QAC: DATA QUALITY-AWARE UNCERTAINTY CALIBRATION FOR TRUSTWORTHY AI

Anonymous authors
Paper under double-blind review

ABSTRACT

A classifier's predicted probability is its statement of uncertainty about the label, and calibration makes this statement match observed frequencies. Post-hoc calibrators use only the model's output, so two samples with the same output receive the same probability. Consider two patients whose laboratory values are almost the same, but one patient's value is wrong because the blood sample was stored too long before analysis: the model gives both the same prediction and the same confidence, although the second prediction is more likely to be wrong. Laboratories and other data pipelines often record such problems. We call a per-sample value that describes how reliable a data point's input or label is its data quality. When it is not available at training time, it cannot be added to the model's input, but it can be given to the calibrator. A standard result implies that, for the best possible predictor, a data-quality value can improve the prediction only if it carries information about the label that the input does not. We propose *Quality-Aware Uncertainty Calibration* (QAC), a second calibration stage that can follow any existing calibrator. For each class, it estimates on the calibration set how often the class is correct given the first-stage probability and the quality score; the direction of the effect of quality is learned, not assumed. On seven datasets, nine base calibrators and ten seeds, and compared with the same stage given a constant quality score, QAC lowers the Brier score by 0.001–0.016 on the four datasets where quality is informative (a simulation and three datasets where quality is annotator agreement). It does not help on the two datasets whose quality score is computed from the input, and gives no significant gain on a clinical dataset. The gain is mainly in separating correct from wrong predictions (error-detection AUROC $+0.021$ to $+0.059$), while the separation between classes improves about a quarter as much or less.

1 INTRODUCTION

Consider a model that predicts a disease from patient data, and two patients whose data are very similar, so that the model gives them the same output. Suppose now that one patient's laboratory value is wrong, for example because the blood sample was stored too long before analysis, or because the analyser had failed its quality-control check. The wrong value still lies in the normal range, so the two records look alike and the model cannot tell which one is corrupted. It classifies both patients the same way and is equally confident about both predictions. Ideally, it would be less confident about the patient whose data are corrupted.

Suppose we can characterise such problems by a value recorded for each data point, such as the time the sample waited or the result of the instrument check. We call this value the *data quality* of the data point, or its quality score, and write it q. It describes how reliable the recorded input or its label is. It can describe the input (*measurement quality*), as in this example, or the label (*label ambiguity*), for example how much annotators agreed on it in a crowdsourced dataset (Peterson et al., 2019; Nie et al., 2020). Clinical data pipelines record many such checks (Kahn et al., 2016; Weiskopf & Weng, 2013; Hripcsak et al., 2015). We would like the model's confidence to use q.

When q is available at training time, it can be added to the model's input. This is not always feasible. Often q is not available when the model is trained but is available for held-out labelled data and at inference, because it is defined after training or because the model is pretrained and cannot be retrained. We therefore propose to use q in post-hoc calibration. A post-hoc calibrator

