Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Lightweight Calibration-Selection Policy After a Probe Fit for Selective Classification on OpenML Tabular Tasks

Haolun Tang
Shanghai Dianji University
Jingyi Zhan
Shanghai Dianji University
Yan Feng
Shanghai Dianji University
Zhipeng Chen
chenzhipeng@sdju.edu.cn

Shanghai Dianji University

Article

Keywords: post-hoc calibration, selective classification, OpenML, abstention, calibration selection

Posted Date: September 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10426790/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Lightweight Calibration-Selection Policy After a Probe Fit for Selective Classification on OpenML Tabular Tasks

Haolun Tang[1], Jingyi Zhan[1], Yan Feng[1*], Zhipeng Chen[1*]

[1]School of Electronic and Information Engineering, Shanghai Dianji University, Shanghai, China.

*Corresponding author(s). E-mail(s): fengyan@sdju.edu.cn; chenzhipeng@sdju.edu.cn;
Contributing authors: tanghl026@163.com; 2274252753@qq.com;

Abstract

Post-hoc calibration is often treated as a default step after training probabilistic classifiers, but that default is less clear when calibration labels are limited and the downstream objective includes abstention quality as well as probability accuracy. This paper studies a lightweight calibration-selection policy for selective classification on OpenML-style tabular tasks, where a probe calibration fit is used to decide whether calibration should be retained or skipped for a given model-condition case. The empirical record covers 15 datasets, 652 completed base evaluation cases, and 1,304 non-raw rows with both negative log-likelihood (NLL) and area under the risk-coverage curve (AURC) available. Within this benchmark, both tested calibrators improve mean NLL relative to raw predictions on average, yet calibration can also worsen mean AURC, so better probability fit does not imply better abstention behavior. The analysis therefore evaluates both row-level gates and a deployed case-level action selector that chooses exactly one action from {raw, temperature scaling, Dirichlet calibration}. The best joint case-level selector chooses a non-raw calibrator in 55.8% of completed cases, reaches mean ΔNLL $\mathbf{-0.0404}$, mean ΔAURC $\mathbf{-1.20 \times 10^{-4}}$, and reduces NLL harm to 6.1%. Among the tested meta-models, logistic regression is the strongest candidate, outperforming a decision tree under NLL, AURC, and a joint selection criterion. The evidence supports framing calibration on small tabular validation splits as a conditional probe-and-select intervention rather than a universally beneficial or purely compute-saving post-processing step.

Keywords: post-hoc calibration, selective classification, OpenML, abstention, calibration selection

1 Introduction

Post-hoc calibration is widely treated as a routine final step for probabilistic classification because many modern learners produce confidence scores that are useful for ranking yet systematically distorted as probability estimates (Zadrozny and Elkan 2001, 2002; Niculescu-Mizil and Caruana 2005; Guo et al. 2017; Kull et al. 2019). That default becomes harder to justify when the downstream objective includes abstention. In selective classification, the system must decide not only which class to predict but also which inputs should be deferred when uncertainty is high, so performance depends on the coverage–risk trade-off induced by the confidence scores rather than on probabilistic fidelity alone (Chow 1970; Bartlett and Wegkamp 2008; El-Yaniv and Wiener 2010; Geifman and El-Yaniv 2017, 2019; Chen et al. 2019). The practical hope is that better calibrated probabilities should also support better abstention decisions. Yet that implication is not automatic, especially on small tabular validation splits where calibration labels are scarce and a post-hoc transform can improve negative log-likelihood while leaving the ranking structure relevant to selective prediction unchanged or even degraded (Naeini et al. 2015; Kull et al. 2017; Guo et al. 2017).

This tension motivates a more conditional view of calibration in the OpenML-style tabular regime studied here. Researchers often train a base classifier, hold out a modest calibration split, and then apply a standard post-hoc method with little task-specific justification, even though this is precisely the regime in which additional score transformation may be unstable (Vanschoren et al. 2013; Casalicchio et al. 2019). When validation labels are few, every calibration attempt consumes limited evidence, and the relevant downstream criterion may combine probability estimation with abstention quality rather than privileging either one in isolation (Zadrozny and Elkan 2002; Naeini et al. 2015; Kull et al. 2017). The central problem in this paper is therefore narrower than asking whether calibration improves probabilistic predictions in principle, which is already well established. Instead, for the two studied calibrators on small tabular tasks, the question is whether applying calibration is worthwhile once its effect is judged jointly through negative log-likelihood (NLL) and area under the risk-coverage curve (AURC). The empirical record analyzed here comprises 15 datasets, 652 completed base evaluation cases, and 1,304 non-raw model-condition rows with both NLL and AURC available, making it possible to examine calibration as a contingent intervention rather than a universal default.

Existing research supplies the main ingredients for this question, but not the particular decision framing pursued here. The calibration literature has developed a range of post-hoc mappings, from early multiclass probability transformations to Bayesian binning and beta calibration (Zadrozny and Elkan 2001, 2002; Naeini et al. 2015; Kull et al. 2017, 2019). Parallel work on reject-option and selective classification has shown that abstention quality depends on how well confidence scores separate correct from incorrect predictions across operating points, not merely on whether those scores satisfy a proper scoring rule (Chow 1970; Bartlett and Wegkamp 2008; El-Yaniv and Wiener 2010; Geifman and El-Yaniv 2017). OpenML-style benchmarks make it feasible to study such interactions systematically across heterogeneous tabular tasks (Vanschoren et al. 2013; Casalicchio et al. 2019). What remains comparatively underexplored is the model-selection question that arises once calibration is treated as

optional: after fitting a lightweight probe calibrator on a small held-out split, can inexpensive task and calibration diagnostics identify when keeping the calibrated variant is preferable to retaining the raw model? Framed this way, the paper does not claim a fully prospective rule that can decide before any calibration work is done. It studies a lightweight calibration-selection policy after a probe fit, because some informative signals about calibration utility arise only once the candidate transform has been fit on the held-out calibration data.

Within that scope, the paper studies whether a compact meta-policy can support benchmark-conditional guidance about when calibration is worth retaining. Using the 1,304 non-raw rows derived from 652 base cases across 15 OpenML tabular datasets, the analysis evaluates policy choices against oracle-style references under NLL, AURC, and a joint selection criterion. The practical point is intentionally modest. Small absolute AURC differences can still matter when accumulated across repeated model-selection decisions, but improvements in NLL do not by themselves establish better selective behavior. The contributions are therefore as follows:

1. We formulate calibration selection for selective classification as a decision problem in which the target is the net value of retaining a fitted calibrator, not calibration quality in isolation and not a zero-cost decision made before any calibration fit.
2. We provide an empirical study on 15 OpenML-style tabular tasks comprising 652 completed base evaluation cases and 1,304 non-raw model-condition rows with both NLL and AURC measurements, thereby restricting the paper's claims to the studied tabular benchmark regime and the two evaluated post-hoc calibrators.
3. We show that, among the tested meta-models, a lightweight logistic-regression policy is the strongest candidate, outperforming a decision tree under NLL, AURC, and the joint criterion in the summarized benchmark results.
4. We compare policy behavior with oracle-style references and interpret regret alongside small absolute AURC differences, yielding practical guidance about when a probe-fit calibration step appears worth retaining on small tabular validation splits.

2 Related Work

Research on probability calibration has largely studied post-hoc correction as a way to improve probabilistic outputs after a classifier has already been trained, especially when the base learner is optimized more for discrimination than for probabilistic fidelity. Early work by Zadrozny and Elkan (2001) and Zadrozny and Elkan (2002) established influential reduction-based approaches for converting classifier scores into calibrated probabilities, and subsequent analyses showed that calibration quality depends strongly on the learner, score shape, and class balance rather than behaving uniformly across settings (Niculescu-Mizil and Caruana 2005). Later methods expanded the design space through Bayesian binning and related uncertainty-aware schemes (Naeini et al. 2015), while Guo et al. (2017) highlighted how even accurate modern models can remain substantially miscalibrated. The same literature also makes clear that calibrator choice is regime dependent: beta calibration was proposed precisely to relax assumptions that can make logistic-style mappings brittle under skewed score distributions (Kull et al. 2017), and Dirichlet calibration extends this

concern to multiclass settings in which apparently favorable proper-score averages may mask uneven confidence corrections across the simplex (Kull et al. 2019). This line of work therefore motivates calibration selection in a broad sense, but it usually evaluates *which* mapping yields better probabilities, not *whether* calibration should be applied at all for a fixed trained model and a limited calibration split. For that reason, the broader post-hoc families cited here serve mainly as methodological context. The empirical claims in this paper are intentionally narrower and concern only the two calibrators actually studied, temperature scaling and Dirichlet calibration, which provide a compact comparison between a simple scalar transformation and a more flexible multiclass parametric mapping (Guo et al. 2017; Kull et al. 2019).

