machine learning & knowledge extraction

MDPI

Article

Calibration, Architecture, and Distribution Shift in Predictive Uncertainty Estimation

Dina Šabanović *, Tea Krčmar, Zdravko Krpić and Ivica Lukić

Faculty of Electrical Engineering, Computer Science and Information Technology Osijek, Josip Juraj Strossmayer University of Osijek, 31000 Osijek, Croatia; tea.krcmar@ferit.hr (T.K.); zdravko.krpic@ferit.hr (Z.K.); ivica.lukic@ferit.hr (I.L.)

* Correspondence: dina.sabanovic@ferit.hr

Abstract

Reliable uncertainty quantification matters for high-stakes tabular classification, yet the comparative influence of architecture and post-hoc calibration on uncertainty quality remains underexplored, particularly outside in-distribution conditions. We present a matched-protocol benchmark on 36 OpenML-CC18 datasets comparing three GBDTs, a single MLP, MC-Dropout, and deep ensembles of three, five, and ten members under five post-hoc calibration methods, with additional evaluation under Gaussian feature shift, symmetric label noise, split-conformal coverage, and a sub-comparison against TabPFN v2. All paired comparisons use Wilcoxon signed-rank tests with Holm corrections within pre-specified research questions at $\alpha = 0.05$. Using this benchmark, calibrator choice had a larger practical influence on uncertainty metrics than the difference between calibrated GBDTs and five-member neural ensembles. We observe a clear architecture-by-calibrator interaction: Dirichlet calibration was generally strongest on the evaluated GBDTs under low-to-moderate class imbalance, whereas temperature scaling was generally strongest on the evaluated neural models. Under controlled covariate shift, the in-distribution ordering reversed, with multinomial logistic recalibration showing the strongest performance among the tested calibrators on the evaluated neural models under heavy perturbation, and under high class imbalance, the preference for Dirichlet calibration on the evaluated GBDTs weakened. GBDTs were run at recommended defaults rather than tuned per dataset, and the covariate-shift protocol uses synthetic Gaussian noise rather than naturalistic out-of-distribution data; the shift-related findings should be read as directional indicators within this protocol. Calibration and architecture should therefore be selected jointly. The preferred calibrator depends on the model family in-distribution and changes again under perturbation.

Keywords: Bayesian approximations; conformal prediction; covariate shift; deep ensembles; Dirichlet calibration; distribution shift; gradient boosting; tabular data; tabular foundation models; uncertainty quantification

Check for updates

Academic Editor: Francesco Folino

Received: 8 May 2026
Revised: 25 June 2026
Accepted: 26 June 2026
Published: 28 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Reliable uncertainty estimation is critical for machine-learning systems operating in high-stakes domains. In clinical decision support, well-calibrated confidence estimates allow clinicians to triage predictions and escalate uncertain cases to human review [1,2]. In credit scoring and fraud detection, predicted probabilities drive threshold-based decision policies whose economic outcomes depend directly on calibration quality [3]. In

autonomous systems, overconfident predictions remain a primary failure mode [4]. A classifier is calibrated when its predicted probabilities align with empirical frequencies: a confidence of 80% should correspond to an empirical accuracy of approximately 80% [3,5]. Calibration and accuracy are distinct properties and must be evaluated independently. This distinction is especially relevant for tabular classification, which dominates real-world machine-learning deployments [6].

Two methodological gaps motivate this work. First, the calibration properties of gradient-boosted decision trees (GBDTs), the dominant model family on tabular benchmarks [7,8], have received substantially less attention than their accuracy, and the widely held assumption that deep ensembles produce inherently better-calibrated uncertainty estimates has rarely been tested under matched post-hoc calibration. Second, most comparative studies of tabular uncertainty methods restrict evaluation to in-distribution conditions, leaving open whether their conclusions transfer to the covariate shift, label noise, and coverage-based deployment scenarios that motivate uncertainty estimation in the first place [9]. Several promising directions are absent from prior calibration-versus-architecture comparisons on tabular data, including Dirichlet calibration [10], Monte-Carlo dropout as a lightweight Bayesian approximation [11], conformal prediction [12,13], and tabular foundation models such as TabPFN v2 [14]. Full variational Bayesian neural networks [15] are not evaluated here, owing to their per-dataset training cost on tabular workloads, and remain an open direction for future work.

This paper addresses both gaps with a matched-protocol benchmark. Eight models spanning three GBDT implementations, two single-network neural variants, and three ensemble sizes are evaluated under five post-hoc calibration methods on 36 OpenML-CC18 datasets. This results in a total of 7740 in-distribution runs. The same configurations are then re-evaluated under Gaussian feature-shift, symmetric label noise, and split-conformal prediction, and contrasted against TabPFN v2 on its compatible subset of 24 datasets. All paired comparisons use the Wilcoxon signed-rank test with explicit, pre-specified Holm corrections defined separately within each research question, following the controlled benchmarking methodology of Demšar [16] and Benavoli et al. [17]. We additionally log training time for each run and introduce a cost-weighted composite calibration score that surfaces the joint trade-off across NLL, ECE, and Brier [18].

The contributions of this paper are as follows:

1. A controlled factorial benchmark over 8 models, 5 calibration methods, and 36 OpenML-CC18 datasets [19] (7740 in-distribution runs), analysed under paired Wilcoxon tests with explicit Holm corrections within each pre-specified research question.
2. An architecture-by-calibrator interaction that has not been systematically characterised in prior tabular work: Dirichlet calibration [10] is comparable to, and in some cases improves upon, temperature scaling on the evaluated GBDTs but provides limited benefit on the evaluated neural models, and this pattern is stable across most of the NLL/ECE/Brier weight simplex.
3. Robustness analyses covering Gaussian feature shift at three intensities, symmetric label noise at two levels, and stratification by class-imbalance ratio, documenting changes in the in-distribution calibrator ordering: multinomial logistic recalibration [20] yields the strongest median performance for the evaluated neural models under the heavy synthetic-Gaussian covariate-shift protocol tested here, while Dirichlet calibration is no longer favoured for the evaluated GBDTs under high class imbalance.
4. A split-conformal coverage analysis [12,13] at 90% target coverage and a sub-comparison against TabPFN v2 [14] on its 24 compatible datasets, separating marginal-coverage validity, which is achieved by every conformal-evaluated model, from set-

size efficiency, which favours CatBoost on this benchmark, and providing a matched-calibration evaluation of a tabular foundation model against classical baselines.
5. Per-run training-time measurements, an MLP hyperparameter sensitivity ablation over 15 configurations per dataset [21], and a cost-weighted composite calibration score that operationalises the NLL/ECE/Brier trade-off under practitioner-chosen weight profiles.

The remainder of the paper is organised as follows. Section 2 reviews calibration, deep ensembles, conformal prediction, and tabular foundation models. Section 3 describes the experimental protocol. Section 4 reports in-distribution results, Section 5 reports robustness under covariate shift, label noise, and class imbalance, Section 6 reports conformal coverage, and Section 7 reports the TabPFN sub-comparison, training-time analysis, and the cost-weighted recommendation. Section 8 synthesises findings, discusses limitations, and concludes.

2. Related Work

2.1. Post-Hoc Calibration of Probabilistic Classifiers

Guo et al. [3] formalised the empirical study of neural-network calibration, demonstrating that modern networks are systematically overconfident and that temperature scaling reliably reduces this overconfidence with a single scalar parameter. Earlier, Platt [20] proposed fitting a logistic regression on classifier outputs, and Zadrozny and Elkan [22] introduced isotonic regression as a non-parametric alternative, offering greater flexibility but with an elevated risk of overfitting on small calibration sets [5]. For multi-class settings, the Dirichlet calibration of Kull et al. [10] extends Platt-style recalibration to the simplex through a full $K \times K$ affine transformation in log-space, with off-diagonal and intercept regularization (ODIR) to mitigate overfitting. Despite being introduced explicitly to address shortcomings of temperature and Platt scaling in the multi-class regime, Dirichlet calibration is rarely included in comparative calibration studies on tabular data, and its interaction with GBDT-style probability estimates remains untested.

The expected calibration error (ECE) is the most widely used calibration metric but is sensitive to bin granularity [23] and biased relative to true calibration error in small samples [24]. To mitigate the bin-granularity sensitivity, the present study reports a mean ECE across three bin widths, $B \in \{10, 15, 20\}$, following Nixon et al. [23]. Negative log-likelihood and Brier score are reported in parallel as strictly proper scoring rules that respond to both calibration and sharpness [3]. Fredsgaard and Schmidt [25] show that jointly tuning ensemble-level hyperparameters such as weight decay, temperature scaling, and early stopping at the ensemble level rather than per member can improve both performance and calibration.

2.2. Deep Ensembles, Bayesian Approximations, and Conformal Prediction

Deep ensembles aggregate the predictions of multiple independently trained networks and remain the standard scalable baseline for predictive uncertainty in classification [26]. Fort et al. [27] showed ensemble members explore fundamentally different posterior modes, and Wilson and Izmailov [28] argued that deep ensembles marginalise over the weight posterior. Ovadia et al. [9] found that deep ensembles degrade more gracefully than Bayesian approximations under dataset shift, motivating the covariate-shift evaluation included here. On tabular data, Gorishniy et al. [29] recently introduced TabM, a parameter-efficient ensembling framework that achieves state-of-the-art results on standard tabular deep learning benchmarks. Bayesian neural networks [15] offer a principled alternative but are computationally demanding for tabular workloads. Monte-Carlo dropout [11] approximates Bayesian inference by retaining dropout at inference time. It is included here

as a computationally inexpensive Bayesian proxy whose calibration on tabular data has rarely been benchmarked against full ensembles under matched calibration.

Conformal prediction [12] provides a complementary, distribution-free guarantee: given exchangeable calibration and test data, a split-conformal predictor returns prediction sets whose marginal coverage equals the target level $1 - \alpha$ in expectation. Adaptive prediction sets (APSs) [13] extend this to a classification-conditional setting with size-efficient sets. Although conformal coverage and post-hoc calibration target distinct notions of reliability, namely set valued marginal coverage and point probability accuracy, they are rarely evaluated together.

2.3. Tabular Benchmarking and Tabular Foundation Models

Modern tabular benchmarking compares model families on standardised suites under uniform splits and metrics. Grinsztajn et al. [7] and McElfresh et al. [8] confirmed that GBDTs retain an accuracy advantage over deep learning on medium-scale tabular data, while also finding that light hyperparameter tuning can be more consequential than the model-family choice itself, a finding the present hyperparameter ablation revisits directly. Holzmüller et al. [30] demonstrated that strong pre-tuned MLPs are competitive with GBDTs, and Kadra et al. [21] showed that a well-regularised MLP can match top GBDT performance. We use the OpenML-CC18 suite [19] under its official manifest as a standardised reference. TabPFN v2 [14], a prior-data-fitted transformer trained on millions of synthetic tabular tasks, performs classification by in-context learning and bypasses per-dataset gradient training. Its post-hoc calibration behaviour and standing under matched calibration against classical baselines have not previously been characterised. The paired non-parametric testing methodology advocated by Demšar [16] and refined by Benavoli et al. [17] is adopted here through paired Wilcoxon signed-rank tests [31] with pre-specified Holm corrections [32] separately within each research question.