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

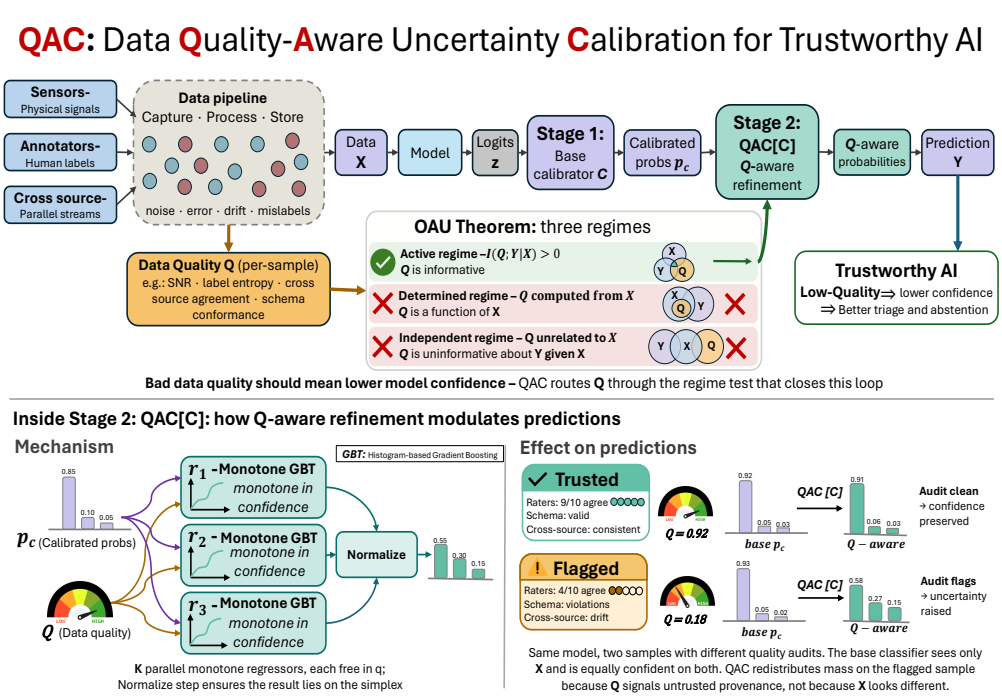

Figure 1: Overview of QAC. *Top*: the data pipeline produces the input X and a per-sample data-quality score Q. The model sees only X; a base calibrator C (Stage 1) maps its logits z to probabilities p_c, and QAC (Stage 2) refines p_c using Q. Q can help only in the active case, where it carries information about the label Y beyond X (Section 2). *Bottom*: Stage 2 fits one gradient-boosted regressor per class, monotone in p_c and free in Q, and normalises the outputs (Section 3); two samples with the same base probabilities but different quality receive different outputs.

is fitted after training on a held-out labelled set, the calibration set, and maps the model's output to probabilities such that, among samples assigned probability 0.9, about 90% are correct (Guo et al., 2017; Niculescu-Mizil & Caruana, 2005). It does so by counting, at each level of the model's confidence, how often the model was correct. Because it sees only the model's output, it pools clean and corrupted records, and the two patients above receive the same probability.

We propose **Quality-Aware Uncertainty Calibration** (QAC), a second calibration stage that takes the probabilities of any existing post-hoc calibrator, which we call the base calibrator, together with q (Figure 1, Section 3). QAC makes the same count separately for each level of quality: if records of low quality were wrong more often in the calibration set, their confidence is lowered. A second stage can change the results even without q, simply by calibrating twice, so we compare QAC with the same stage given a constant q.

We also study which values can serve as data quality, because not every value helps. Aleatoric uncertainty is the uncertainty about the label that remains even for the best model that sees the input. Observing q removes the part of it about which q carries information beyond the input, which we call the *observable* part; the rest is the *intrinsic* part (Section 2). So q helps only if it tells us something about the label that the input does not. A value computed from the input, such as the noise level of an audio clip, adds nothing that a good model cannot learn itself. Moreover, a better quality score does not always mean fewer errors. In an intensive-care dataset where q has several components (eICU, Section 4), three components go with slightly more errors when their score is better (Section 3), so QAC estimates the direction of the effect of q from data instead of assuming it. We also show how to check before use whether a given q is informative (Section 4.2).

The components of QAC are not new: the split of aleatoric uncertainty follows from the chain rule of entropy (Cover et al., 2006), and calibration with quality features exists in speaker verification (Thienpondt et al., 2021) and online advertising (Pan et al., 2020; Wei et al., 2022), where the features are

computed from the input. The closest work (Cheng et al., 2025) fixes the direction of the effect of its quality index; QAC estimates it.

Contributions. The rest of the paper follows the questions above in order.

- *When quality can help* (Section 2). We define data quality as a per-sample value outside the model input and use the split of aleatoric uncertainty into intrinsic and observable parts to state when such a value can improve a calibrated prediction, and why.
- *How to use it* (Section 3). We propose QAC, a second calibration stage for any post-hoc calibrator that estimates the effect of quality from data, and a constant-q control that separates this effect from that of calibrating twice.
- *Evidence* (Section 4). We test QAC on seven datasets with nine base calibrators and ten seeds. The quality score is informative on four of them, and on these QAC lowers the Brier score relative to the constant-q control, mainly by improving error detection. On the two datasets whose quality score is computed from the input it does not help, and on eICU its gain is not significant. The direction of a quality effect cannot be inferred from its name.