A parallel literature studies selective classification, abstention, reject-option learning, and confidence-threshold tuning, where the decision problem is not only whether a prediction is correct but also whether the system should predict at all. Classical foundations for reject-option classification were laid by Chow (1970) and later formalized in statistical learning terms by Bartlett and Wegkamp (2008) and El-Yaniv and Wiener (2010). Contemporary selective prediction methods operationalize this idea through confidence-based selection rules, with Geifman and El-Yaniv (2017) showing that high-confidence classifiers can be tuned to meet explicit risk targets and Geifman and El-Yaniv (2019) integrating selection directly into end-to-end training. Related analyses of deep confidence estimators and modern architectures likewise emphasize that useful uncertainty scores are task dependent and can be systematically misleading if their ranking of hard examples is poor (Hein et al. 2019; Chen et al. 2019). This literature is adjacent to, but distinct from, post-hoc calibration selection. Most threshold-tuning and selective-prediction methods take a confidence score as given and learn where to abstain, or they learn a dedicated selector jointly with the predictor. The present decision problem arises one step earlier: whether to replace the raw score with a calibrated one before any abstention threshold is chosen. That distinction matters because selective criteria such as coverage-conditioned risk or area under the risk-coverage curve depend on confidence ordering, not only on global probabilistic fidelity. A post-hoc transform can therefore improve negative log-likelihood while worsening abstention behavior if it perturbs the ranking that the downstream thresholding rule relies on.

Benchmarking work on heterogeneous tabular tasks provides the empirical backdrop for studying that interaction under realistic small-validation conditions. The OpenML ecosystem was designed to support standardized, shareable experiments across diverse datasets and learners, enabling comparisons that are both broader and more reproducible than single-dataset studies (Vanschoren et al. 2013; Casalicchio et al. 2019). That infrastructure is particularly relevant here because tabular tasks vary sharply in dimensionality, class skew, sample size, and signal-to-noise structure, all of which can alter whether a post-hoc mapping helps or harms decision quality. Prior calibration studies often report improvements on curated benchmarks, but the evidence base is thinner when the downstream criterion includes abstention quality and the calibration split is small enough that estimation noise becomes a first-order concern (Guo et al. 2017; Kull et al. 2019). In that setting, a compact benchmark with 15 datasets, 652 base evaluation cases, and 1,304 non-raw model-condition rows allows

the calibration decision itself to be studied under repeated but resource-constrained conditions. The resulting contribution is correspondingly narrow: not a comprehensive ranking of post-hoc calibrators, and not a new abstention architecture, but a benchmark-conditional analysis of when the two studied calibrators appear worth applying and when leaving raw scores untouched better preserves selective behavior.

3 Method

This study treats post-hoc calibration as a conditional intervention on a fixed benchmark of multiclass tabular prediction problems, rather than as a universally beneficial default. The recovered artifact bundle contains 15 OpenML datasets in the benchmarking tradition of Vanschoren et al. (2013) and Casalicchio et al. (2019): *letter*, *optdigits*, *pendigits*, *satimage*, *page-blocks*, *yeast*, *mfeat-factors*, *mfeat-fourier*, *mfeat-karhunen*, *waveform-5000*, *segment*, *dna*, *jungle_chess_2pcs_raw_endgame_complete*, *texture*, and *mnist_784*. These tasks were retained from a curated OpenML candidate list after applying the implemented filters of at least 1,000 and at most 100,000 rows, at least three classes, at least two predictor columns, and at most 20% missing entries. The benchmark is therefore entirely multiclass; no binary task enters the reported analysis.

For each retained dataset, a stratified 20% test split is held fixed. The remaining 80% pool is repartitioned under three resamples and five calibration fractions, $f \in \{0.01, 0.02, 0.05, 0.10, 0.20\}$, where f is the fraction of the 80% pool assigned to calibration and the complement is used as proper training data. Each dataset-resample-fraction combination is evaluated with three stored base-learner families: logistic regression, random forest, and histogram gradient boosting. This design yields 675 intended base cases ($15 \times 3 \times 5 \times 3$), of which 652 complete successfully in the preserved artifacts. Because each completed base case is paired with two candidate calibrators, the meta-level analysis operates on 1,304 non-raw candidate rows. The upstream artifact bundle preserves the base-learner identities but not a machine-readable dump of their training hyperparameters, so the present manuscript is exactly reproducible for the split protocol, calibrator fitting, and meta-policy construction, but only family-level auditable for the upstream base-model fits. All empirical claims should therefore be read as conditional on these fixed stored base predictions rather than as a fully re-instantiated end-to-end training recipe.

The downstream calibration study is correspondingly narrow. Since every retained task is multiclass, the only candidate post-hoc calibrators actually exercised are temperature scaling (Guo et al. 2017) and Dirichlet calibration (Kull et al. 2019). Temperature scaling fits a single scalar temperature $T > 0$ by minimizing calibration-split negative log-likelihood over $T \in [0.05, 20]$, implemented by golden-section search on $\log T$ with at most 48 iterations and early stopping when the bracket width falls below 10^{-4}. Dirichlet calibration uses a diagonal log-probability map

$$\tilde{p}(x) = \mathrm{softmax}(b + s \odot \log p(x)), \tag{1}$$

initialized at $s = \mathbf{1}$ and $b = \mathbf{0}$ and optimized for 40 gradient steps with learning rate 0.05 and an L_2 penalty of 10^{-4} on $s - \mathbf{1}$. Accordingly, the paper does not make

claims about post-hoc calibration in general; it studies only these two calibrators on multiclass OpenML-style tabular tasks.

3.1 Benchmark construction, selective classification, and calibration-decision targets

Selective classification is implemented by the standard confidence-threshold mechanism (Geifman and El-Yaniv 2017; Hein et al. 2019; Chen et al. 2019). For an input x, the base predictor emits a probability vector $p(x) \in \Delta^{K-1}$, the predicted label is

$$\hat{y}(x) = \arg \max_{k \in \{1,\dots,K\}} p_k(x), \tag{2}$$

and the confidence score is

$$s(x) = \max_{k \in \{1,\dots,K\}} p_k(x). \tag{3}$$

For a threshold $\tau \in [0, 1]$, the system predicts when $s(x) \geq \tau$ and abstains otherwise. This simple reject rule is deliberate: any improvement or degradation in selective behavior must arise from how calibration changes the scale or ordering of the confidence scores, rather than from a separately learned abstention model (Geifman and El-Yaniv 2019).

Each non-raw row j corresponds to a tuple (d, b, f, r, a) consisting of dataset, base learner, calibration fraction, resample, and candidate calibrator $a \in \{\text{temperature}, \text{Dirichlet}\}$. Let M_j^{raw} denote the stored uncalibrated predictor for that tuple's base case and let M_j^a denote the same predictor after fitting calibrator a on the associated calibration split. The two primary binary targets are

$$z_j^{\text{NLL}} = \mathbb{I}\big[\text{NLL}(M_j^a) < \text{NLL}(M_j^{\text{raw}})\big] \tag{4}$$

and

$$z_j^{\text{AURC}} = \mathbb{I}\big[\text{AURC}(M_j^a) < \text{AURC}(M_j^{\text{raw}})\big]\,. \tag{5}$$

The first target asks whether the candidate calibrator improves probability quality on the held-out test split; the second asks whether it improves the ranking-sensitive selective objective on that same split. Because each base case contributes one row for temperature scaling and one row for Dirichlet calibration, the learned selector solves two binary raw-versus-candidate decisions per base case rather than a multinomial choice among several calibrators.