2.4. Synthesis and the Gap Addressed by This Work

The literature contains five recurring gaps that this study fills within a single protocol. First, GBDT calibration is rarely compared under matched post-hoc adjustment. Second, Dirichlet calibration, Monte-Carlo dropout, and conformal prediction are rarely evaluated jointly against deep ensembles on tabular data. Third, calibrator orderings observed in-distribution are rarely tested for stability under covariate or label perturbations, particularly given Ovadia et al.'s [9] finding that uncertainty methods reorder under shift. Fourth, the ensemble-size effect for $M > 5$ and the choice between post-softmax and member-level temperature scaling are seldom quantified. Fifth, tabular foundation models such as TabPFN v2 have not been compared to classical baselines under matched calibration.

3. Methodology

3.1. Overview of the Experimental Protocol

The end-to-end study comprises five sequential stages, described below. Stages 1–3 are run once per dataset and seed. Stage 4 is run once per dataset, seed, and perturbation setting. Stage 5 aggregates across seeds and datasets. Training time is recorded per (model, dataset, seed) run and is analysed together with the cost-weighted recommendation in the results section.

Stage 1. Data preparation. Load an OpenML-CC18 task, integer-encode categoricals, drop constant/all-missing columns, impute remaining missing values by column-wise median, and produce the stratified $X_{\text{train}}/X_{\text{val_es}}/X_{\text{val_cal}}/X_{\text{test}}$ split described in Section 3.3. Fit `StandardScaler` on X_{train} only.

Stage 2. Model fitting. Train one of eight models on X_{train} with early stopping on $X_{\text{val_es}}$. These models comprise three GBDTs, two single-network neural variants, and three ensemble sizes, all described in Section 3.4. Log training time: TabPFN v2 is invoked by in-context inference on the 24 compatible datasets and has no gradient-training cost.

Stage 3. Calibrator fitting. For each model, produce predictions on $X_{\text{val_cal}}$ and fit five post-hoc calibrators on those predictions: Identity, temperature scaling, MLR, isotonic regression, and Dirichlet, each described in Section 3.5. The calibrator never observes X_{train} or $X_{\text{val_es}}$.

Stage 4. Multi-condition evaluation. Apply the fitted (model, calibrator) pair to four test conditions: (a) in-distribution X_{test}, (b) X_{test} perturbed by additive Gaussian noise at $\sigma \in \{0.5, 1.0, 2.0\}$, (c) X_{test} from a model trained on X_{train} with symmetric label noise at $\eta \in \{0.1, 0.2\}$, and (d) split-conformal sets at $1 - \alpha = 0.90$. Record NLL, ECE (mean over $B \in \{10, 15, 20\}$), Brier, accuracy, empirical coverage, and average set size.

Stage 5. Statistical analysis. For each (dataset, model, calibrator) triple, take the per-dataset median across the 5 seeds. Across the 36 datasets, run paired two-sided Wilcoxon signed-rank tests with rank-biserial effect sizes. Apply Holm corrections separately within each pre-specified family: H1 ($k = 39$), H2 ($k = 72$), H3 ($k = 45$) at $\alpha = 0.05$. Additional analyses, namely member-level temperature scaling and the hyperparameter ablation, fall outside the pre-specified H1–H3 test families and are therefore reported with uncorrected p-values.

Figure 1 provides a compact schematic overview of the workflow described above.

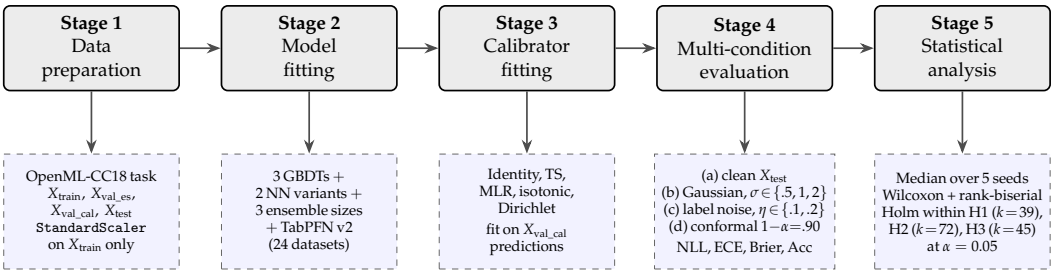


Figure 1. Schematic of the five-stage experimental pipeline. The top row shows the sequential workflow, and the bottom row summarises the principal inputs, configurations, and outputs associated with each stage.

3.2. Research Questions and Factorial Design

This study examines three prespecified hypotheses:

1. No model class exhibits a statistically significant raw-uncertainty advantage over any other model under matched evaluation (H1).
2. Post-hoc calibration methods differ significantly in their impact on uncertainty quality within each model family (H2).
3. After applying the same post-hoc calibration, no statistically significant uncertainty gap remains between deep ensembles and any comparison model (H3).

The design is fully crossed over three factors: model class (LightGBM, XGBoost, CatBoost, SingleMLP, MC-Dropout, DeepEnsemble at $M \in \{3, 5, 10\}$), calibration method (Identity, temperature scaling, MLR, isotonic regression, Dirichlet), and random seed ($\{0, 1, 2, 3, 4\}$). As a result, there are $8 \times 5 \times 5 = 200$ conditions per dataset, or 7200 runs across the 36 OpenML-CC18 tasks. An additional 540 runs apply member-level temperature scaling to the three ensemble sizes, as described in Section 4.6, bringing the total to

7740 in-distribution runs. TabPFN v2 is evaluated on the 24 datasets meeting its inference constraints ($\leq$10,000 training samples, $\leq$500 features, $\leq$10 classes).

3.3. Datasets and Splitting

The OpenML-CC18 benchmark suite (study ID 99) [19] is used, stratifying tasks into three dataset sizes: small ($n < 1000$), medium ($1000 \leq n < 10{,}000$), and large ($n \geq 10{,}000$). For each size, 12 tasks are sampled (seed = 0), resulting in 36 datasets. Datasets in which any class contains fewer than 20 samples are excluded. Categorical features are integer-encoded, while columns that are constant or contain only missing values are removed. Missing values are imputed using the column-wise median. All features are standardised using scikit-learn's StandardScaler [33], fitted exclusively on X_{train}.

A stratified 60/20/20 split is adopted, following Guo et al. [3] and Gorishniy et al. [34]. The 60% training portion is further divided 90/10 into X_{train} ($\approx$54% overall) and the early-stopping validation partition $X_{\text{val_es}}$ ($\approx$6%); the 40% held-out portion is split equally into the calibrator-fitting partition $X_{\text{val_cal}}$ (20%) and the test partition X_{test} (20%). The key invariant is that $X_{\text{val_es}}$ and $X_{\text{val_cal}}$ are disjointed with strictly separate roles, avoiding the calibration leakage that can arise when the same validation partition is used for both early stopping and calibrator fitting. Table 1 summarises the partition proportions and roles.

Table 1. Data-split protocol. All splits are stratified by class label and fixed per dataset and seed.

Partition	Fraction	Purpose
X_{train}	$\approx$54%	Model weight learning
$X_{\text{val_es}}$	$\approx$6%	Early stopping only, never used for calibration
$X_{\text{val_cal}}$	20%	Calibrator fitting only, never seen during training
X_{test}	20%	Final evaluation, strictly held out throughout

3.4. Models and Hyperparameters

A single, fixed configuration is used per model class across all 36 datasets, since per-dataset hyperparameter optimisation could confound model-family differences with tuning quality, and GBDTs are considerably less sensitive to hyperparameter selection than MLPs. The hyperparameter sensitivity ablation in Section 4.7 directly assesses whether this design choice influences the main results. The selected hyperparameters follow related-work recommendations and widely adopted defaults for tabular benchmarks [7,8,21,34]. Three GBDT implementations (LightGBM [35], XGBoost [36], CatBoost [37]) cover the families that dominate Kaggle-style leaderboards and recent academic comparisons. SingleMLP and MC-Dropout cover the deterministic and Bayesian-approximation neural baselines [11,26], and deep ensembles at $M \in \{3, 5, 10\}$ cover the practical range of ensemble cardinalities [26,29]. TabPFN v2 [14] represents tabular foundation models. Table 2 summarises the configurations.

The deep-ensemble predictive distribution is

$$\hat{p}_{\text{ens}}(y \mid x) = \frac{1}{M} \sum_{m=1}^{M} \hat{p}_m(y \mid x). \tag{1}$$

All experiments were implemented in Python 3.12, using the following package versions: PyTorch 2.11.0, LightGBM 4.3.0, XGBoost 3.2.0, CatBoost 1.2.10, scikit-learn 1.6.1, dirichletcal 0.5.3, TabPFN v2 8.0.8, openml 0.15.1, numpy 2.0.2, pandas 2.2.2, scipy 1.16.3, and statsmodels 0.14.6.

Table 2. Model configurations. A single fixed configuration per model class is applied across all 36 datasets. The hyperparameter sensitivity ablation in Section 4.7 confirms the main conclusions are not artefacts of this choice.

Model	Configuration
LightGBM	lr 0.05, num_leaves 63, min_child_samples 20, feature/bagging fraction 0.8, L_1/L_2 reg 0.1, up to 500 rounds with early stopping
XGBoost	analogous settings (tree booster, max_depth 6, eta 0.05, subsample 0.8)
CatBoost	depth 6, lr 0.05, iterations 500, l2_leaf_reg 3.0, early stopping, trained on the same preprocessed feature matrix as the other models
SingleMLP	$d_{\text{in}} \to 256 \to 128 \to K$, BN, ReLU, 10% dropout, AdamW (lr 10^{-3}, wd 10^{-4}), cosine annealing 200 epochs, ℓ_2 clip 1.0, early-stop patience 20
MC-Dropout	same as SingleMLP, dropout retained at inference, $T = 30$ stochastic passes averaged
DeepEnsemble	M independent SingleMLP instances, $M \in \{3, 5, 10\}$, predictive distribution: mean of post-softmax member outputs
TabPFN v2	pre-trained transformer, in-context inference on training set, no per-dataset gradient training

3.5. Calibration and Evaluation

The uncertainty quality of each (model, calibrator) combination is assessed using three complementary metrics, following the multi-metric evaluation practice established in modern probabilistic-classification studies [3,23]. Negative log-likelihood

$$\text{NLL} = -\frac{1}{n}\sum_{i=1}^{n} \log \hat{p}(y_i \mid x_i) \tag{2}$$

is a strictly proper scoring rule sensitive to both miscalibration and sharpness. Expected calibration error

$$\text{ECE} = \sum_{b=1}^{B} \frac{|B_b|}{n} \left|\text{acc}(B_b) - \text{conf}(B_b)\right| \tag{3}$$

is computed at $B \in \{10, 15, 20\}$ equal-width bins and averaged. Predicted probabilities are clipped to $[\epsilon, 1-\epsilon]$ with $\epsilon = 10^{-8}$ before binning, so no confidence coincides exactly with a bin boundary. The Brier score

$$\text{Brier} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} \left(\hat{p}_{ik} - \mathbf{1}[y_i = k]\right)^2 \tag{4}$$

is a bounded proper scoring rule less sensitive to extreme probability values than NLL. Accuracy is recorded for reference but is not subject to hypothesis testing.

