Joseph E. Gonzalez. Skipnet: Learning dynamic routing in convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 409–424, 2018.

Zuxuan Wu, Tushar Nagarajan, Abhishek Kumar, Steven Rennie, Larry S. Davis, Kristen Grauman, and Rogerio Feris. Blockdrop: Dynamic inference paths in residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8817–8826, 2018.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699, 2002. doi: 10.1145/775047.775151.

Lucas Zimmer, Marius Lindauer, and Frank Hutter. Auto-PyTorch: Multi-fidelity metalearning for efficient and robust AutoDL. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(9):3079–3090, 2021. doi: 10.1109/TPAMI.2021.3067763.

A Proof of Theorem 1

The ROC curve $\{(\mathrm{FPR}(t), \mathrm{TPR}(t)) : t \in \mathbb{R}\}$ satisfies $\mathrm{AUC} = \int_0^1 \mathrm{TPR}(u)\, du$ where $u = \mathrm{FPR}$. Suppose no point on the ROC satisfies (2), i.e., $\mathrm{TPR}(u) < C(\pi, \alpha) \cdot u$ for all $u \in (0, 1]$. Then:

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(u)\, du < \int_0^1 C \cdot u\, du = \frac{C}{2}.$$

Taking the contrapositive: if $\mathrm{AUC} \geq C/2 = \Phi_c(\pi, \alpha)$, then there exists a point satisfying (2). Since $\mathrm{AUC} \leq 1$, we clip at 1. Note this is a necessary condition for infeasibility, not sufficient for feasibility: $\mathrm{AUC} \geq \Phi_c$ guarantees some point on the ROC has high enough likelihood ratio, but a non-concave ROC could still have that point at a degenerate threshold.

B Additional Results

Regression conformal baseline. A regression conformal baseline using Ridge regression to predict degradation $d(x)$ with split conformal intervals, routing when the upper bound is below τ, violates α on 35–62% of instances (Table 2), substantially more often than gate conformal. This occurs because predicting the magnitude of degradation from input features is more difficult than predicting a binary safe or unsafe label. Figure 3 visualizes this across the full α range: gate conformal stays near the $\delta = 0.10$ line, while regression conformal and naive thresholding remain far above it, converging only at very permissive α.

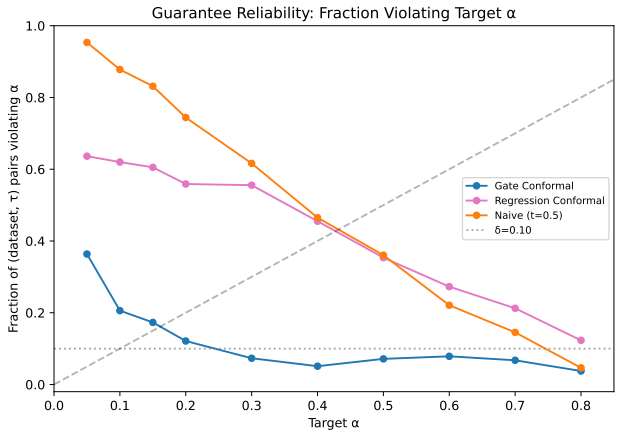

Figure 3: Fraction of (dataset, τ) pairs violating target α. Gate conformal stays near the $\delta = 0.10$ line across all α values.

C Additional Model Pair: XGBoost vs. Decision Tree

To verify that our conclusions are not specific to the random forest / decision tree pairing used in the main text, we repeated the full conformal routing evaluation using XGBoost as the expensive model and a decision tree as the surrogate on the same 35 OpenML regression datasets. All data splits, calibration sizes, tolerance levels τ, and risk levels α were kept identical.

Guarantee reliability. The conformal guarantee remains well-calibrated across risk levels:

- $\alpha = 0.10$: 14.3% violation rate.
- $\alpha = 0.20$: 12.1% violation rate.
- $\alpha = 0.30$: 14.5% violation rate.
- $\alpha = 0.50$: 5.1% violation rate.

For $\alpha \geq 0.20$, violation rates remain close to the nominal confidence parameter $\delta = 0.10$, matching the behavior observed in the random forest setting. In contrast, regression conformal violates α at substantially higher rates (31%–37% for $\alpha \in \{0.10, 0.20, 0.30\}$).

Coverage comparison. Across practical operating points ($\tau \geq 0.5$), gate conformal achieves higher coverage than regression conformal while maintaining comparable or lower violation rates. For example, at $\tau = 1.0$, $\alpha = 0.2$, gate coverage is 0.419 compared to 0.310 for regression conformal.

These results indicate that the validity and coverage advantages of classification-based conformal routing are not specific to tree ensembles.

D Additional Model Pair: MLP vs. Decision Tree

We further repeated all experiments using a multilayer perceptron (MLP) as the expensive model and a decision tree surrogate, keeping all experimental settings identical.

Guarantee reliability. Violation rates remain near the target confidence level:

- $\alpha = 0.10$: 10.3% violation rate.
- $\alpha = 0.20$: 16.5% violation rate.
- $\alpha = 0.30$: 13.6% violation rate.
- $\alpha = 0.50$: 5.5% violation rate.

As in the main experiment, regression conformal violates α far more frequently (50%–71% for $\alpha \in \{0.10, 0.20, 0.30\}$).

Coverage comparison. At practical tolerances, gate conformal consistently achieves higher coverage. For example, at $\tau = 1.0$, $\alpha = 0.2$, gate coverage is 0.471 versus 0.342 for regression conformal, with lower violation.

These experiments show that the proposed routing framework is robust across tree-based, boosting-based, and neural network black-box models.