2 When quality can help

Setup. We write X for the model input, $Y \in \{1, \dots, K\}$ for the label and Q for the quality score as random variables, and x, y, q for their observed values. C is a post-hoc calibrator of the model's outputs, and $\mathrm{QAC}[C]$ denotes any second calibration stage that takes $(C(X), Q)$ as input; Section 3 gives our construction.

2.1 Which quality scores can help?

Return to the two patients. The model sees their measurements X and must predict the disease Y. Even the best possible model is uncertain when the measurements are compatible with both outcomes. This remaining uncertainty is the aleatoric uncertainty, usually measured by the conditional entropy $H(Y \mid X)$ (Depeweg et al., 2018; Hüllermeier & Waegeman, 2021). It depends on what is measured (Der Kiureghian & Ditlevsen, 2009; Smith et al., 2024): if we also observe the data quality Q, part of it may disappear. The chain rule of entropy (Cover et al., 2006) splits it exactly:

$$\underbrace{H(Y \mid X)}_{\text{aleatoric given } X} = \underbrace{H(Y \mid X, Q)}_{\textbf{intrinsic}} + \underbrace{I(Q; Y \mid X)}_{\textbf{observable}}.$$

The intrinsic part remains even when both the measurements and their quality are known. The observable part, the conditional mutual information, is what knowing the quality removes. It is the only part of the aleatoric uncertainty that a quality score can reduce.

This split tells us which values can serve as data quality. If Q carries information about Y beyond X, the *active* case, a predictor that uses Q can do better. If Q is a function of X, the *determined* case, for example a noise estimate computed from the recording itself, the best predictor from X already contains it. If Q is not a function of X but says nothing about Y once X is known, the *independent* case, it does not help either (Proposition 6, Appendix B). The record that a sample waited too long before analysis can fall in the active case. A real model is not the best predictor, so a value computed from the input can still help if the model has not learned it; the ΔR^2 of Section 4.2 measures this.

2.2 How much can it help?

Section 2.1 says whether Q can help. To say how much, we need a way to score predicted probabilities. A scoring rule L assigns a loss to a predicted distribution and an observed label. It is *strictly proper* if its expected value is minimised only by the true distribution; the Brier score and the log-loss are examples. The next theorem compares the best predictor that uses X with the best predictor that uses (X, Q).

Theorem 1 (Observable Aleatoric Uncertainty (OAU))**.** *If L is strictly proper and $I(Q; Y \mid X) > 0$, then $\min_g \mathbb{E}[L(g(X, Q), Y)] < \min_f \mathbb{E}[L(f(X), Y)]$. The difference equals $\mathbb{E}_{(X,Q)}[d_\varphi(P(Y \mid X), P(Y \mid X, Q))]$, where $P(Y \mid X)$ and $P(Y \mid X, Q)$ are the conditional distributions of Y given*

X and given (X, Q), and d_φ is the Bregman divergence associated with L (for the log-loss, the Kullback–Leibler divergence).

For the log-loss, this gap equals $I(Q; Y \mid X)$, the observable part of Section 2.1 (Corollary 7, Appendix B).

This inequality is a standard consequence of propriety (Savage, 1971; Gneiting & Raftery, 2007; Bröcker, 2009; Kull & Flach, 2015): conditioning on more information cannot increase the expected loss. We do not claim it as a contribution; we state it because it defines the quantity that QAC targets. Two practical consequences are proved in Appendix B. If Q is available at calibration and inference but not during training, a post-hoc stage is the only way to use it (Corollary 5). If Q is also available during training, a post-hoc stage on $(C(X), Q)$ is as good as a model trained on (X, Q) exactly when $C(X)$ keeps all the information in X about Y given Q; otherwise the difference is the residual grouping loss of C given Q (Corollary 8). For a calibrator that uses Q after C, two further questions matter: how much of its gain can come from recalibrating C alone, and when Q cannot help at all. The next two statements answer them.

Corollary 2 (Reliability–resolution bound)**.** *For any base calibrator C and any $\mathrm{QAC}[C]$ that takes $(C(X), Q)$ as input,*

$$\mathrm{Brier}(C) - \mathrm{Brier}(\mathrm{QAC}[C]) \;\le\; \mathrm{Reliability}(C) + \big(\mathrm{Resolution}(\mathrm{QAC}[C]) - \mathrm{Resolution}(C)\big),$$

where the terms are those of the Bröcker decomposition of the Brier score (Bröcker, 2009), applied class-wise.