Five post-hoc calibration methods are evaluated, fitted exclusively on $X_{\text{val_cal}}$ predictions and applied to test predictions. Identity returns raw probabilities clipped to $[\epsilon, 1-\epsilon]$. Temperature scaling [3] fits a single scalar $T > 0$ via the limited-memory BFGS optimiser (L-BFGS) [38] to minimise NLL on $X_{\text{val_cal}}$:

$$\hat{p}_k^{\text{cal}} = \frac{\exp(\log \hat{p}_k / T)}{\sum_j \exp(\log \hat{p}_j / T)}. \tag{5}$$

For models that do not produce raw pre-softmax logits (the deep ensemble after averaging, and the GBDTs), temperature scaling is applied to a pseudo-logit $z_k = \log \hat{p}(y = k \mid x)$. This pseudo-logit approximation is applied consistently across all model families and represents the standard prediction-averaging protocol, making it the most operationally relevant

comparison for H3. Isotonic regression and Dirichlet calibration provide convergent evidence under non-temperature calibration methods. Multinomial logistic recalibration (MLR) [20] fits a multinomial logistic regression on predicted probability vectors as a multi-class extension of Platt scaling (`sklearn LogisticRegression`, solver lbfgs, $C = 1.0$). Isotonic regression [22] fits one `IsotonicRegression` per class against binary one-vs-rest labels, with renormalisation at inference. Dirichlet calibration [10] fits a $K \times K$ affine transformation in log-simplex with off-diagonal-and-intercept regularisation (ODIR).

3.6. Statistical Analysis

For each (dataset, model, calibrator) triple, the five seed-level results are first aggregated to a per-dataset median to reduce seed variance. The resulting $n = 36$ per-dataset medians constitute the paired observations for hypothesis testing, where the natural unit of replication is the dataset rather than the seed. All pairwise comparisons use the two-sided paired Wilcoxon signed-rank test [31]. Holm corrections [32] are applied separately within each pre-specified test family: H1 ($k = 39$), H2 ($k = 72$), H3 ($k = 45$) at $\alpha = 0.05$, controlling the family-wise error rate within each research question. The k values decompose as follows. H1 ($k = 39$) tests thirteen pairwise raw-model comparisons across three metrics: each of three neural representatives (SingleMLP, MC-Dropout, deep ensemble at $M = 5$) against each of the three GBDTs (nine pairs), each of the three ensemble cardinalities against SingleMLP (three pairs), and MC-Dropout against SingleMLP (one pair). H2 ($k = 72$) tests four calibrators against raw within each of six model families (three GBDTs + SingleMLP, MC-Dropout, $M = 5$ ensemble) on three metrics. H3 ($k = 45$) tests the $M = 5$ ensemble against five comparison models under three matched calibrators (temperature, isotonic, Dirichlet) on three metrics. Comparisons between ensemble cardinalities form a separate family of $k = 18$ tests. The hyperparameter ablation in Section 4.7 and the member-level temperature-scaling comparison in Section 4.6 fall outside these families and therefore report uncorrected p-values, clearly labelled as additional analyses. Practical magnitude is reported as the signed rank-biserial correlation

$$r = \frac{T^+ - T^-}{n(n+1)/2}, \tag{6}$$

with conventional thresholds $|r| < 0.3$ (small), 0.3–0.5 (medium), and >0.5 (large). For paired calibrator-vs-raw comparisons, Win/Tie/Loss (W/T/L) counts over the 36 datasets are additionally reported, with a tolerance of 10^{-3} for NLL and a strict comparison for ECE.

4. Results: In-Distribution

4.1. Overview

Figure 2 presents per-dataset median values of the three primary uncertainty metrics across all 40 model–calibrator conditions. Four patterns emerge: (1) In-distribution, MLR consistently degrades NLL and ECE across all model families, (2) isotonic regression frequently improves ECE but degrades NLL, (3) temperature scaling and Dirichlet calibration occupy the best two cells in nearly every column, with their relative ranking dependent on model family, and (4) Brier-score differences are smaller in magnitude than NLL differences. These patterns are statistically evaluated in subsequent analyses.

The heatmap shows the architecture-by-calibrator interaction examined statistically in Section 4.3: Dirichlet calibration produces the deepest blue cells for the three GBDTs but does not for the three neural model classes.

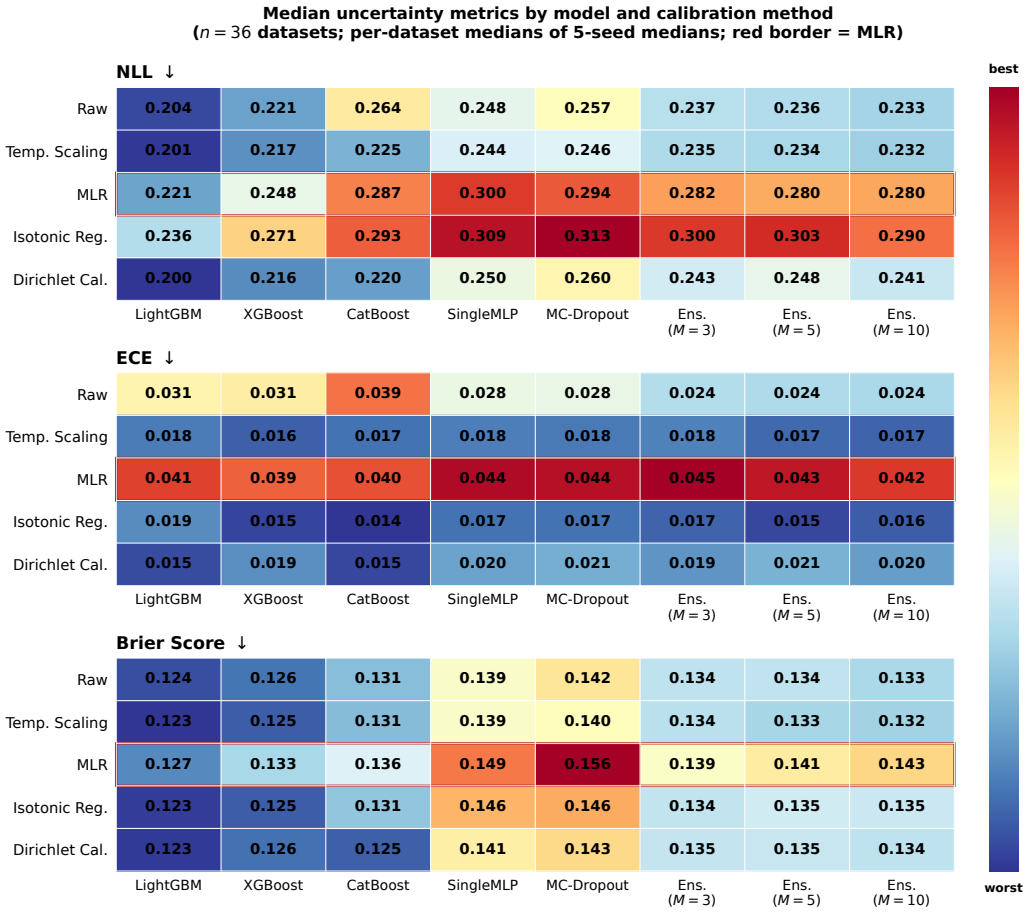


Figure 2. Per-dataset median uncertainty metrics by model and calibration method ($n = 36$ datasets, per-dataset medians of 5-seed medians). Three stacked panels show NLL (**top**), ECE (**middle**), and Brier score (**bottom**). Cell values are the median across datasets, and colour is normalised within each metric (blue = best, red = worst). The MLR row is highlighted with a red border to mark the calibrator that consistently degrades all metrics.

4.2. H1: Raw Uncertainty Across Model Classes

Table 3 presents the raw, uncalibrated H1 comparisons under Holm correction over the H1 family ($k = 39$ tests). Three results hold consistently. First, no neural model achieves a statistically significant raw-uncertainty advantage over any of the three GBDTs on any evaluated metric (all $|r| \leq 0.27$, $p_{\text{adj}} = 1.0$). Second, the deep ensemble at every cardinality $M \in \{3, 5, 10\}$ demonstrates a significant raw advantage over the fixed-configuration SingleMLP on NLL and Brier, with uniformly large effects (full statistics in Table 3). Third, MC-Dropout exhibits no significant difference from SingleMLP on any metric, indicating that, within this benchmark, MC-Dropout provides no measurable uncertainty benefit over a single deterministic MLP at the configurations tested.

Pairwise increments between ensemble sizes are tested separately as an additional ensemble-size analysis ($k = 18$). Under Holm correction within this analysis, $M = 10$ significantly improves NLL over both $M = 5$ and $M = 3$, and improves Brier over $M = 3$; the $M = 5$ ensemble also improves Brier over $M = 3$ (all four pairs: large effects, $p_{\text{adj}} \leq 0.015$). All ECE differences across ensemble sizes are non-significant.

These findings strengthen and extend the original H1 conclusion across three dimensions: the null over LightGBM also holds against XGBoost and CatBoost as alternative GBDT implementations, against MC-Dropout as a Bayesian approximation, and across three ensemble cardinalities.

Table 3. H1 statistical results: pairwise raw model comparisons under Holm correction over the H1 family ($k = 39$ tests, $n = 36$ datasets per pair). Only the six significant comparisons ($p_{adj} < 0.05$) are listed.

Metric	Comparison	Med A	Med B	Diff	p_{adj}	r	Mag.
NLL	Ens. $M = 3$ vs. SingleMLP	0.2373	0.2476	−0.0103	<0.001	−0.847	large
NLL	Ens. $M = 5$ vs. SingleMLP	0.2357	0.2476	−0.0119	<0.001	−0.835	large
NLL	Ens. $M = 10$ vs. SingleMLP	0.2331	0.2476	−0.0145	<0.001	−0.871	large
Brier	Ens. $M = 3$ vs. SingleMLP	0.1342	0.1389	−0.0047	<0.001	−0.928	large
Brier	Ens. $M = 5$ vs. SingleMLP	0.1342	0.1389	−0.0047	<0.001	−0.919	large
Brier	Ens. $M = 10$ vs. SingleMLP	0.1330	0.1389	−0.0059	<0.001	−0.943	large

Notable near-significant trend (failed Holm correction in this family): Ens. $M = 5$ vs. CatBoost on ECE ($r = -0.571$, $p_{adj} = 0.068$). All ECE pairwise comparisons in the H1 family are non-significant. Ensemble-size effects ($M = 3$, $M = 5$, $M = 10$) are tested separately as an additional ensemble-size analysis ($k = 18$).