The probability-quality metric is the usual multiclass negative log-likelihood,

$$\text{NLL} = -\frac{1}{n} \sum_{i=1}^{n} \log p_{y_i}(x_i), \tag{6}$$

where $p_{y_i}(x_i)$ is the predicted probability assigned to the true class of test example i. Lower values are preferred. NLL is a proper scoring rule and is highly sensitive

to overconfident mistakes, but it does not by itself guarantee preservation of the confidence ordering that selective classification uses (Niculescu-Mizil and Caruana 2005; Guo et al. 2017).

To define the joint target, let $\mathcal{A}_c$ denote the set of available actions for base case c, consisting of the raw predictor and its available candidate calibrators. For each metric $m \in \{\mathrm{NLL}, \mathrm{AURC}\}$, define the oracle-relative regret of action $a \in \mathcal{A}_c$ by

$$r_c^m(a) = U_c^m(a) - \min_{a' \in \mathcal{A}_c} U_c^m(a'), \tag{7}$$

with lower utility values preferred. Since NLL and AURC live on different numerical scales, the implementation normalizes each metric by the largest observed within-case regret,

$$s_c^m = \max\left(\{r_c^m(a') : a' \in \mathcal{A}_c\} \cup \{10^{-12}\}\right), \tag{8}$$

and then forms the joint regret

$$J_c(a) = \frac{r_c^{\mathrm{NLL}}(a)}{s_c^{\mathrm{NLL}}} + \frac{r_c^{\mathrm{AURC}}(a)}{s_c^{\mathrm{AURC}}}. \tag{9}$$

For the candidate row j associated with action a_j, the joint label is

$$z_j^{\mathrm{joint}} = \mathbb{I}\left[J_{c(j)}(a_j) < J_{c(j)}(\mathrm{raw})\right]. \tag{10}$$

This definition makes the joint selector conservative in a useful sense: calibration is labeled positive only when the candidate action improves the normalized sum of oracle-relative NLL and AURC regrets for its own base case.

3.2 Meta-features, probe-fit scope, and leave-one-dataset-out training

The reported selector should be interpreted as a lightweight calibration-selection policy after a probe fit, not as a universal skip-before-fit rule. The reason is implementation-level and explicit: the meta-feature vector contains both signals that are available before fitting the candidate calibrator and diagnostics that can only be computed after fitting and applying that calibrator on the calibration split. At the same time, the construction remains leakage-safe with respect to the test split: dataset identifiers, oracle annotations, and all test-set outcomes are excluded from the feature vector and enter only through the supervision labels defined above.

Formally, let $\phi_j \in \mathbb{R}^d$ denote the feature vector for non-raw row j. The training code excludes `dataset_slug`, `task_id`, `data_id`, `dataset_name`, `fit_params_json`, `class_counts_json`, `oracle_best_nll`, `oracle_best_aurc`, all test-set metrics, and all target labels. The retained feature set contains 46 numeric variables and 3 categorical variables. The categorical variables are `model`, `calibrator`, and `task_type`. The numeric variables available before fitting the candidate calibrator are `fraction`, `resample`, `calibration_prob_rows_normalized`, `calibration_rows`, `calibration_size`, `class_count`, `classes_observed_in_calibration`,

imbalance_ratio, max_class_count, mean_class_count, min_class_count, missing_class_fraction_in_calibration, proper_train_size, raw_cal_aurc, raw_cal_nll, samples_per_class_max, samples_per_class_mean, samples_per_class_min, task_missing_fraction, task_n_categorical_features, task_n_classes, task_n_features_used, task_n_numeric_features, task_n_rows, and test_size. The additional probe-fit diagnostics, which require fitting and applying the candidate calibrator on the calibration split, are calibrated_cal_aurc, calibrated_cal_nll, confidence_pearson_raw_vs_cal, confidence_spearman_raw_vs_cal, delta_aurc_calibration, delta_nll_cal_boot_ci_lower, delta_nll_cal_boot_ci_upper, delta_nll_cal_boot_mean, delta_nll_cal_boot_p_harm, delta_nll_cal_boot_p_improve, delta_nll_cal_boot_std, delta_nll_cal_observed, delta_nll_calibration, max_abs_conf_shift, mean_abs_conf_shift, mean_rank_distortion, median_abs_conf_shift, median_rank_distortion, top1_prediction_change_rate, and top10_confidence_jaccard. The bootstrap summaries are computed with 200 resamples and seed 42. Consequently, the selector uses only calibration-set information, but it does not use only pre-calibration information.

The meta-policy is trained on the 1,304 non-raw rows only. For objective $m \in \{\text{NLL}, \text{AURC}, \text{joint}\}$, it first learns a candidate-transition gate

$$\pi_m(\phi_j) = \mathbb{I}\big[\Pr(z_j^m = 1 \mid \phi_j) \geq \tfrac{1}{2}\big], \tag{11}$$

where $\pi_m(\phi_j) = 1$ means "apply the candidate calibrator represented by row j" and $\pi_m(\phi_j) = 0$ means "keep the raw predictor." Numeric features are passed as sparse dictionary features; if a numeric value is missing, the implementation inserts a companion missingness indicator and omits the base value. Categorical features are one-hot encoded by dictionary vectorization, with unseen or empty values mapped to a dedicated missing token.

Because deployment requires one action per base case rather than two independent binary decisions, the revised analysis also collapses the row-level gate into a case-level action selector. For each completed base case c, the action set is

$$\mathcal{A}_c = \{\text{raw}, \text{temperature}, \text{Dirichlet}\}.$$

For mode m, the selector keeps raw if no candidate has $\pi_m(\phi_j) = 1$; if one or both candidates are accepted, it chooses the accepted calibrator with the larger out-of-fold probability $\Pr(z_j^m = 1 \mid \phi_j)$. Ties are broken deterministically by the stored calibrator order. This conversion enforces the deployed action space and makes the case-level metrics directly comparable with always-raw, always-temperature, always-Dirichlet, and oracle case-level baselines.

Two meta-model families are evaluated. The primary model is logistic regression with dictionary vectorization, MaxAbs scaling, liblinear optimization, class_weight=balanced, max_iter=2000, and decision threshold 0.5. The nonlinear baseline is a decision tree with dictionary vectorization, max_depth=3,

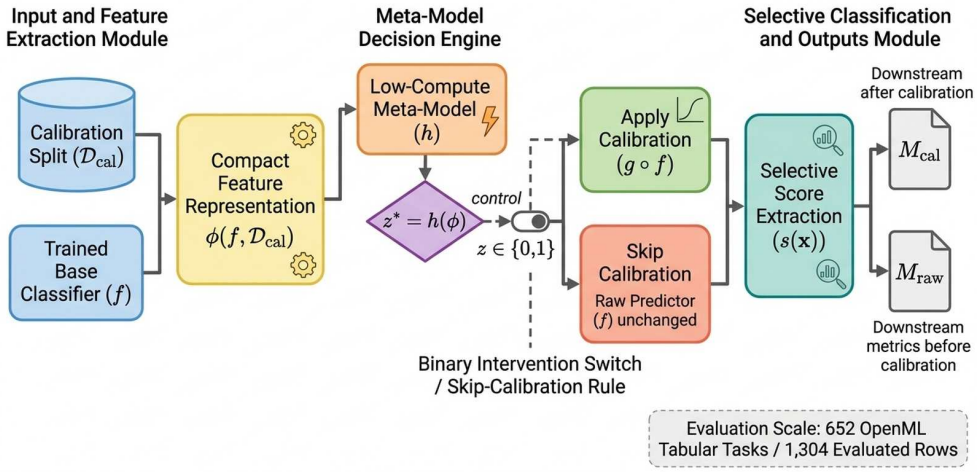


Fig. 1 Method pipeline for conditional post-hoc calibration on multiclass OpenML-style tabular tasks. A stored base probabilistic classifier is evaluated first; each candidate calibrator is then probe-fit on a small calibration split; calibration-set diagnostics and paired raw-versus-calibrated outcomes are computed; and a leave-one-dataset-out meta-policy predicts whether that specific candidate calibrator should be applied or whether the raw model should be retained. The learned selector is therefore a calibration-selection policy after a probe fit, trained on 652 completed base cases and 1,304 non-raw candidate rows.