Reliability measures how far the predicted probabilities are from the observed frequencies; any recalibration of C can remove it without Q (Niculescu-Mizil & Caruana, 2005). Resolution measures how much the observed frequency varies with the prediction; no per-class recalibration of C can increase it, but Q can.

Corollary 3 (Quality determined by the calibrated score)**.** *If $Q = h(C(X))$ for a measurable function h, every $\mathrm{QAC}[C]$ is a function of $C(X)$ alone and cannot do better than a flexible recalibration of C.*

How quality changes a prediction: a closed-form case. The results above say when and how much quality can help, but not why. Theorem 9 (Appendix B.4) gives a squared-loss analogue of Theorem 1 that can be computed exactly, in the setting of the patient example. Let the true value of a measurement be $V \sim \mathcal{N}(\mu, \tau^2)$, and let the recorded value be $X = V + \sigma\xi$, where ξ is standard normal noise and the noise level σ differs between samples, independently of V. The quality is the precision $Q = 1/\sigma^2$. The best estimate of the true value is

$$\mathbb{E}[V \mid X, Q] \;=\; \frac{\mu/\tau^2 + Q\,X}{1/\tau^2 + Q},$$

a weighted average of the typical value μ and the recorded value X, with more weight on X when its quality is higher. For a corrupted sample, with low Q, the estimate moves toward the typical value. A model that sees only X must give the same estimate to all samples with the same recorded value. Its excess loss is exactly the gap of Theorem 1, $\mathbb{E}_X[\mathrm{Var}_{Q|X}(\mathbb{E}[V \mid X, Q])]$: the variance of the best estimate across samples with the same recorded value but different quality. The same holds for the binary label $\mathbf{1}\{V > 0\}$, where low quality moves the predicted probability toward the base rate. On simulated data, post-hoc QAC reaches this gap to four decimal places at $n = 10^5$.

3 How QAC works

Idea. Section 2 describes the best possible predictor; this section describes what a calibrator can do in practice. We have a trained classifier f that maps an input x to logits $z = f(x)$, and we do not retrain it. A standard calibrator counts, on held-out data, how often f is correct at each confidence, and so pools the clean and the corrupted patient records. QAC makes the same count separately for each level of quality, which is possible because q is recorded for every calibration sample.

Definition. Let C be a post-hoc calibrator that maps logits z to class probabilities. We use nine: temperature scaling (TS), isotonic regression (Iso), Bayesian binning into quantiles (BBQ), Platt

scaling, beta calibration, histogram binning, spline calibration, Dirichlet calibration (Dir) and top-versus-all isotonic calibration (TvA-iso) (Guo et al., 2017; Niculescu-Mizil & Caruana, 2005; Naeini et al., 2015; Platt et al., 1999; Kull et al., 2017; Zadrozny & Elkan, 2001; Gupta et al., 2020; Kull et al., 2019; Le Coz et al., 2024). The calibration set is $\{(z_i, y_i, q_i)\}_{i=1}^{n}$, where $q_i \in \mathcal{Q}$ is the quality score of sample i, a scalar or a vector. QAC$[C]$ has two stages. *Stage 1* fits C on $\{(z_i, y_i)\}$ and gives $\hat{p}_i := C(z_i)$. *Stage 2* fits one regressor for each class k:

$$r_k = \arg\min_{r \in \mathcal{R}_\uparrow} \sum_{i=1}^{n} \big(r(\hat{p}_{ik}, q_i) - \mathbf{1}\{y_i = k\}\big)^2.$$

Here $\mathcal{R}_\uparrow$ is the set of functions that are non-decreasing in their first argument and unconstrained in q. The prediction for class k is $\text{QAC}[C](z, q)_k = r_k(C(z)_k, q) / \sum_j r_j(C(z)_j, q)$. We fit each r_k with scikit-learn's histogram gradient boosting (Pedregosa et al., 2011), with `monotonic_cst=` $[+1, 0, \ldots, 0]$ and the same hyperparameters on every dataset (Appendix D). Fitting takes less than one second per class for $n \leq 18{,}000$. Both stages are fitted on the same calibration set. QAC does not retrain the model; it uses only the model's logits and q, so a practitioner can keep the calibrator they already use as Stage 1. An earlier one-stage variant and four ways of giving q to the model during training are compared in Appendices H.4 and H.7.