4.3. H2: Calibration Effects Within Model Families

Within-family calibration effects are stronger and more systematic than raw model-family differences. Table 4 reports H2 results for the six main model families under Holm correction over the H2 family ($k = 72$ tests). The principal new finding in this section is a clear architecture-by-calibrator interaction for Dirichlet calibration: it is the strongest calibrator on ECE for every GBDT and on Brier for LightGBM and CatBoost, but provides no significant improvement on any metric for any neural model. The magnitude is practically large: switching from MLR to Dirichlet reduces ECE by 52–62% and NLL by 9.5–23.2% across all eight models (per-model values in Table 4). The near-identical patterns across LightGBM, XGBoost, and CatBoost suggest that the finding holds across the three GBDT implementations evaluated here. Three further patterns are visible directly in Table 4: temperature scaling significantly improves NLL for every GBDT and for both single-network neural models (with a medium effect on the deep ensemble that fails Holm correction after the family expansion); MLR significantly worsens NLL across all six families with uniformly large effect sizes; and isotonic regression significantly worsens NLL for all six families while leaving ECE and Brier mostly unchanged.

Table 4. H2 statistical results across the six main model families ($n = 36$, Holm correction over the H2 family of $k = 72$ tests). Each row reports a comparison of Raw versus a single calibrator. The five-member deep ensemble is used as the ensemble representative in this table, consistent with the H3 post-calibration comparison. Bold = significant ($p_{adj} < 0.05$).

Model	Calibrator	NLL p_{adj}	NLL r	ECE p_{adj}	ECE r	Brier p_{adj}	Brier r
LightGBM	Temp	**0.029**	**−0.622**	**0.002**	**−0.742**	**0.002**	**−0.748**
LightGBM	MLR	**<0.001**	**+0.970**	**0.029**	**+0.622**	**0.029**	**+0.622**
LightGBM	Isotonic	**<0.001**	**+0.796**	0.330	−0.471	1.000	+0.138
LightGBM	Dirichlet	1.000	−0.336	**0.026**	**−0.628**	**0.021**	**−0.640**
XGBoost	Temp	**0.014**	**−0.658**	**0.001**	**−0.769**	0.082	−0.593
XGBoost	MLR	**<0.001**	**+0.955**	**0.008**	**+0.682**	**0.016**	**+0.652**
XGBoost	Isotonic	**<0.001**	**+0.865**	0.326	−0.474	1.000	+0.084
XGBoost	Dirichlet	1.000	−0.309	**0.021**	**−0.640**	0.067	−0.577
CatBoost	Temp	**<0.001**	**−0.883**	**<0.001**	**−0.874**	**<0.001**	**−0.919**
CatBoost	MLR	**<0.001**	**+0.808**	1.000	+0.333	0.840	+0.384
CatBoost	Isotonic	**0.035**	**+0.610**	**0.001**	**−0.754**	0.825	−0.393
CatBoost	Dirichlet	0.067	−0.577	**<0.001**	**−0.823**	**0.001**	**−0.757**
SingleMLP	Temp	**0.008**	**−0.685**	0.440	−0.447	0.825	−0.393
SingleMLP	MLR	**<0.001**	**+0.997**	**<0.001**	**+0.778**	**<0.001**	**+0.790**
SingleMLP	Isotonic	**<0.001**	**+0.883**	0.798	−0.399	1.000	+0.261
SingleMLP	Dirichlet	1.000	−0.093	1.000	−0.222	1.000	−0.063

Table 4. *Cont.*

Model	Calibrator	NLL p_{adj}	NLL r	ECE p_{adj}	ECE r	Brier p_{adj}	Brier r
MC-Dropout	Temp	**0.007**	**−0.691**	0.221	−0.505	0.840	−0.384
MC-Dropout	MLR	**<0.001**	**+0.997**	**<0.001**	**+0.790**	**<0.001**	**+0.892**
MC-Dropout	Isotonic	**<0.001**	**+0.949**	1.000	−0.306	1.000	+0.132
MC-Dropout	Dirichlet	1.000	−0.003	1.000	−0.354	1.000	−0.024
Ens. $M = 5$	Temp	0.275	−0.489	0.159	−0.526	0.323	−0.477
Ens. $M = 5$	MLR	**<0.001**	**+0.997**	**<0.001**	**+0.826**	**<0.001**	**+0.862**
Ens. $M = 5$	Isotonic	**<0.001**	**+0.919**	0.708	−0.411	0.438	+0.450
Ens. $M = 5$	Dirichlet	1.000	+0.171	1.000	−0.195	1.000	+0.099

Per-Dataset Reliability of Temperature Scaling

A focus on central tendency understates the consistency of the temperature-scaling effect, as the small median improvements (-0.0003 to -0.0055) are overshadowed by the broader across-dataset NLL range. Figure 3 presents the paired per-dataset difference $\Delta\text{NLL} = \text{NLL}(\text{Temp. Scaling}) - \text{NLL}(\text{Raw})$. Temperature scaling improves NLL on the majority of datasets for every model class (W/T/L counts and uncorrected p-values are annotated per row in the figure, with all uncorrected $p \leq 0.014$). CatBoost shows the largest effect ($\Delta_{\text{med}} = -0.0055$, $p < 0.0001$), consistent with its raw probabilities exhibiting the largest miscalibration in the benchmark (raw ECE $= 0.0385$ in Figure 2).

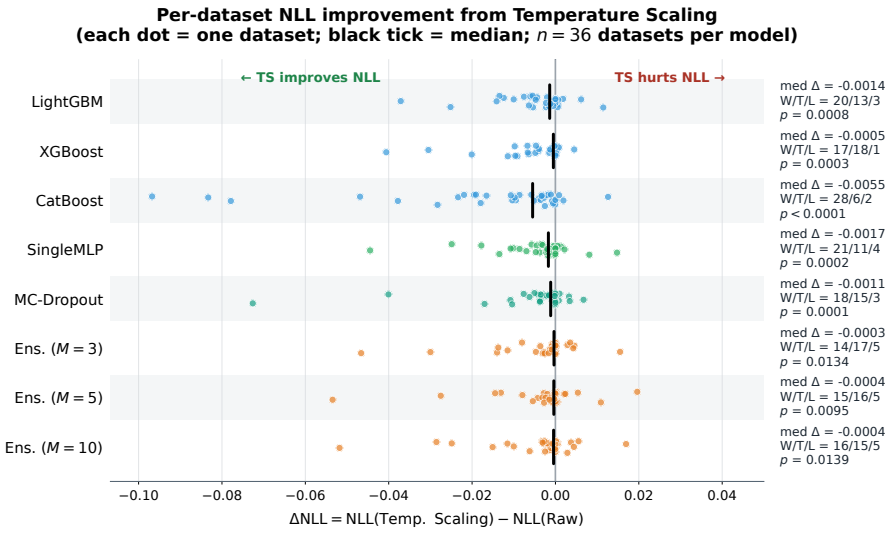


Figure 3. Per-dataset NLL improvement from temperature scaling. Each dot is one dataset (median over 5 seeds). The black tick is the median across datasets, and the annotation panel on the right reports the median Δ, W/T/L counts, and uncorrected Wilcoxon p.

4.4. Calibration Effects on ECE: Per-Dataset View

Figure 4 presents the joint $(\Delta\text{ECE}, \Delta\text{NLL})$ plane. Temperature scaling improves ECE on the majority of datasets for every model (28/36 for LightGBM, 28/36 for XGBoost, 30/36 for CatBoost, 23/36 for SingleMLP, 27/36 for MC-Dropout, and 25–27/36 across the three ensemble sizes). Isotonic regression produces similar ECE improvements but also leads to NLL degradation on the same datasets, as seen in the "isotonic trap" visible in the upper-left quadrant of Figure 4a. Temperature scaling shows no such trade-off (Figure 4b): both metrics move toward improvement together. Optimising any single calibration metric in isolation can therefore mask large degradations in the other proper scoring rules, motivating the later cost-weighted composite analysis.

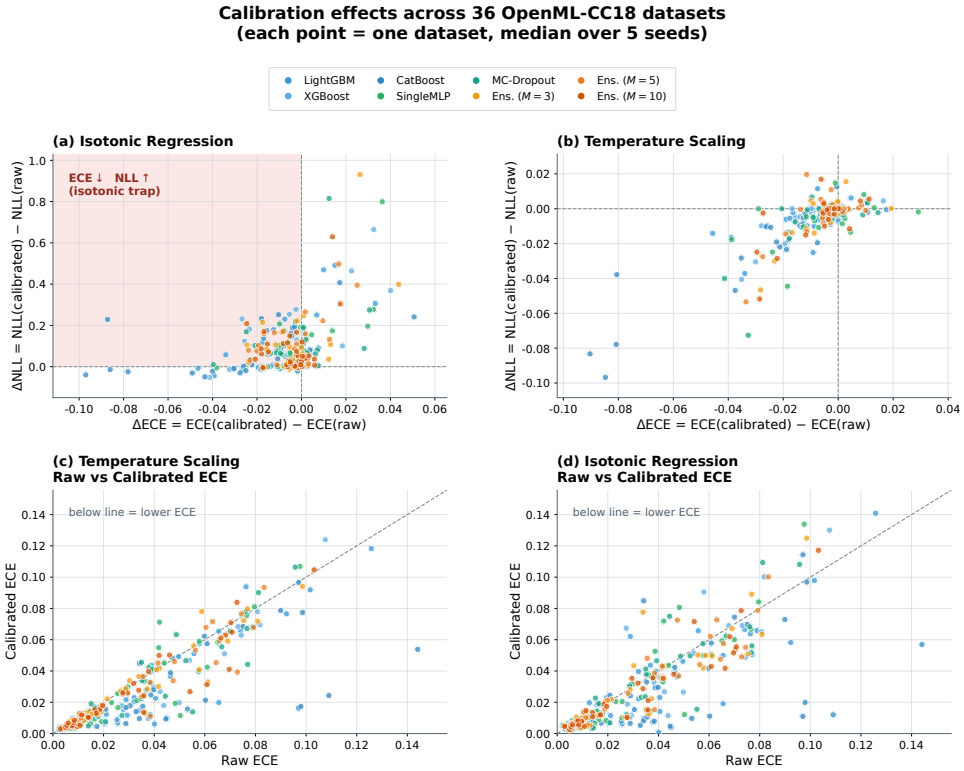


Figure 4. Four-panel diagnostic of calibration effects across 36 OpenML-CC18 datasets (each point = one dataset, median over 5 seeds, with dots coloured by model class). (**a**) ΔNLL vs. ΔECE under isotonic regression. The shaded upper-left quadrant marks the "isotonic trap". (**b**) The same for temperature scaling: both metrics move together, near the origin. (**c**) Raw vs. calibrated ECE under temperature scaling. (**d**) The same for isotonic regression.