`min_samples_leaf=10`, `class_weight=balanced`, and the same 0.5 threshold on the positive class. Evaluation uses leave-one-dataset-out (LODO) splitting: each fold withholds one entire dataset, trains on the remaining 14 datasets, and produces out-of-fold probabilities for every row of the held-out dataset. If a training fold contains only one class for the objective under study, the implementation falls back to a constant prediction on that fold. Candidate meta-models are ranked separately for NLL, AURC, and joint labels by their mean out-of-fold objective values, and logistic regression is selected in all three modes in the preserved results. The overall pipeline is summarized in Figure 1.

3.3 Selective metrics, operating-scope limits, and oracle-relative regret

The selective-classification analysis follows the standard risk-coverage formulation (Chow 1970; Bartlett and Wegkamp 2008; El-Yaniv and Wiener 2010; Geifman and El-Yaniv 2017). For a threshold τ, coverage is

$$C(\tau) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[s(x_i) \geq \tau] \,, \tag{12}$$

and selective risk on the retained set is

$$R(\tau) = \frac{\sum_{i=1}^{n} \mathbb{I}[s(x_i) \geq \tau]\, \mathbb{I}[\hat{y}(x_i) \neq y_i]}{\sum_{i=1}^{n} \mathbb{I}[s(x_i) \geq \tau]}. \tag{13}$$

The implementation summarizes this tradeoff by the empirical area under the risk-coverage curve, computed by sorting test examples in decreasing confidence order and averaging the cumulative error rate:

$$\mathrm{AURC} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{i} \sum_{\ell=1}^{i} \mathbb{I}\left[\hat{y}_{(\ell)} \neq y_{(\ell)}\right]. \tag{14}$$

Lower AURC is preferred. This metric is especially relevant here because a calibrator may improve NLL while perturbing the confidence ordering that selective classification depends upon (Geifman and El-Yaniv 2017; Chen et al. 2019). The absolute AURC deltas in this benchmark are often small, so later sections interpret them in raw units and relative-to-oracle context rather than treating every signed change as practically large.

The preserved artifact bundle does not expose `risk_at_cov_*`, `coverage_*`, or threshold-specific operating summaries. The selective evaluation is therefore limited to full-curve AURC and oracle-relative regret rather than fixed-coverage risk comparisons. This restriction should be kept in mind when interpreting the selective-classification evidence: the reported method speaks to whole-curve ranking quality, not to a particular operating point.

To quantify the cost of an incorrect calibration decision, the analysis also reports regret relative to the best realized action for each row. If U_j^{π} denotes the utility obtained when policy π chooses either the raw model or the candidate calibrator for row j, and

$$U_j^{\star} = \min\{U_j^{\mathrm{raw}}, U_j^{a}\} \tag{15}$$

is the row-wise oracle utility for the criterion under study, then

$$\mathrm{Regret}(\pi) = \frac{1}{m} \sum_{j=1}^{m} \left(U_j^{\pi} - U_j^{\star}\right). \tag{16}$$

This regret is reported separately for NLL and AURC, and the joint objective uses the normalized within-case sum defined above. Regret is more informative than binary meta-label accuracy alone because a false positive on a nearly neutral row is operationally different from missing a large calibration gain.

Table 1 summarizes the core objects of the method. The central methodological claim is correspondingly narrow. The paper does not establish a universal low-compute rule that can decide on calibration before any candidate calibrator is fit. It evaluates a leakage-safe, probe-fit calibration-selection policy that uses calibration-set diagnostics to decide whether one of two studied multiclass calibrators is likely to outperform the raw predictor on either NLL, AURC, or their normalized joint regret.

Table 1 Summary of the principal objects in the conditional calibration pipeline.

Object	Description
M^{raw}	Stored base probabilistic classifier before post-hoc calibration
M^a	Candidate calibrated counterpart for calibrator a on the same base case
$s(x)$	Confidence score given by the maximum predicted class probability
$z^{\text{NLL}}, z^{\text{AURC}}$	Binary targets indicating whether the candidate calibrator beats raw
$J_c(a)$	Normalized within-case joint regret used for the joint target
$\pi_m(\phi)$	LODO meta-policy for objective $m \in \{\text{NLL}, \text{AURC}, \text{joint}\}$
NLL	Probability-quality metric based on multiclass log-loss
AURC	Selective-classification metric integrating cumulative risk over coverage
Regret(π)	Average utility gap between the policy and row-wise oracle selection

4 Experiments

4.1 Experimental Setup

This empirical study is deliberately benchmark-conditional. It evaluates a lightweight calibration-selection policy on 15 OpenML-style multiclass tabular tasks: 6-letter, 28-optdigits, 32-pendigits, 182-satimage, 30-page-blocks, 181-yeast, 12-mfeat-factors, 14-mfeat-fourier, 16-mfeat-karhunen, 60-waveform-5000, 36-segment, 40670-dna, 41027-jungle-chess-2pcs-raw-endgame-complete, 40499-texture, and 554-mnist-784. These tasks were curated from a larger OpenML candidate list using simple manifest-level filters: 1,000–100,000 rows, at least three classes, at least two predictive features, and overall missing-value fraction at most 0.2. The resulting benchmark spans 1,484–70,000 rows and 3–26 classes. This is a narrow but reproducible tabular regime in the sense emphasized by Vanschoren et al. (2013) and later benchmark-oriented OpenML work (Casalicchio et al. 2019), and the claims in this section are confined to that regime and to the two studied post-hoc calibrators.

The split protocol has two stages. Each dataset is first stratified with seed 42 into a fixed proper-train/test decomposition equivalent to the released 60/20/20 train/calibration/test manifest split. The downstream calibration sweep then re-partitions the non-test pool for each resample: for calibration fractions $f \in \{0.01, 0.02, 0.05, 0.10, 0.20\}$ and resamples $r \in \{0, 1, 2\}$, a calibration subset of size f times the proper-train pool is reserved and the remainder is used to fit the base learner, while the 20% test split is held fixed. Thus the calibration size varies from 16 to 320 examples on 12-mfeat-factors, from 25 to 510 on 40670-dna, and from 160 to 3200 on 6-letter. Combining 15 datasets, 3 base-model families, 5 calibration fractions, and 3 resamples yields 675 planned base cases.

The base-model artifacts cover three off-the-shelf learner families, identified in the stored prediction directories as logistic regression, random forest, and histogram-based gradient boosting. The released benchmark artifacts preserve the model-family labels but not a fuller hyperparameter ledger for these raw predictors, so the empirical claims are tied to these fixed artifactized base models rather than to a separately retuned learner suite. Each completed base case is evaluated in raw form and after two post-hoc calibrators. Temperature scaling is implemented as a scalar temperature on log probabilities, with T chosen by golden-section search over $T \in [0.05, 20]$ for up to 48

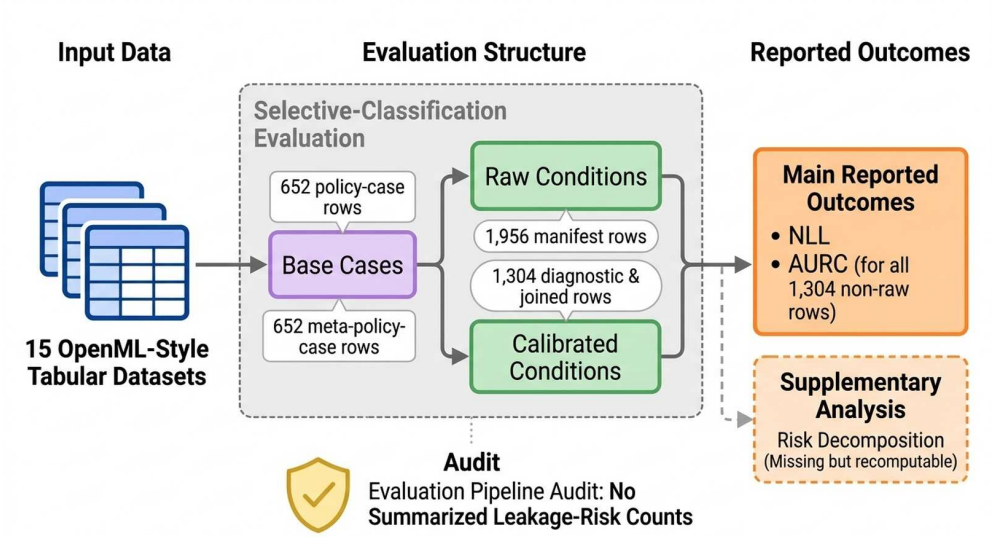


Fig. 2 Experimental pipeline for conditional post-hoc calibration on the OpenML tabular benchmark. Each task contributes a fixed held-out test split, a resampled proper-train pool, raw predictions from a stored base learner, and up to two candidate calibration transforms before final evaluation under NLL and AURC.

iterations. Dirichlet calibration is implemented as a classwise affine transform of log probabilities followed by softmax, optimized for 40 gradient steps with learning rate 0.05 and L_2 shrinkage 10^{-4} toward unit scales. Because every dataset in the benchmark has more than two classes, both calibrators are available throughout. Figure 2 summarizes the resulting pipeline.