What Stage 2 estimates. Under squared loss, the best predictor of a $0/1$ target is its conditional mean. If $P(Y = k \mid C(X)_k = u, Q = v)$ increases with u and the regressor is flexible enough, r_k therefore approaches

$$r_k(u, v) \;\to\; P\big(Y = k \mid C(X)_k = u,\; Q = v\big),$$

the probability that class k is the true label among samples to which C assigned probability u for class k and whose quality score is v. For the predicted class, this is the probability that the prediction is correct, so it is what we call the relation between quality and error. Histogram binning estimates the same quantity without v. For two classes the final normalisation is exact; for more classes it is an approximation. QAC does not assume that low quality lowers confidence. It lowers confidence only where low-quality calibration samples were wrong more often.

Example: E_toy. The simulated sensor example shows this at work. Each sample is a two-dimensional point from one of two Gaussian classes, measured by two sensors. The classifier receives sensor 1 only, which is corrupted by large noise for 20% of samples, and q measures how closely the two sensors agree; it is close to an indicator of corruption. A corrupted reading looks like an ordinary point, so the corruption is invisible in the input but visible in q: E_toy is in the active case of Section 2.1. Among test samples with a base confidence between 0.85 and 0.95 (mean 0.91), 95% are correct when the sensors agree and 72% when they disagree (Appendix E.1). Stage 2 assigns these two groups 0.96 and 0.74; a calibrator that uses only the model's output must assign both the same value.

Why monotone in $\hat{p}$ but not in q. Two design choices remain: how r_k depends on C's probability, and how it depends on q. At a fixed quality score, a higher Stage-1 probability for class k should not lead to a lower value of r_k. The constraint enforces this, keeps the ranking of samples for class k within a quality level (approximately, after normalisation), and regularises the fit. Removing the constraint and fitting on the full probability vector gives a similar Brier score but a higher top-label calibration error on four of five datasets (Appendix E.3). For q, the direction is not known in advance. In the clinical dataset eICU (Section 4.1), three of the five quality dimensions whose relation to error is stable across seeds go with slightly more errors when their score is better (Appendix E.2). A constraint that lower quality must lower confidence would be wrong for them, so Stage 2 estimates the direction from the data.

Why the refinement is effective. Samples with the same Stage-1 probability can have different chances of being correct. When q predicts this difference, Stage 2 gives them different confidences. This raises the resolution term of the Brier score, which no per-class recalibration of C can do (Corollary 2), and it reorders samples across quality levels by their chance of being correct, which is why error detection improves most (Section 4.4).

4 EXPERIMENTS

4.1 SETUP

What we test. Sections 2 and 3 make four claims, which we test in order. First, a quality score helps only if it carries information about the label beyond the model's input (Section 4.2). Second, when it does, QAC improves calibration, and the gain is due to q rather than to calibrating twice (Section 4.3). Third, the gain comes from ranking predictions by their chance of being correct, so it should appear mainly in error detection (Section 4.4). Fourth, quality does not help when it is computed from the input (Section 4.5). The closed-form case of Section 2.2 and the E_toy example of Section 3 show why the mechanism works; the experiments check that it does on real data.

Datasets and quality scores. We use seven datasets (Appendix C). Four have a measurement-quality score: E_toy (agreement between two sensors, Section 3); eICU, intensive-care stays from 41 hospitals (Pollard et al., 2018), with in-hospital mortality as the label and five record-quality checks per patient as q; Audio MUSAN, spoken keywords mixed with MUSAN noise (Snyder et al., 2015), with the measured signal-to-noise ratio of each clip as q; and Camelyon17, tumour tissue in images from several hospitals (Koh et al., 2021), with an image-quality vector computed from the image as q. Three have a label-ambiguity score derived from several annotations of each test item: ChaosNLI (100 annotations (Nie et al., 2020)), CIFAR-10H (about 50 annotations (Peterson et al., 2019)) and CIFAR-10N (noisy training labels (Wei et al., 2021), with the CIFAR-10H annotations as q). CIFAR-10H and CIFAR-10N share the same annotations and are not independent evidence.