4.5. H3: Post-Calibration Gap Between Model Classes

Table 5 presents the H3 family ($k = 45$ tests, Holm-corrected). Under every matched calibrator and metric, no comparison between the five-member deep ensemble and any of the three GBDTs reaches statistical significance ($|r| \leq 0.32$ for all 27 GBDT comparisons, all $p_{\text{adj}} = 1.0$). Thus, within the fixed-configuration design and the 36 OpenML-CC18 datasets considered here, we do not detect a statistically significant uncertainty-quality advantage of five-member neural ensembles over calibrated GBDTs. This should be interpreted as absence of evidence for an ensemble advantage under the evaluated protocol, rather than as a formal equivalence result. Against the two single-network neural models, the ensemble retains a significant advantage on NLL and Brier under temperature scaling and Dirichlet calibration, with large effect sizes, and on Brier under isotonic regression, as shown in Table 5. ECE differences are non-significant under all conditions. Under the fixed-hyperparameter design, calibration does not close the ensemble–single-MLP gap on likelihood-based metrics in this benchmark. Section 4.7 examines this gap under per-dataset MLP tuning; see Section 8.3 for the broader hyperparameter-fairness caveat.

4.6. Standard vs. Member-Level Temperature Scaling

For deep ensembles, the standard practice of applying a single temperature to the post-softmax member average can be replaced by fitting per-member temperatures T_m and averaging the rescaled softmaxes [25]. Table 6 reports paired comparisons across the three ensemble sizes (p-values are uncorrected, since this comparison falls outside the main H1/H2/H3 families). Rank-biserial correlations are positive across all sizes and metrics: on the majority of datasets, member-level temperature scaling produces worse uncertainty quality than the standard post-softmax variant. The degradation is largest and

most significant for ECE at $M = 10$. On this benchmark, member-level temperature scaling therefore does not provide a reliable improvement over the standard protocol.

Table 5. H3 statistical results: five-member deep ensemble vs. comparison models under matched calibration ($n = 36$, Holm correction over the H3 family of $k = 45$ tests). Only the 10 significant rows ($p_{\text{adj}} < 0.05$) are listed. All 27 Ens. vs. GBDT comparisons (3 GBDTs × 3 calibrators × 3 metrics) yield $p_{\text{adj}} = 1.0$ with $|r| \leq 0.32$, and all 15 ECE comparisons (5 comparison models × 3 calibrators) are non-significant.

Met.	Comparison	Ens.	Comp.	Diff	p_{adj}	r	Mag.
NLL	Ens vs. SingleMLP (Temp)	0.2340	0.2444	−0.0104	<0.001	−0.913	large
NLL	Ens vs. SingleMLP (Dirichlet)	0.2480	0.2500	−0.0020	<0.001	−0.814	large
NLL	Ens vs. MC-Dropout (Temp)	0.2340	0.2455	−0.0115	<0.001	−0.904	large
NLL	Ens vs. MC-Dropout (Dirichlet)	0.2480	0.2601	−0.0121	<0.001	−0.847	large
Brier	Ens vs. SingleMLP (Temp)	0.1326	0.1389	−0.0063	<0.001	−0.904	large
Brier	Ens vs. SingleMLP (Dirichlet)	0.1347	0.1411	−0.0064	<0.001	−0.985	large
Brier	Ens vs. SingleMLP (Isotonic)	0.1352	0.1455	−0.0103	0.002	−0.727	large
Brier	Ens vs. MC-Dropout (Temp)	0.1326	0.1396	−0.0070	<0.001	−0.916	large
Brier	Ens vs. MC-Dropout (Dirichlet)	0.1347	0.1432	−0.0085	<0.001	−0.940	large
Brier	Ens vs. MC-Dropout (Isotonic)	0.1352	0.1463	−0.0111	0.002	−0.736	large

Notable near-significant: Ens vs. SingleMLP (Isotonic) on NLL: $r = -0.517$, $p_{adj} = 0.210$.

Table 6. Member-level versus standard (post-softmax) temperature scaling across the three ensemble sizes ($n = 36$ datasets, uncorrected paired Wilcoxon). Positive r indicates that member-level scaling is worse than standard scaling on the majority of datasets.

Model	Metric	Med (Member)	Med (Standard)	r	p	Mag.
Ens. $M = 3$	NLL	0.2348	0.2350	+0.604	0.001	large
Ens. $M = 3$	ECE	0.0224	0.0181	+0.225	0.245	small
Ens. $M = 3$	Brier	0.1338	0.1338	+0.480	0.011	medium
Ens. $M = 5$	NLL	0.2327	0.2340	+0.610	0.001	large
Ens. $M = 5$	ECE	0.0220	0.0169	+0.219	0.258	small
Ens. $M = 5$	Brier	0.1332	0.1326	+0.571	0.002	large
Ens. $M = 10$	NLL	0.2324	0.2318	+0.724	0.0001	large
Ens. $M = 10$	ECE	0.0242	0.0166	+0.697	0.0001	large
Ens. $M = 10$	Brier	0.1323	0.1324	+0.348	0.069	medium

4.7. Hyperparameter Sensitivity Ablation

The fixed-MLP design is the principal threat to internal validity: GBDTs are substantially less sensitive to hyperparameter selection than MLPs [7,8]. We evaluate 15 distinct MLP configurations per dataset, sampled from the grid

$$\begin{aligned}\text{hidden layers} &\in \{(64,32),(128,64),(256,128),(512,256),(256,128,64),(512,256,128)\},\\ \text{learning rate} &\in \{10^{-4}, 5\times10^{-4}, 10^{-3}, 3\times10^{-3}\},\\ \text{dropout} &\in \{0, 0.05, 0.1, 0.2, 0.3\},\\ \text{weight decay} &\in \{0, 10^{-5}, 10^{-4}, 10^{-3}\},\end{aligned}$$

with the fixed configuration (config 0) included as a 15th anchor. All 15 are trained over 5 seeds on the same 36 datasets, calibrated with temperature scaling on $X_{\text{val_cal}}$, and evaluated on X_{test}. *SingleMLP_Tuned* is the oracle-best configuration per dataset, selected by minimum mean validation NLL. The fixed configuration achieved the lowest validation NLL on only 1 of 36 datasets (task: `sick`). The median validation NLL improvement from tuning was −0.010 (IQR −0.028 to −0.006). This is an oracle upper bound, as practical deployment would require nested cross-validation because both configuration selection and calibrator fitting currently use $X_{\text{val_cal}}$. Table 7 reports per-dataset median post-temperature-

scaling metrics for the fixed and oracle-tuned MLP configurations alongside the GBDT and deep-ensemble baselines.

Table 7. Per-dataset median post-temperature-scaling metrics ($n = 36$ datasets) and secondary Wilcoxon tests (p-values uncorrected, outside the main H1/H2/H3 families). Tuning closes the SingleMLP–ensemble gap on NLL ($p = 0.451$, small effect) and the SingleMLP–LightGBM gap ($p = 0.636$, small effect).

Model	NLL ↓	ECE ↓	Brier ↓
LightGBM	0.2010	0.0178	0.1229
XGBoost	0.2173	0.0164	0.1250
CatBoost	0.2247	0.0174	0.1307
SingleMLP_Fixed	0.2438	0.0179	0.1389
SingleMLP_Tuned	0.2127	0.0159	0.1327
DeepEnsemble ($M = 5$)	0.2340	0.0169	0.1326

Per-dataset tuning significantly improves SingleMLP on both NLL and Brier with large effect sizes (Table 7), confirming the fixed configuration is genuinely suboptimal. Despite this improvement, the tuned MLP does not significantly differ from LightGBM on any metric, nor from the deep ensemble, and numerically outperforms the ensemble on median post-TS NLL (0.2127 vs. 0.2340). The ensemble–single-MLP gap of Sections 4.2 and 4.5 therefore appears partly attributable to the fixed-HP design, with the caveat that the oracle tuning used here upper-bounds what deployable nested cross-validation would achieve.

5. Results: Robustness Under Distribution Shift, Label Noise, and Class Imbalance

The in-distribution results above establish a clear calibrator ordering on X_{test} drawn from the training distribution. This section examines whether that ordering persists when deployment conditions differ or when the dataset structure deviates from the balanced-class average. Two perturbations are considered: additive Gaussian noise on the test features (Section 5.1) and symmetric label noise on the training labels (Section 5.2). Both run at a single seed and are reported descriptively as a qualitative robustness check, with no hypothesis-test correction performed. A third analysis (Section 5.3) stratifies the in-distribution five-seed results by class-imbalance ratio and therefore inherits the full statistical power of Section 4.3.

5.1. Covariate Shift via Gaussian Feature Noise

Figure 5 traces median test-set NLL across the 36 datasets as Gaussian feature-shift intensity σ increases from 0 (in-distribution) through 0.5, 1.0, and 2.0. Three patterns dominate. First, the in-distribution best calibrator is no longer best at moderate or heavy shift: for all four neural-model variants (SingleMLP, MC-Dropout, deep ensembles at $M \in \{3, 5, 10\}$), MLR achieves the lowest median NLL at both $\sigma = 1.0$ and $\sigma = 2.0$, with reductions of roughly 35–40% relative to temperature scaling at $\sigma = 2.0$. Per-dataset Win/Tie/Loss counts confirm this is not a median-driven artefact: at $\sigma = 2.0$, MLR beats temperature scaling on 33–34 of 36 datasets for every neural model, and on 26–30 of 36 at $\sigma = 1.0$.

Second, isotonic regression already exhibits the largest NLL degradation in-distribution and collapses further under shift. SingleMLP NLL under isotonic rises from 0.31 in-distribution to 5.21 at $\sigma = 2.0$. This pattern is consistent with the rank-based mechanism behind isotonic regression: by pooling adjacent probability values, it produces sharp probabilities that minimise ECE but assign near-zero mass to classes that turn out to be correct after shift, sending the log-likelihood arbitrarily large on those examples. Third, the GBDT

panels show a more nuanced ordering. At $\sigma = 1.0$, the four non-isotonic calibrators bunch within 0.05 NLL of each other for LightGBM and CatBoost. At $\sigma = 2.0$, MLR moves into the lead for LightGBM and XGBoost but CatBoost still prefers temperature scaling. The smaller GBDT reversal is consistent with GBDT raw probabilities being intrinsically less extreme than neural-network softmaxes.