Of the 675 planned base cases, 652 complete successfully, yielding 1,304 non-raw candidate-transition rows and 652 raw companions. The 23 missing case pairs are not treated as innocuous missing-at-random noise: the audit shows that they are concentrated in later-fraction 554-mnist-784 runs, so all aggregate claims below are conditional on the 652 completed base cases rather than on the full nominal design. Within that completed set, both NLL and AURC are available for all 1,304 non-raw rows. By contrast, the repaired artifacts do not expose `risk_at_cov_*`, `coverage_*`, or threshold summaries, so fixed-coverage selective-risk analysis cannot be reported from this benchmark record. The selective-classification evidence is therefore limited to AURC and oracle-regret summaries rather than operating-point-specific risk curves. Because the checkout does not preserve reruns or imputed outcomes for those 23 missing case pairs, the paper intentionally reports complete-case estimates rather than a pseudo-complete benchmark average. A stronger missing-case sensitivity audit would bracket the aggregate conclusions under best-case and worst-case assumptions for the absent NLL/AURC outcomes, but that calculation is not available in the preserved artifact bundle and is therefore not claimed here.

At the decision level the benchmark compares two families of selectors. The first is a transparent rule-based baseline evaluated leave-one-dataset-out: for each held-out

task, the rule is tuned on the remaining tasks over three calibration-split diagnostics for a specific calibrator, namely minimum samples per observed class, the upper endpoint of a 200-round bootstrap interval for calibration-split ΔNLL, and the Spearman correlation between raw and calibrated confidence scores. Candidate thresholds are drawn from empirical quantiles and endpoints, and the selected rule minimizes training-set mean ΔNLL, then mean NLL regret to oracle, then NLL harm rate, and finally mean ΔAURC. The second family is a learned meta-policy, also evaluated leave-one-dataset-out, on the 1,304 non-raw rows. Each row corresponds to a particular raw-to-calibrator transition for one $(\text{dataset}, \text{base model}, f, r)$ case, so the learned gate first predicts whether that specific transition should be executed. A separate case-level deployment reduction then chooses exactly one action per base case: raw if neither candidate is accepted, otherwise the accepted calibrator with the larger out-of-fold probability for the mode under evaluation. This reduction is reported separately because it is the action space that a deployed calibration-selection policy would actually use.

The learned meta-policy uses 45 numeric and 3 categorical calibration-split descriptors. The numeric inventory comprises fraction, resample, calibration size, calibration rows, proper-train size, test size, task row count, task class count, task feature counts, task missing fraction, class count, classes observed in calibration, minimum/mean/maximum class counts, minimum/mean/maximum samples per observed class, imbalance ratio, missing-class fraction in calibration, raw and calibrated calibration-split NLL and AURC, calibration-split ΔNLL and ΔAURC, the observed, bootstrap-mean, bootstrap-standard-deviation, bootstrap-95% interval, and bootstrap harm/improvement probabilities for calibration-split ΔNLL, raw-versus-calibrated confidence Pearson and Spearman correlations, mean/median/maximum absolute confidence shift, mean/median rank distortion, top-1 prediction change rate, top-10 confidence Jaccard overlap, and the count of renormalized calibration-probability rows. The categorical fields are calibrator identity, base-model family, and task type. Numeric missingness is represented by feature-specific missing indicators, categorical missingness by a dedicated missing level, and the categorical variables are one-hot encoded.

Among candidate meta-models, logistic regression and a shallow decision tree are tested. The logistic gate uses a one-hot vectorizer, max-abs scaling, balanced class weights, the `liblinear` solver, a default L_2 penalty with $C = 1$, and `max_iter` $= 2000$. The tree uses balanced class weights, `max_depth` $= 3$, and `min_samples_leaf` $= 10$. Three decision targets are considered. The NLL and AURC targets are binary labels indicating whether calibration yields $\Delta\text{NLL} < 0$ or $\Delta\text{AURC} < 0$ on the test set. The joint target is defined through casewise normalized regret. For a case c and action $a \in \{\text{raw}, \text{temperature}, \text{Dirichlet}\}$, let

$$r_{\text{NLL}}(a; c) = \text{NLL}(a; c) - \min_{a'} \text{NLL}(a'; c), \qquad r_{\text{AURC}}(a; c) = \text{AURC}(a; c) - \min_{a'} \text{AURC}(a'; c).$$

These regrets are scaled within case by $s_{\mathrm{NLL}}(c) = \max_{a'} r_{\mathrm{NLL}}(a'; c) \vee 10^{-12}$ and $s_{\mathrm{AURC}}(c) = \max_{a'} r_{\mathrm{AURC}}(a'; c) \vee 10^{-12}$, and the joint regret is

$$J(a; c) = \frac{r_{\mathrm{NLL}}(a; c)}{s_{\mathrm{NLL}}(c)} + \frac{r_{\mathrm{AURC}}(a; c)}{s_{\mathrm{AURC}}(c)}.$$

The joint label is one when the calibrated action has smaller $J(a; c)$ than raw, and mean joint regret is the corresponding average selected value over evaluation rows. This construction keeps the two regret components comparable even though the absolute scales of NLL and AURC differ sharply across datasets and model families.

Because several of these features, especially rank distortion and bootstrap ΔNLL diagnostics, are computed only after fitting and applying a candidate calibrator on the calibration split, the selector should be interpreted as a lightweight calibration-selection policy after probe fitting, not as a zero-cost pre-calibration screen. The computational saving comes from avoiding downstream use of unhelpful calibrators and from using a simple selector, not from eliminating the need to inspect the candidate transform altogether.

4.2 Main Results

Within this benchmark and for these two calibrators, unconditional calibration improves probability fit on average but does not improve abstention quality on average. Across the 652 completed base cases, Dirichlet calibration achieves mean ΔNLL of -0.0337 and temperature scaling achieves mean ΔNLL of -0.0450, both relative to the corresponding raw model output, yet their mean ΔAURC values are $+8.83 \times 10^{-4}$ and $+3.96 \times 10^{-4}$, respectively. The same tension appears in harm frequencies. Dirichlet calibration worsens NLL in 21.3% of cases and worsens AURC in 46.0% of cases, while temperature scaling worsens NLL in 29.1% of cases and worsens AURC in 52.3% of cases. These numbers support a narrower claim than a generic anti-calibration story: on these OpenML tabular tasks, and for these two calibrators, improvements in log-loss do not reliably transfer to the ranking-sensitive selective-classification objective (El-Yaniv and Wiener 2010; Geifman and El-Yaniv 2017; Chen et al. 2019).

Table 2 also clarifies the baseline landscape. The rule-based skip baseline reduces NLL harm relative to unconditional calibration, but it gives back much of the average NLL benefit and still worsens AURC on average. Always raw trivially avoids harm relative to itself, but it incurs the largest regret to the NLL oracle among the case-level deployment baselines. The learned gates therefore matter only if they can sit between those two extremes: closer to unconditional calibration on average NLL, but closer to always raw on failure control.