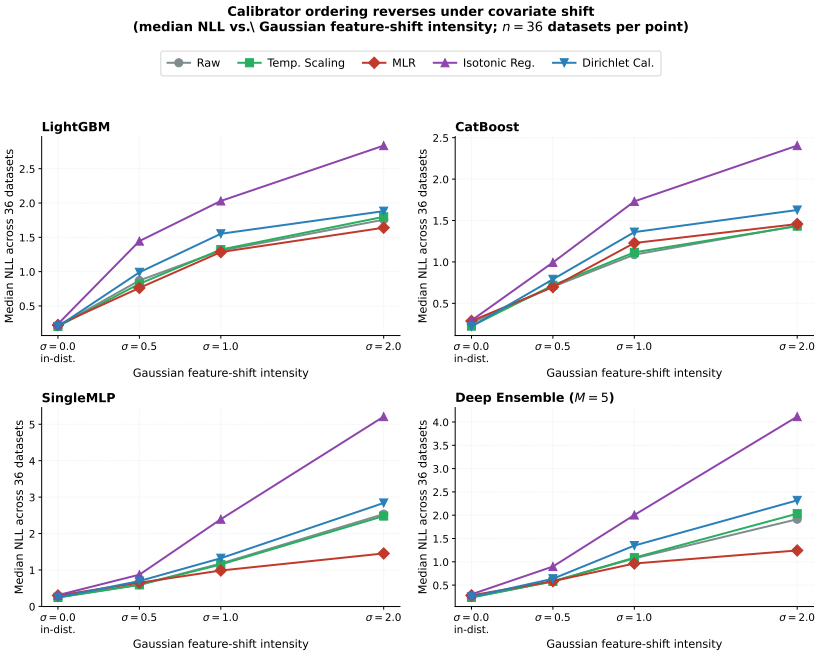


Figure 5. Calibrator ordering reverses under covariate shift. Each panel shows median NLL across the 36 datasets as Gaussian feature-shift intensity σ increases from 0 (in-distribution, leftmost) to 2 (heavy shift, rightmost). Five lines per panel correspond to the five calibration methods. The in-distribution point ($\sigma = 0$) is the per-dataset median across five seeds, and shifted points are at a single seed across the same 36 datasets.

A mechanistic interpretation, expanded in Section 8.2, is that under heavy shift the in-distribution corrections learned by temperature scaling and Dirichlet calibration become miscalibrated themselves, while MLR's richer $K \times K$ transformation can pull aggressively toward the marginal class distribution, which is closer to correct under heavy shift than any miscalibrated training-distribution estimate. Within this synthetic shift protocol, the in-distribution calibrator ranking does not transfer to heavily perturbed test data; Section 8.3 discusses why this finding should be read as a directional indicator under controlled shift rather than a recommendation for naturalistic out-of-distribution settings.

5.2. Symmetric Label Noise

Table 8 reports median test-set metrics under symmetric label noise at $\eta \in \{0.1, 0.2\}$ across six representative datasets (`kc2`, `cmc`, `cylinder-bands`, `MiceProtein`, `pendigits`, `har`). Here, "symmetric" denotes uniform resampling of the η-fraction over all K classes (including the original), so the realised flip rate is $\eta(K-1)/K$ rather than η exactly. For $K = 2$, this halves the effective rate and for $K = 10$ it leaves it nearly unchanged. Three observations follow. First, the in-distribution calibrator ordering is preserved under label noise: temperature scaling and Dirichlet calibration remain the strongest calibrators, while MLR and isotonic regression remain weakest. This contrasts with the covariate-shift finding: because label noise corrupts X_{train} but leaves X_{test} exchangeable with $X_{\text{val_cal}}$, the calibrator fitted on $X_{\text{val_cal}}$ remains valid for the test distribution. Second, model-level robustness varies markedly. At $\eta = 0.2$, CatBoost with temperature scaling achieves the lowest NLL (0.274), narrowly beating CatBoost/Dirichlet (0.286) and Ens. $M = 5$/Dirichlet (0.307). The

five-member ensemble degrades by 1.6× between $\eta = 0.1$ and $\eta = 0.2$ whereas CatBoost degrades only by 1.1×. Third, the ensemble–single-MLP gap that survived matched calibration in Section 4.5 also persists under label noise on this subset.

Table 8. Median uncertainty metrics under symmetric label noise on X_{train}, computed at a single seed across six datasets. Lower is better for all metrics.

		$\eta = 0.1$			$\eta = 0.2$		
Model	**Calibrator**	**NLL**	**ECE**	**Brier**	**NLL**	**ECE**	**Brier**
LightGBM	Raw	0.255	0.049	0.147	0.342	0.065	0.193
LightGBM	Temp	0.257	0.045	0.152	0.318	0.041	0.187
LightGBM	MLR	0.351	0.092	0.172	0.420	0.079	0.215
LightGBM	Isotonic	0.359	0.046	0.152	0.382	0.049	0.172
LightGBM	Dirichlet	0.247	0.043	0.134	0.312	0.057	0.170
CatBoost	Raw	0.365	0.118	0.186	0.395	0.147	0.217
CatBoost	Temp	0.251	0.040	0.150	0.274	0.045	0.159
CatBoost	MLR	0.398	0.085	0.199	0.430	0.083	0.217
CatBoost	Isotonic	0.336	0.042	0.147	0.361	0.054	0.162
CatBoost	Dirichlet	0.243	0.031	0.143	0.286	0.042	0.160
SingleMLP	Raw	0.254	0.070	0.140	0.425	0.098	0.211
SingleMLP	Temp	0.235	0.043	0.136	0.372	0.050	0.192
SingleMLP	MLR	0.320	0.097	0.159	0.437	0.085	0.224
SingleMLP	Isotonic	0.286	0.029	0.137	0.380	0.039	0.174
SingleMLP	Dirichlet	0.245	0.037	0.146	0.320	0.065	0.175
Ens. $M = 5$	Raw	0.238	0.065	0.132	0.406	0.093	0.195
Ens. $M = 5$	Temp	0.217	0.045	0.128	0.358	0.051	0.175
Ens. $M = 5$	MLR	0.311	0.093	0.154	0.413	0.094	0.207
Ens. $M = 5$	Isotonic	0.331	0.050	0.145	0.555	0.045	0.176
Ens. $M = 5$	Dirichlet	0.226	0.041	0.133	0.307	0.048	0.155

Within this benchmark, the calibrator preference therefore depends on the perturbation setting. Temperature scaling and Dirichlet calibration remained the strongest calibrators under label noise and under any perturbation leaving the test feature distribution intact, while under the synthetic covariate-shift protocol tested here MLR was the more robust choice on the evaluated neural models despite its in-distribution failure mode.

5.3. Stratification by Class-Imbalance Ratio

To address whether the calibrator orderings of Section 4.3 hold under class imbalance, we stratify the 36 datasets by the ratio of the largest to the smallest class size, computed from the full available labels. Three buckets emerge: *low* ($\leq$2, 20 datasets), *medium* (2–5, 5 datasets), and *high* (>5, 11 datasets). Median post-calibration NLL per model and calibrator is recomputed within each bucket. Table 9 reports the most diagnostic cells.

Three patterns hold within this benchmark. First, in the low imbalance bucket the in-distribution ordering of Section 4.3 is preserved: Dirichlet is the strongest calibrator on every GBDT evaluated, temperature scaling is strongest on the deep ensemble, and MLR is the weakest in every (bucket, model) cell. In the medium bucket, the GBDT preference narrows: temperature scaling and Dirichlet calibration are essentially tied, with temperature scaling marginally ahead on all three GBDTs by less than 0.01 NLL, while MLR remains weakest. Second, in the high-imbalance bucket the GBDT ordering reverses: relative to raw probabilities, Dirichlet calibration degrades LightGBM NLL by 28%, XGBoost by 37%, and CatBoost by 36%, while temperature scaling remains roughly neutral on all three GBDTs. The deep ensemble shows a smaller but directionally consistent shift. Third, MLR remains weakest in every bucket, confirming that its in-distribution failure is not driven by imbalance.

Table 9. Median NLL by imbalance bucket for four representative models ($n = 36$ datasets stratified into 20/5/11 by max/min class-size ratio $\leq$2, 2–5, >5). Bold marks the best calibrator within each (bucket, model) row. The Dirichlet-on-GBDT advantage observed in Section 4.3 holds in the low and medium buckets but reverses at high imbalance, where raw or temperature-scaled probabilities are best.

Bucket	Model	Raw	Temp	Dirichlet	MLR
Low ($\leq$2)	LightGBM	0.2447	0.2406	**0.2383**	0.2746
	XGBoost	0.2576	0.2504	**0.2490**	0.2828
	CatBoost	0.3141	0.3081	**0.2744**	0.3229
	Ens. $M = 5$	0.2581	**0.2498**	0.2662	0.3292
Medium (2–5)	LightGBM	0.3843	**0.3838**	0.3844	0.4130
	XGBoost	0.4131	**0.4039**	0.4061	0.4176
	CatBoost	0.3879	**0.3781**	0.3867	0.4096
	Ens. $M = 5$	0.3742	0.3744	**0.3715**	0.4017
High (>5)	LightGBM	**0.0893**	0.0881	0.1146	0.1205
	XGBoost	0.0841	**0.0839**	0.1156	0.1395
	CatBoost	**0.0866**	0.0875	0.1176	0.1281
	Ens. $M = 5$	**0.1201**	0.1225	0.1266	0.1810

A mechanistic interpretation matches the small-calibration-set diagnosis unified in Section 8.2: under heavy class imbalance, $X_{\text{val_cal}}$ contains very few minority-class examples, so the $K \times K$ Dirichlet transformation cannot be estimated reliably when one or more of its rows are anchored by a handful of observations, while the single-parameter temperature transformation degrades gracefully. Raw GBDT probabilities, already sharply peaked toward the majority class, are not improved by any flexible re-mapping. The practical implication is that the Dirichlet-on-GBDT recommendation of Section 4.3 should be guarded by an imbalance check: at max/min class-size ratios above approximately five, raw or temperature-scaled GBDT probabilities are at least as good as, and often better than, Dirichlet-calibrated ones.

6. Results: Split-Conformal Coverage at 90% Target

For each (dataset, model, seed), we fit a split-conformal predictor on $X_{\text{val_cal}}$ using adaptive prediction sets (APS) [13] at target coverage $1 - \alpha = 0.90$, then evaluate marginal empirical coverage and average prediction-set size on X_{test}. The conformal evaluation uses $M = 5$ as the ensemble representative; $M = 3$ and $M = 10$ are omitted. Table 10 reports per-model summaries across the 36 datasets. All six models achieve nominal marginal coverage at or above 0.90 on the median dataset, with median empirical coverage equal to 1.0 across the benchmark (reflecting that the calibration procedure is conservative when test-set difficulty closely matches $X_{\text{val_cal}}$). The key differentiator is set-size efficiency. CatBoost produces the sharpest median set size (2.47), is the sharpest model on 19 of 36 datasets, and dominates the deep ensemble (2.79), LightGBM (2.79), MC-Dropout (2.78), XGBoost (2.83), and SingleMLP (2.83). The minimum empirical coverage across datasets is 0.958 for CatBoost and $\geq$0.967 for all other models, well within the standard sample-size noise band around the 0.90 target.

The set-size advantage of CatBoost on the conformal track is consistent with its raw probabilities being the most miscalibrated of any model in the benchmark (raw ECE = 0.039 in Figure 2), yet also the sharpest after temperature scaling: post-calibration sharpness translates directly to sharper conformal sets at fixed coverage. Marginal-coverage validity is a property of the conformal procedure (achieved by all models), while set-size efficiency is a property of the underlying classifier. Within this benchmark, classifiers with sharp, well-separated raw scores such as CatBoost produced smaller prediction sets at fixed coverage,

suggesting that set-size efficiency is worth considering alongside marginal coverage when selecting a model for conformal deployment rather than treating any model that meets the coverage target as equivalent.

Table 10. Split-conformal results at 90% target coverage (APS, $n = 36$ datasets, mean across 5 seeds then summary across datasets). All models achieve nominal marginal coverage. CatBoost produces the sharpest median prediction sets.

Model	Median Empirical Coverage	Median Avg Set Size	Sharpest on (of 36)
CatBoost	1.000	2.467	19
DeepEns. $M = 5$	1.000	2.786	5
MC-Dropout	1.000	2.778	4
XGBoost	1.000	2.826	4
LightGBM	1.000	2.788	3
SingleMLP	1.000	2.829	1

7. Results: TabPFN v2 on Its Compatible Scope

TabPFN v2 [14] is evaluated on the 24 of 36 datasets meeting its inference constraints ($\leq$10,000 training samples, $\leq$500 features, $\leq$10 classes). Using this restricted scope, we compute the same five calibration methods and the same three uncertainty metrics, with TabPFN trained once per dataset and seed by in-context inference. Table 11 reports the resulting calibration metrics. Because TabPFN's seed-level variance is dominated by occasional in-context-inference fluctuations on a small number of datasets, the table reports per-dataset means across seeds. The median-then-median protocol of Section 3.6 produces qualitatively identical calibrator orderings (raw NLL = 0.086, isotonic NLL = 0.189, other cells within $\pm$0.005). Two findings dominate. First, TabPFN v2 raw probabilities are already extremely well-calibrated, with median NLL = 0.099, ECE = 0.030, Brier = 0.049, and accuracy = 0.969. These results are numerically better than those of any classical model on the same 24-dataset subset, with CatBoost post Dirichlet calibration—the best classical baseline on this subset—achieving 0.122 NLL for reference. Temperature scaling provides a marginal improvement on NLL (0.098 vs. 0.099), consistent with TabPFN's in-context-learning outputs being already well-calibrated on this subset.

Table 11. TabPFN v2 calibration on its 24-dataset compatible scope. Raw and temperature-scaled outputs are best on NLL and Brier. Dirichlet and isotonic minimise ECE but at large NLL cost. The isotonic trap is severe here.

Calibrator (TabPFN, 24 ds)	NLL ↓	ECE ↓	Brier ↓	Acc.
None (raw)	0.099	0.030	**0.049**	**0.969**
Temperature scaling	**0.098**	0.030	0.050	0.969
Dirichlet	0.100	**0.027**	0.056	0.965
MLR	0.138	0.062	0.063	0.962
Isotonic	0.246	0.028	0.062	0.961

Second, the calibration ordering on TabPFN differs from the classical-model ordering. Dirichlet moderately worsens NLL on TabPFN (0.100 vs. 0.099 raw), unlike its consistent improvement on GBDTs. MLR significantly worsens NLL (0.138) and ECE (0.062), and isotonic catastrophically degrades NLL (0.246). This pattern is consistent with the diagnosis that calibrator effectiveness scales inversely with raw calibration quality: when raw outputs are already well-calibrated, fitting any parametric or non-parametric calibrator on a finite $X_{\text{val_cal}}$ introduces estimation noise that exceeds the residual miscalibration being corrected. For TabPFN v2 on its 24-dataset compatible subset, raw outputs or a single-parameter temperature correction were the safest defaults among the five calibrators tested. Whether

this generalises to other tabular foundation models or to TabPFN v2 outside its compatible scope is an open question.

7.1. Training Time

Training time (seconds) is logged for every (model, dataset, seed) run. Table 12 summarises per-run wall-clock training time across the thirty-six datasets and five seeds, giving $n = 180$ runs per model. Medians represent dataset-typical runtime, whereas means are strongly affected by the largest datasets and therefore better reflect aggregate batch-experiment cost. Three observations follow. First, the lightest GBDTs are faster than the neural models in median terms: LightGBM (0.36 s) and XGBoost (0.48 s) outpace SingleMLP (1.92 s) and MC-Dropout (1.98 s), while CatBoost remains in the low-second range (1.14 s). Second, deep ensembles scale approximately linearly with ensemble size: the median runtime increases from 6.15 s at $M = 3$ to 10.18 s at $M = 5$ and 19.58 s at $M = 10$. By median runtime, the $M = 10$ ensemble is approximately 54× slower than LightGBM. Third, mean runtimes are much larger and more compressed across models because they are dominated by the largest datasets; for example, LightGBM has a median runtime below one second but a mean runtime of 41.05 s.

Table 12. Per-run training time across 36 datasets and five seeds ($n = 180$ runs per model). Medians represent dataset-typical runtime, while means are affected by large-dataset outliers and reflect aggregate wall-clock cost in batch experimentation.

Model	Median (s)	Mean (s)
LightGBM	0.36	41.05
XGBoost	0.48	73.50
CatBoost	1.14	38.52
SingleMLP	1.92	9.80
MC-Dropout	1.98	9.90
DeepEns. $M = 3$	6.15	30.34
DeepEns. $M = 5$	10.18	49.35
DeepEns. $M = 10$	19.58	98.88

7.2. Cost-Weighted Composite Calibration Score

Single-metric reporting can mislead practitioners when calibrators trade off NLL against ECE (the isotonic trap of Section 4.4). To surface the joint trade-off, we define a min–max-normalised composite cost C_w computed per dataset and then aggregated,

$$C_w(\text{model}, \text{cal}) = w_1 \cdot \text{NLL}_{\text{norm}} + w_2 \cdot \text{ECE}_{\text{norm}} + w_3 \cdot \text{Brier}_{\text{norm}}, \tag{7}$$

where each metric is min–max-scaled per dataset across all (model, calibrator) cells so that the best cell receives 0 and the worst receives 1, and the weights satisfy $w_1 + w_2 + w_3 = 1$. Table 13 reports the best cell under three weight profiles that correspond to common deployment intents: NLL-dominant (likelihood-driven decisions such as cost-sensitive risk scoring), ECE-dominant (threshold-policy applications where confidence–accuracy alignment matters most), and balanced. Dirichlet-calibrated LightGBM wins all three in-distribution weight profiles. Under covariate shift, MLR is the dominant calibrator across all profiles: the $M = 3$ and $M = 10$ deep ensembles win the NLL-dominant and balanced cells at $\sigma = 1$ and $\sigma = 2$, respectively, and CatBoost wins the ECE-dominant cell at $\sigma = 2$. The composite makes explicit a recommendation otherwise implicit across the tables: there is no single "best" cell, and weight profiles should be set from deployment economics rather than left implicit.

Table 13. Best (model, calibrator) cell under three weight profiles for the composite C_w (Equation (7)). The preferred cell changes under both the NLL/ECE/Brier weighting and the evaluation setting. In-distribution values computed from Figure 2, and shifted values from the Section 5.1 medians.

Profile $(w_{\text{NLL}}, w_{\text{ECE}}, w_{\text{Brier}})$	In-distribution	At $\sigma = 1$	At $\sigma = 2$
NLL-dominant $(0.6, 0.2, 0.2)$	LightGBM/Dirichlet	Ens. $M = 3$/MLR	Ens. $M = 10$/MLR
Balanced $(0.34, 0.33, 0.33)$	LightGBM/Dirichlet	Ens. $M = 3$/MLR	Ens. $M = 10$/MLR
ECE-dominant $(0.2, 0.6, 0.2)$	LightGBM/Dirichlet	Ens. $M = 3$/MLR	CatBoost/MLR

Beyond the three weight profiles of Table 13, we additionally sweep the full (w_1, w_2, w_3) simplex in 0.1 increments (66 weight triples per model) and record which calibrator minimises C_w for each triple. Table 14 reports the per-model dominance fractions. The architecture-by-calibrator interaction holds across essentially the entire simplex: CatBoost selects Dirichlet on 98.5% of all weight triples, and every neural model selects temperature scaling (or its member-level variant) on at least 94%. XGBoost is the only model with a non-trivial isotonic-regression region (16.7% of the simplex), reflecting that the isotonic-trap trade-off is steeper for XGBoost than for the other GBDTs.

Table 14. Per-model calibrator dominance across the in-distribution weight simplex. For each model, entries report the fraction of the 66 $(w_{\text{NLL}}, w_{\text{ECE}}, w_{\text{Brier}})$ triples on the 0.1-step simplex on which each calibrator minimises the composite C_w. The temperature-scaling column for ensembles aggregates standard and member-level variants. Bold marks the dominant calibrator per row.

Model	Dirichlet	Temp/Temp-Member	Isotonic	Raw
CatBoost	**98.5%**	0.0%	1.5%	0.0%
LightGBM	**68.2%**	31.8%	0.0%	0.0%
XGBoost	4.5%	**78.8%**	16.7%	0.0%
SingleMLP	0.0%	**97.0%**	1.5%	1.5%
MC-Dropout	0.0%	**98.5%**	1.5%	0.0%
DeepEns. $M = 3$	0.0%	**94.0%**	6.1%	0.0%
DeepEns. $M = 5$	0.0%	**96.9%**	3.0%	0.0%
DeepEns. $M = 10$	0.0%	**98.5%**	1.5%	0.0%

8. Discussion, Limitations, and Conclusions

8.1. Synthesis

Five findings summarise the benchmark.

1. Within the evaluated in-distribution setting, we do not detect a statistically significant uncertainty-quality advantage of deep ensembles over calibrated GBDTs on the primary metrics. The stronger performance of ensembles over the fixed SingleMLP appears partly related to the fixed MLP hyperparameter configuration, as the tuning ablation narrows this gap.
2. Within this benchmark, calibration effects are model-family dependent. Dirichlet calibration is most favourable for the evaluated GBDTs, whereas temperature scaling is most favourable for the evaluated neural models.
3. The in-distribution rankings observed here do not transfer unchanged to the perturbation settings tested. Under the synthetic Gaussian covariate-shift protocol, multinomial logistic recalibration becomes more competitive for the evaluated neural models despite its poor in-distribution behaviour, and under high class imbalance the advantage of Dirichlet calibration for the evaluated GBDTs weakens. Whether the same reversal would hold under naturalistic distribution shift is not tested here.
4. In the split-conformal track, conformal prediction separates coverage validity from set-size efficiency. Several evaluated models achieve nominal marginal coverage, but CatBoost produces the smallest prediction sets in this benchmark.

5. TabPFN v2 performs strongly on its compatible subset, although the tested post-hoc calibrators do not improve its primary uncertainty metrics on that subset.

Taken together, within the scope of the evaluated protocol, uncertainty quality does not appear to be determined by model architecture alone. It depends on the interaction between model family, calibration method, class balance, perturbation setting, and the metric trade-off considered by the practitioner.

Several practical implications follow from this pattern. Within TabPFN v2's compatible scope, raw probabilities were competitive with calibrated alternatives, so post-hoc calibration provided little benefit under those constraints in this benchmark. Outside that scope, Dirichlet calibration on a GBDT minimised the in-distribution composite, with the caveat that this advantage weakens above approximately 5:1 class imbalance, where temperature scaling becomes the more robust choice. For conformal deployment on this benchmark, CatBoost produced the sharpest prediction sets at fixed coverage. Under tight compute budgets, LightGBM was statistically indistinguishable from the $M = 10$ deep ensemble on the primary uncertainty metrics at roughly $1/60$ the wall-clock cost, and did not yield a statistically detectable uncertainty advantage in this benchmark.