On the candidate-transition frame, the learned leave-one-dataset-out policy is predictable enough to recover most of calibration's NLL benefit while sharply reducing its failure rate. The NLL-targeted logistic gate selects calibration in 61.4% of non-raw rows and reaches mean selected ΔNLL of -0.0398, nearly matching the -0.0393 mean attained by always calibrating on the same 1,304-row frame, but with NLL harm reduced from 25.2% under always calibrate to 6.5%. When the objective is switched to AURC, the same logistic family selects calibration in 40.3% of rows and attains

Table 2 Main experimental summaries. Negative ΔNLL and ΔAURC indicate improvement relative to raw probabilities on the same evaluation unit. The first block is case-level (652 completed base cases); the second is candidate-transition level (1304 non-raw rows), where each base case contributes one row per calibrator.

Method	Rows	Mean ΔNLL	NLL harm (%)	Mean ΔAURC	AURC harm (%)
Case-level baselines					
Always raw	652	0	0.0	0	0.0
Rule-based skip	652	−0.0136	19.6	$+4.76 \times 10^{-4}$	34.5
Always temperature scaling	652	−0.0450	29.1	$+3.96 \times 10^{-4}$	52.3
Always Dirichlet calibration	652	−0.0337	21.3	$+8.83 \times 10^{-4}$	46.0
Oracle selector	652	−0.0836	0.0	-7.38×10^{-4}	0.0
Learned leave-one-dataset-out gates					
Logistic gate, NLL objective	1304	−0.0398	6.5	$+1.33 \times 10^{-4}$	26.1
Decision tree gate, NLL objective	1304	−0.0265	7.0	$+1.87 \times 10^{-5}$	25.9
Logistic gate, AURC objective	1304	−0.0148	4.9	-9.25×10^{-5}	13.0
Decision tree gate, AURC objective	1304	−0.0129	8.0	-1.72×10^{-5}	13.1

mean selected ΔAURC of -9.25×10^{-5}, while AURC harm falls to 13.0% from 49.2% under unconditional calibration. Because the absolute AURC deltas are small, they are more informative when read together with oracle regret than in isolation: under the AURC-targeted mode, mean AURC regret to oracle falls from 1.38×10^{-3} for always calibrate to 6.45×10^{-4}, and under the joint criterion mean joint regret falls from 0.9431 to 0.9047. These are descriptive benchmark gains, not evidence for a universal abstention improvement guarantee.

The action-matched case-level reduction preserves the same qualitative picture while removing the ambiguity of two independent binary gates. Table 3 reports one-action-per-case summaries over the 652 completed base cases. The joint case-level selector chooses a calibrator in 55.8% of cases, with 46.6% assigned to temperature scaling and 9.2% assigned to Dirichlet calibration. It attains mean ΔNLL −0.0404, slightly better than always-Dirichlet and close to always-temperature, while reducing NLL harm to 6.1% compared with 29.1% for always-temperature and 21.3% for always-Dirichlet. It also attains a small negative mean ΔAURC of -1.20×10^{-4}. The AURC-focused case-level selector is safer on AURC harm, at 14.6%, and reaches mean ΔAURC -1.60×10^{-4}, but gives back more NLL improvement. These case-level results validate the deployed action space more directly than the row-level gate: the learned approach is not merely making independent accept/reject decisions for each calibrator, but can be converted into a single action over $\{\text{raw}, \text{temperature}, \text{Dirichlet}\}$ with bounded regret to the case-level oracle.

Among the two tested meta-model families, logistic regression is the stronger and more stable candidate. In the NLL mode it outperforms the decision tree on the primary objective, achieving mean selected ΔNLL of −0.0398 against −0.0265, lower mean joint regret (0.8683 against 0.9096), and higher out-of-fold AUC (0.8118 against 0.7587). In the AURC mode it again produces the better primary objective value, with mean selected ΔAURC of -9.25×10^{-5} compared with -1.72×10^{-5}, and it also yields lower AURC regret to oracle, 6.45×10^{-4} versus 7.20×10^{-4}. The joint criterion points in the same direction: logistic regression reaches mean joint regret

Table 3 Case-level action-matched policy summaries over the 652 completed base cases. Each policy chooses exactly one action from raw, temperature scaling, and Dirichlet calibration. Negative deltas indicate improvement relative to raw.

Policy	Non-raw (%)	Mean ΔNLL	NLL harm (%)	Mean ΔAURC	AURC harm (%)
Always raw	0.0	0	0.0	0	0.0
Always temperature	100.0	−0.0450	29.1	$+3.96 \times 10^{-4}$	52.3
Always Dirichlet	100.0	−0.0337	21.3	$+8.83 \times 10^{-4}$	46.0
Case-level NLL selector	68.3	−0.0271	7.8	$+5.62 \times 10^{-4}$	31.4
Case-level AURC selector	46.9	−0.0235	7.8	-1.60×10^{-4}	14.6
Case-level joint selector	55.8	−0.0404	6.1	-1.20×10^{-4}	22.7

0.9047, whereas the decision tree reaches 0.9303. The ranking is therefore not that linear models win every secondary statistic, but that the simpler regularized gate is the better primary selector in all three optimization modes tested here.

The robustness evidence is stronger for NLL than for AURC. A separate case-level leave-one-dataset-out audit of the logistic selector improves mean ΔNLL on 14 of the 15 held-out datasets, but mean ΔAURC is negative on only 8 of 15, with median held-out mean ΔAURC -1.81×10^{-5}. The available uncertainty summaries point in the same direction without justifying stronger inferential language: bootstrap summaries by calibration fraction show persistent negative mean ΔNLL for Dirichlet calibration and more fragile temperature-scaling gains at the smallest fractions, whereas the AURC intervals remain small in raw units and sometimes overlap zero. The released artifacts therefore support the descriptive conclusion that conditional calibration is useful for managing failure rates and oracle regret, but they do not support a stronger claim that AURC improvements are large or uniformly stable across operating regimes. Figure 3 shows that the underlying NLL/AURC trade-off is distributed across tasks rather than driven by a single outlier.

4.3 Mechanism Analysis

The mechanism analysis is descriptive rather than causal. It addresses why NLL can improve while AURC worsens by tracking how post-hoc calibration perturbs the confidence ordering that generates the risk–coverage curve (Chow 1970; El-Yaniv and Wiener 2010; Geifman and El-Yaniv 2017). Selective classification depends on that ordering because abstention is implemented by thresholding confidence scores. Post-hoc calibration, by contrast, is optimized for probability fidelity and can reshape score geometry in ways that are benign for NLL but harmful for ranking-sensitive abstention quality (Niculescu-Mizil and Caruana 2005; Guo et al. 2017; Kull et al. 2019). The benchmark diagnostics make that trade-off visible on the calibration split before test outcomes are consulted.

For Dirichlet calibration, the Spearman association between mean rank distortion and ΔAURC is 0.218, while the corresponding association for temperature scaling is 0.212; larger rank distortion is therefore systematically associated with worse selective performance for both calibrators. Calibration size and mean samples per class move in the opposite direction. For Dirichlet calibration, their Spearman associations

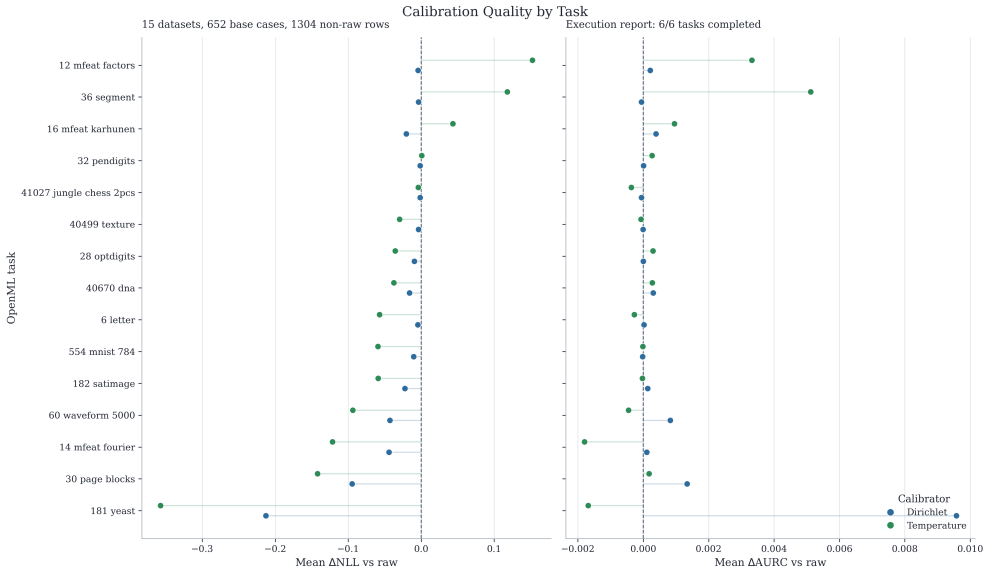


Fig. 3 Task-level changes in NLL and AURC induced by calibration and skip policies across the 15 OpenML benchmarks. The selective-classification evidence is necessarily AURC-based here because fixed-coverage risk summaries are not available in the supplied artifacts.

with ΔAURC are −0.238 and −0.229, and for temperature scaling they are −0.150 and −0.135. The descriptive pattern is therefore consistent across the two calibrators studied here: larger calibration support tends to reduce abstention harm, whereas transformations that more strongly disturb confidence ranking tend to increase it. This does not establish a causal mechanism, but it does show that the NLL/AURC disagreement is structured rather than idiosyncratic.