8.2. A Unifying Small-Calibration-Set Mechanism

The two reversals identified above share a single underlying mechanism. Both temperature scaling (one parameter) and Dirichlet calibration (a $K \times K$ affine transformation) are fitted on $X_{\text{val_cal}}$, but they differ in how much support they require: the single-parameter scalar transformation is well-estimated from a handful of examples per class, whereas the $K \times K$ Dirichlet transformation has $K(K+1)$ free parameters and needs adequate coverage of every off-diagonal entry. Under heavy covariate shift, the in-distribution calibration set is no longer representative of test-time inputs, so the additional flexibility of Dirichlet becomes a liability. MLR's even richer transformation can compensate by pulling aggressively toward the marginal class distribution. Under high class imbalance, the calibration set itself contains too few minority-class examples to fit any flexible transformation reliably. In both cases, the single-parameter temperature transformation degrades gracefully, whereas higher-capacity calibrators degrade sharply, explaining why the reversals are largest for neural models (extreme raw softmaxes) and largest for GBDTs at high imbalance (sharp majority-class peaks).

8.3. Limitations

Five limitations bound the scope of these conclusions. First, although Section 5.3 identifies a reversal at imbalance >5, the high-imbalance bucket relies on 11 datasets, so the precise threshold should be regarded as qualitative. The OpenML-CC18 suite also lacks datasets with sparse features or temporal structure, so robustness rankings under those conditions remain to be tested. Second, the covariate-shift analysis relies on additive isotropic Gaussian noise applied to standardised features and is a controlled stress test, not a realistic out-of-distribution evaluation. Naturalistic distribution shifts (population or sub-population shift, temporal drift, sensor- or instrument-dependent noise) typically exhibit between-feature dependencies, class-conditional structure, and changes in the marginal class distribution that the Gaussian protocol does not reproduce. The MLR-over-temperature-scaling reversal reported here should therefore be read as a directional indicator under synthetic shift, not as a deployment recommendation for any specific real-world OOD setting. Replicating it under benchmarks with naturalistic shift, such as temporal drift in financial or clinical tabular data or class-conditional sub-population shifts, would be a necessary precondition before MLR could be recommended as a production default. Third, the label-noise analysis uses uniform random-label replacement on a six-

dataset subset, and class-conditional and instance-dependent noise may behave differently. Fourth, the main benchmark uses a single fixed hyperparameter configuration per model class across all 36 datasets, chosen from related-work defaults. This places all model–calibrator combinations on equal footing for this fixed-protocol comparison, but does not represent the strongest possible version of any individual model. The MLP ablation in Section 4.7 uses oracle selection on $X_{\text{val_cal}}$ and therefore represents an upper bound on what nested cross-validation could realistically deliver. Deployable tuning would close the SingleMLP–ensemble gap less completely. We did not run equivalent oracle-tuning sweeps for the three GBDT families, so the headline GBDT-vs-ensemble comparison should be read as a comparison at recommended defaults rather than at tuned optima. Whether per-dataset GBDT tuning would also shift the architecture-by-calibrator interaction is untested here. Fifth, this study uses MC-Dropout as a lightweight Bayesian approximation and does not evaluate full variational Bayesian neural networks, evidential deep learning, ensembles of GBDTs, or hybrid stacked architectures. Whether those families exhibit the same calibrator-by-architecture interaction is an open question.

8.4. Conclusions

This study provides a fixed-protocol empirical comparison of tabular uncertainty-estimation strategies across model families, calibration methods, ensemble sizes, controlled perturbations, conformal coverage, and computational cost. Within this benchmark, no single architecture or calibrator dominates: post-hoc calibration can substantially alter uncertainty-quality rankings, and the preferred calibrator depends on both the model family and the expected evaluation setting. Under controlled covariate shift and high class imbalance, the in-distribution preferences change, highlighting the risk of selecting calibration methods using only clean in-distribution validation results. These findings support reporting multiple uncertainty metrics, computational costs, and explicit NLL/ECE/Brier trade-offs rather than relying on a single calibration score or model-family comparison.

Author Contributions: Conceptualization, D.Š.; methodology, D.Š.; software, D.Š.; validation, D.Š., T.K., Z.K. and I.L.; formal analysis, D.Š.; investigation, D.Š.; writing—original draft preparation, D.Š.; writing—review and editing, D.Š., T.K., Z.K. and I.L.; visualization, D.Š. and T.K.; supervision, Z.K. and I.L.; funding acquisition, I.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded and is a part of research for the activities to be carried out for the project "Researching advanced algorithms and innovative business intelligence solutions in the cloud—NPOO.C3.2.R3-I1.04.0128".

Institutional Review Board Statement: Not applicable.

Data Availability Statement: All experimental results, code, configuration files, and dataset manifests are available at https://github.com/dinasabanovic/calibration-uncertainty-benchmark (accessed on 25 June 2026). The OpenML-CC18 benchmark datasets are publicly available via the OpenML platform (https://www.openml.org, study ID 99). The complete dataset manifest is provided in the project repository as `dataset_manifest.csv` and `dataset_manifest.json`.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; Dean, J. A guide to deep learning in healthcare. *Nat. Med.* **2019**, *25*, 24–29. [CrossRef] [PubMed]
2. Kompa, B.; Snoek, J.; Beam, A.L. Second opinion needed: Communicating uncertainty in medical machine learning. *npj Digit. Med.* **2021**, *4*, 4. [CrossRef] [PubMed]

3. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017; pp. 1321–1330.
4. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete problems in AI safety. *arXiv* **2016**, arXiv:1606.06565. [CrossRef]
5. Niculescu-Mizil, A.; Caruana, R. Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning (ICML), Bonn, Germany, 6–11 August 2005; pp. 625–632. [CrossRef]
6. Shwartz-Ziv, R.; Armon, A. Tabular data: Deep learning is not all you need. *arXiv* **2021**, arXiv:2106.03253. [CrossRef]
7. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why tree-based models still outperform deep learning on tabular data. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 507–520. [CrossRef]
8. McElfresh, D.; Khandagale, S.; Valverde, J.; Vishak Prasad, C.; Ramakrishnan, G.; Goldblum, M.; White, C. When do neural nets outperform boosted trees on tabular data? *arXiv* **2023**, arXiv:2305.02997. [CrossRef]
9. Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.V.; Lakshminarayanan, B.; Snoek, J. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *arXiv* **2019**, arXiv:1906.02530. [CrossRef]
10. Kull, M.; Perello-Nieto, M.; Kängsepp, M.; Filho, T.S.; Song, H.; Flach, P. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. *arXiv* **2019**, arXiv:1910.12656. [CrossRef]
11. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (ICML), New York, NY, USA, 19–24 June 2016; pp. 1050–1059.
12. Vovk, V.; Gammerman, A.; Shafer, G. *Algorithmic Learning in a Random World*; Springer: New York, NY, USA, 2005. [CrossRef]
13. Romano, Y.; Sesia, M.; Candès, E.J. Classification with valid and adaptive coverage. *arXiv* **2020**, arXiv:2006.02544.. [CrossRef]
14. Hollmann, N.; Müller, S.; Purucker, L.; Krishnakumar, A.; Körfer, M.; Hoo, S.B.; Schirrmeister, R.T.; Hutter, F. Accurate predictions on small data with a tabular foundation model. *Nature* **2025**, *637*, 319–326. [CrossRef] [PubMed]
15. Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; Wierstra, D. Weight uncertainty in neural networks. In Proceedings of the 32nd International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015; pp. 1613–1622.
16. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
17. Benavoli, A.; Corani, G.; Demšar, J.; Zaffalon, M. Time for a change: A tutorial for comparing multiple classifiers through Bayesian analysis. *J. Mach. Learn. Res.* **2017**, *18*, 1–36.
18. Javed, A.; Javaid, N.; Asim, Y.; Ali, Z.; Alsamani, B.; Saudagar, A.K.J. An AI based framework using Aquila-optimized capsule network for radar object detection. *Complex Intell. Syst.* **2026**, *12*, 103. [CrossRef]
19. Bischl, B.; Casalicchio, G.; Feurer, M.; Gijsbers, P.; Hutter, F.; Lang, M.; Mantovani, R.G.; van Rijn, J.N.; Vanschoren, J. OpenML benchmarking suites. *arXiv* **2017**, arXiv:1708.03731. [CrossRef]
20. Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.* **1999**, *10*, 61–74.
21. Kadra, A.; Lindauer, M.; Hutter, F.; Grabocka, J. Well-tuned simple nets excel on tabular datasets. *arXiv* **2021**, arXiv:2106.11189. [CrossRef]
22. Zadrozny, B.; Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, 23–26 July 2002; pp. 694–699. [CrossRef]
23. Nixon, J.; Dusenberry, M.W.; Zhang, L.; Jerfel, G.; Tran, D. Measuring calibration in deep learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Long Beach, CA, USA, 15–20 June 2019; pp. 38–41. [CrossRef]
24. Kumar, A.; Liang, P.S.; Ma, T. Verified uncertainty calibration. *arXiv* **2019**, arXiv:1909.10155. [CrossRef]
25. Fredsgaard, L.; Schmidt, M.N. On joint regularization and calibration in deep ensembles. *arXiv* **2025**, arXiv:2511.04160. [CrossRef]
26. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv* **2017**, arXiv:1612.01474. [CrossRef]
27. Fort, S.; Hu, H.; Lakshminarayanan, B. Deep ensembles: A loss landscape perspective. *arXiv* **2019**, arXiv:1912.02757. [CrossRef]
28. Wilson, A.G.; Izmailov, P. Bayesian deep learning and a probabilistic perspective of generalization. *arXiv* **2020**, arXiv:2002.08791. [CrossRef]
29. Gorishniy, Y.; Rubachev, I.; Babenko, A. TabM: Advancing tabular deep learning with parameter-efficient ensembling. In Proceedings of the International Conference on Learning Representations (ICLR), Singapore, 24–28 April 2025.
30. Holzmüller, D.; Grinsztajn, L.; Steinwart, I. Better by default: Strong pre-tuned MLPs and boosted trees on tabular data. *arXiv* **2024**, arXiv:2407.04491. [CrossRef]
31. Wilcoxon, F. Individual comparisons by ranking methods. *Biom. Bull.* **1945**, *1*, 80–83. [CrossRef]
32. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
33. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

34. Gorishniy, Y.; Rubachev, I.; Khrulkov, V.; Babenko, A. Revisiting deep learning models for tabular data. *arXiv* **2021**, arXiv:2106.11959. [CrossRef]
35. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3146–3154.
36. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
37. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. *arXiv* **2018**, arXiv:1706.09516. [CrossRef]
38. Tankaria, H.; Sugimoto, S.; Yamashita, N. A regularized limited memory BFGS method for large-scale unconstrained optimization and its efficient implementations. *arXiv* **2021**, arXiv:2101.04413. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.