The learned gate internalizes the same structure. In the NLL-targeted logistic model, some of the largest-magnitude positive coefficients are attached to histogram-based gradient boosting as the base-model family (+2.819), calibration fraction (+1.528), top-1 prediction change rate (+1.062), and calibration size (+0.841); some of the largest negative coefficients are attached to bootstrap uncertainty in calibration-split ΔNLL (−2.092), bootstrap harm probability (−1.592), mean rank distortion (−1.573), and calibration-split calibrated NLL (−1.438). The signs are informative. The selector becomes more willing to retain calibration when support is larger, and more reluctant when the estimated gain is unstable or when the raw-to-calibrated transformation appears rank-distorting on the calibration split. Because these are probe-fit diagnostics, the operational interpretation is again conditional selection after a cheap calibration probe rather than a pure pre-screening test.

These diagnostics sharpen the interpretation of the main results. The benchmark does not imply that post-hoc calibration is generally harmful, nor does it imply that probability quality and abstention quality are unrelated. Instead, it shows that their relationship is regime-dependent on these tabular tasks: smaller calibration subsets, thinner per-class support, and larger calibration-induced ranking changes are precisely the conditions under which the universal-calibrate heuristic becomes unreliable.

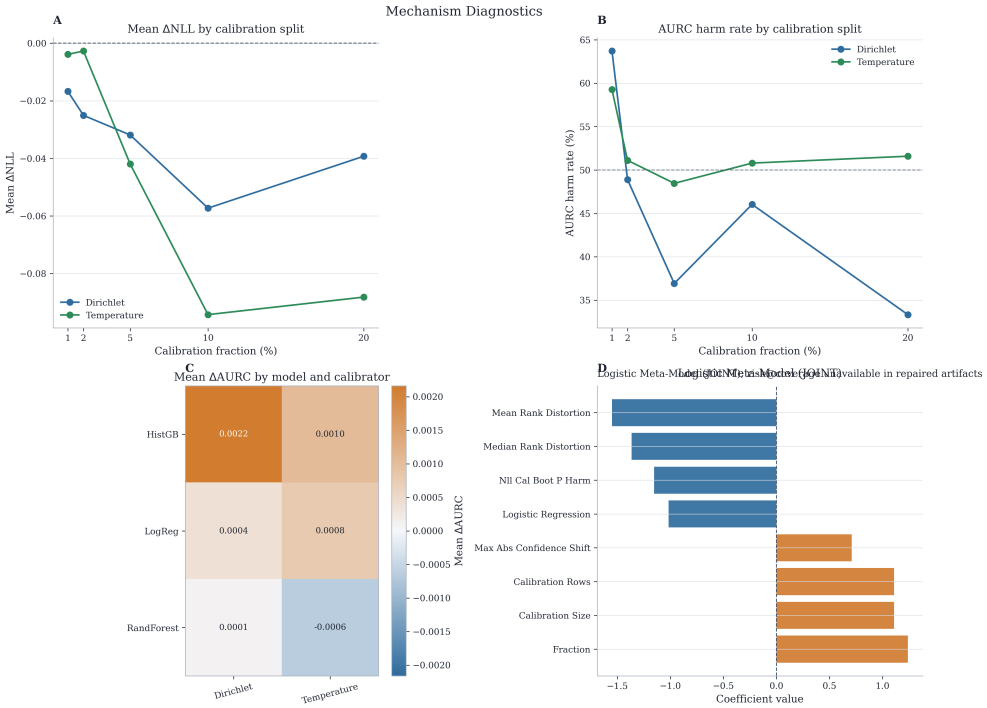


Fig. 4 Descriptive mechanism diagnostics for calibration-induced abstention harm. Larger rank distortion is associated with worse AURC outcomes, whereas larger calibration support and more samples per class are associated with safer calibration decisions.

That conclusion is consistent with the broader calibration literature, which has long emphasized that different post-hoc methods behave differently across sample regimes and model families (Zadrozny and Elkan 2002; Naeini et al. 2015; Guo et al. 2017; Kull et al. 2017, 2019), but the present results extend that point to selective classification, where rank preservation is itself part of the objective. On this benchmark, the practical lesson is therefore narrow but defensible: calibration is best treated as a conditional intervention selected by a lightweight probe-fit rule, not as a mandatory final step in the prediction pipeline.

5 Discussion

Within the benchmark actually studied here, post-hoc calibration should not be treated as an unconditional default for selective classification. Across 15 OpenML-style multiclass tabular tasks, 652 completed base cases, and 1,304 non-raw candidate-transition rows, the two studied calibrators improve mean NLL relative to raw predictions while failing to improve mean AURC on average. That pattern matters because selective classification depends not only on marginal probability accuracy but also on how confidence scores rank uncertain examples for abstention, a relationship further characterized by the diagnostics in Figure 4. The strongest supported claim is

therefore narrow: under these aggregated OpenML-style tabular conditions, and for temperature scaling and Dirichlet calibration, better probabilistic fit does not reliably imply better abstention quality.

The policy result should be interpreted with the same caution. Because the learned selector uses probe-fit diagnostics such as calibration-split ΔNLL summaries, confidence-shift statistics, and rank-distortion features, it is better described as a lightweight calibration-selection policy after probe fitting than as a zero-cost pre-calibration skip rule. Its practical appeal is that a simple regularized logistic-regression gate outperforms the tested shallow decision-tree alternative and can reduce harm relative to unconditional calibration, while remaining much simpler than a large policy learner or an expanded calibrator search. The revised case-level reduction also shows that the row-level gate can be made deployment-shaped: the joint selector chooses exactly one action per base case, obtains mean ΔNLL -0.0404, avoids the positive mean AURC harm of both unconditional calibrators, and reduces NLL harm to 6.1%. But the evidence does not support a broader deployment claim that calibration can be screened out before any candidate transform is fit, and any compute saving is correspondingly limited: it comes from using a cheap selector and from avoiding downstream reliance on unhelpful probe-fitted calibrators, not from eliminating the probe step altogether. This more modest framing is also the one most consistent with the wider literature showing that confidence corrections are objective dependent and that procedures that help one target need not help another (Niculescu-Mizil and Caruana 2005; Guo et al. 2017; Kull et al. 2019).

Several limitations materially constrain the scope of the conclusion. First, the empirical basis is a compact tabular benchmark rather than a broad claim about post-hoc calibration in general: the study covers 15 OpenML-style tasks, three stored base-learner families, and only two post-hoc calibrators, so the results should not be extrapolated to other domains, model classes, or calibration families without further evidence (Vanschoren et al. 2013; Casalicchio et al. 2019). Second, the benchmark is incomplete. Of 675 planned base cases, 23 case pairs are missing, and the audit indicates that these failures are concentrated in later-fraction runs on 554-mnist-784 rather than dispersed uniformly. The reported summaries therefore describe the 652 completed cases, not the full nominal design, and they remain vulnerable to bias if the missing cases are systematically harder or more calibration sensitive. Third, the selective-classification evidence is necessarily incomplete: the artifact record supports NLL, AURC, and oracle-relative regret, but it does not expose fixed-coverage risk, threshold, or coverage summaries. As a result, the paper can speak to ranking-sensitive selective performance through AURC, but not to application-specific operating points where abstention budgets are externally set (Chow 1970; El-Yaniv and Wiener 2010; Geifman and El-Yaniv 2019). Fourth, the absolute AURC differences are small, and the leave-one-dataset-out pattern is less stable for AURC than for NLL, so these effects are best read descriptively and alongside regret rather than as large standalone practical gains. Finally, the mechanism analysis is post-hoc and associational. The observed links between larger rank distortion, smaller calibration support, and worse AURC are informative, but they do not establish a causal account of why a given calibration map harms abstention on a given task.

Even with those caveats, the results connect cleanly to a broader distinction in the selective-prediction literature. Classic reject-option work treated abstention as a decision problem tied to uncertainty ordering (Chow 1970; Bartlett and Wegkamp 2008; El-Yaniv and Wiener 2010), whereas the calibration literature largely evaluates whether predicted probabilities agree with empirical frequencies (Zadrozny and Elkan 2001, 2002; Naeini et al. 2015; Guo et al. 2017; Kull et al. 2017). The present benchmark reinforces that these are related but non-identical objectives: a post-hoc transform can improve log-loss while leaving the risk–coverage trade-off unchanged or worse (Niculescu-Mizil and Caruana 2005; Kull et al. 2019). For modern selective-classification pipelines, the practical implication is therefore not that calibration is broadly undesirable, but that calibration should be evaluated jointly with the abstention objective it is meant to support (Geifman and El-Yaniv 2017; Chen et al. 2019; Geifman and El-Yaniv 2019; Hein et al. 2019). On the evidence available here, that joint evaluation is most defensibly expressed as conditional selection among probe-fitted options rather than as a universal post-processing default.

Several directions for future work follow directly from these limits. The first is broader external validation on additional OpenML collections and on non-tabular settings where representation learning and higher-dimensional score geometry may change the relationship between NLL gains and AURC gains (Vanschoren et al. 2013; Casalicchio et al. 2019; Guo et al. 2017). The second is empirical scope: either claims should remain explicitly confined to temperature scaling and Dirichlet calibration, or the benchmark should be extended to additional standard post-hoc families such as isotonic regression, vector scaling, logistic scaling, beta calibration, and Bayesian-binning-style methods (Zadrozny and Elkan 2002; Naeini et al. 2015; Kull et al. 2017). The third is evaluation: operating-point-specific selective-risk summaries, uncertainty intervals for policy comparisons, and explicit sensitivity analyses for the concentrated missing-case pattern would make the practical meaning of small AURC differences easier to assess. More generally, sharper theory for when post-hoc transforms preserve or disrupt the confidence ordering used by abstention could help explain why gains in NLL coexist with losses in selective quality. If that theory is paired with broader and more complete benchmark evidence, then lightweight probe-fit calibration-selection policies may become a better-grounded component of selective prediction under limited validation data.

6 Conclusion

This paper examined post-hoc calibration as a conditional intervention for selective classification on a deliberately narrow OpenML-style multiclass tabular benchmark (Vanschoren et al. 2013; Casalicchio et al. 2019). Across 15 datasets, 652 completed base evaluation cases, and 1,304 calibrated candidate-transition rows with both NLL and AURC available, the evidence showed that calibration improved probabilistic accuracy on average but did not reliably improve abstention behavior, reinforcing the distinction between calibration objectives and reject-option objectives already emphasized in prior work (Chow 1970; Bartlett and Wegkamp 2008; El-Yaniv and Wiener 2010; Geifman and El-Yaniv 2017). Within this benchmark, and for the two studied

calibrators, logistic regression was the strongest of the tested compact meta-policies, outperforming a shallow decision tree under NLL, AURC, and the joint regret criterion. The revised case-level deployment reduction further shows that this is not merely a row-level diagnostic: the joint selector chooses one action from raw, temperature scaling, and Dirichlet calibration for each base case, reaches mean ΔNLL -0.0404, mean ΔAURC -1.20×10^{-4}, and reduces NLL harm to 6.1%. Read together with oracle-regret summaries, the small absolute AURC differences support a limited practical takeaway: a simple learned selector can reduce harm relative to always calibrating, even when average gains in selective performance remain modest in raw units.

The evidence does not support a broader claim that a generally valid low-compute skip-calibration rule has been established. The learned selector relies on probe-fit diagnostics, including post-calibration quantities such as rank distortion and bootstrap summaries of calibration-split ΔNLL, so it is more accurately interpreted as a lightweight calibration-selection policy after a candidate calibrator has been fit and inspected, not as a purely pre-calibration screening rule (Zadrozny and Elkan 2001, 2002; Niculescu-Mizil and Caruana 2005; Guo et al. 2017). The empirical scope is likewise restricted: the conclusions are conditional on the completed benchmark cases, the 23 missing case pairs are concentrated rather than random, only temperature scaling and Dirichlet calibration are evaluated, and the available artifacts support AURC- and oracle-regret-based selective summaries but not fixed-coverage risk analyses. Under those constraints, the main contribution is benchmark-conditional evidence that the decision to retain or discard a post-hoc calibrator can itself be learned and should be aligned with downstream coverage-risk trade-offs rather than probability quality alone (Naeini et al. 2015; Kull et al. 2017, 2019; Chen et al. 2019; Geifman and El-Yaniv 2019; Hein et al. 2019). Natural next steps are to test whether the same conclusion survives under richer calibrator families, larger model classes, distribution shift, and operating-point-specific selective objectives.

Data availability

The derived data and analysis code supporting the findings of this study are available in Zenodo at https://doi.org/10.5281/zenodo.21535697. The source benchmark datasets are publicly available from OpenML under the dataset identifiers documented in the archived materials.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 62501380), awarded to Zhipeng Chen.

References

Bartlett PL, Wegkamp MH (2008) Classification with a reject option using a hinge loss. Journal of Machine Learning Research 9:1823–1840

Casalicchio G, Bossek J, Lang M, et al (2019) OpenML: An R package to connect to the machine learning platform OpenML. Computational Statistics 34(3):977–991. https://doi.org/10.1007/s00180-017-0742-2

Chen T, Navratil J, Iyengar VS, et al (2019) Confidence scoring using whitebox meta-models with linear classifier probes. In: Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, pp 1467–1475

Chow CK (1970) On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory 16(1):41–46. https://doi.org/10.1109/TIT.1970.1054406

El-Yaniv R, Wiener Y (2010) On the foundations of noise-free selective classification. Journal of Machine Learning Research 11:1605–1641

Geifman Y, El-Yaniv R (2017) Selective classification for deep neural networks. In: Advances in Neural Information Processing Systems 30, pp 4878–4887

Geifman Y, El-Yaniv R (2019) SelectiveNet: A deep neural network with an integrated reject option. In: Proceedings of the 36th International Conference on Machine Learning, pp 2151–2159

Guo C, Pleiss G, Sun Y, et al (2017) On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning, pp 1321–1330

Hein M, Andriushchenko M, Bitterwolf J (2019) Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 41–50, https://doi.org/10.1109/CVPR.2019.00013

Kull M, de Menezes e Silva Filho T, Flach PA (2017) 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, pp 623–631

Kull M, Perelló-Nieto M, Kängsepp M, et al (2019) Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. In: Advances in Neural Information Processing Systems 32, pp 12295–12305

Naeini MP, Cooper GF, Hauskrecht M (2015) Obtaining well calibrated probabilities using bayesian binning. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, pp 2901–2907, https://doi.org/10.1609/AAAI.V29I1.9602

Niculescu-Mizil A, Caruana R (2005) Predicting good probabilities with supervised learning. In: Proceedings of the Twenty-Second International Conference on Machine Learning, pp 625–632, https://doi.org/10.1145/1102351.1102430

Vanschoren J, van Rijn JN, Bischl B, et al (2013) OpenML: Networked science in machine learning. ACM SIGKDD Explorations Newsletter 15(2):49–60. https://doi.org/10.1145/2641190.2641198

Zadrozny B, Elkan C (2001) Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In: Proceedings of the Eighteenth International Conference on Machine Learning, pp 609–616

Zadrozny B, Elkan C (2002) Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 694–699, https://doi.org/10.1145/775047.775151