Models. QAC is post-hoc, so each dataset uses a fixed trained classifier: a small multilayer perceptron on E_toy and eICU, DeBERTa-v3-base (He et al., 2021) on ChaosNLI, ResNet-18 (He et al., 2016) on CIFAR-10H and CIFAR-10N, ResNet-50 on Camelyon17, and an Audio Spectrogram Transformer (Gong et al., 2021) fine-tuned on Speech Commands (Warden, 2018) on Audio MUSAN (Appendix C). We use their logits on the calibration and test sets.

Methods and controls. We apply QAC to each of nine base calibrators: temperature scaling (Guo et al., 2017), isotonic regression (Niculescu-Mizil & Caruana, 2005), BBQ (Naeini et al., 2015), Platt scaling (Platt et al., 1999), beta calibration (Kull et al., 2017), histogram binning (Zadrozny & Elkan, 2001), spline calibration (Gupta et al., 2020), Dirichlet calibration (Kull et al., 2019) and top-versus-all isotonic (Le Coz et al., 2024). For each base, we fit Stage 2 three times on the same splits: with the real q, with a constant q ($q \equiv 0$), and with q shuffled across samples. The constant version calibrates a second time without any quality information. The difference between the real and the constant version is therefore the effect of the quality score. The main table and figure show QAC with Dirichlet calibration as Stage 1, written QAC[Dir]; results for every base calibrator are in the appendix.

Metrics. The Brier score, the mean squared difference between the predicted probabilities and the one-hot label, is our main metric; top-label ECE (15 equal-mass bins) and negative log-likelihood are in the appendix. For selective prediction, AURC is the area under the curve of error rate against the fraction of most confident samples retained (Geifman & El-Yaniv, 2017). Error-detection AUROC measures how well the confidence separates correct from incorrect predictions (Hendrycks & Gimpel, 2016), and class-discrimination AUROC how well the probabilities separate the classes. All results use ten seeds with the same splits for every method, and paired tests across seeds with a Bonferroni correction over the four informative datasets.

4.2 IS QUALITY INFORMATIVE BEYOND THE INPUT?

Section 2 says that q can help only if it carries information about the label that the input does not. Table 1 checks this with two measures computed on the calibration set. The first, ρ_q, is the five-fold cross-validated R^2 of a gradient-boosted regressor that predicts the base calibrator's $0/1$ error from q, in the spirit of goodness-of-fit tests that predict residuals with a flexible learner (Janková et al., 2020). The second, ΔR^2, is the increase in this R^2 when q is added to the model's predicted probabilities $\hat{p}$. It tests whether the model's output already contains the information in q. ΔR^2 is the empirical counterpart of the condition that matters for a post-hoc stage, information in q beyond

Table 1: Does q predict the base calibrator's errors? The error is 1 if the calibrator's predicted class is wrong and 0 otherwise. ρ_q: cross-validated R^2 of predicting this error from q alone (negative: worse than predicting its mean). ΔR^2: gain in R^2 when q is added to the model's probabilities, with the number of seeds (of 10) where it is positive. Calibration set, 10-seed means.

dataset	kind of q	ρ_q	ΔR^2 beyond $\hat{p}$	seeds > 0
E_toy	sensor agreement	0.073	+0.046	10
ChaosNLI	annotator entropy	0.069	+0.048	10
CIFAR-10N	annotator entropy	0.102	+0.028	10
CIFAR-10H	annotator agreement	0.055	+0.010	9
eICU	record quality (5-d)	0.014	+0.003	10
Audio MUSAN	measured SNR	0.026	−0.000	6
Camelyon17	image quality (4-d)	−0.078	−0.018	0

Table 2: Brier score (lower is better) on the five main datasets, mean ± SE over ten seeds. † marks the informative datasets ($\rho_q \geq 0.05$). QAC[Dir] has the lowest Brier score on E_toy, ChaosNLI and CIFAR-10N (4.7%, 7.6% and 4.9% below the best standard calibrator) and ties isotonic regression on CIFAR-10H (0.0870 vs. 0.0867). The last row chooses the base calibrator for each dataset by nested cross-validation on the calibration set (Appendix D). All methods tie on eICU. The two datasets whose q is computed from the input are in Appendix G.

Method	E_toy†	ChaosNLI†	CIFAR-10H†	CIFAR-10N†	eICU
Uncalibrated	$0.2050_{\pm 0.0017}$	$0.5496_{\pm 0.0065}$	$0.0912_{\pm 0.0014}$	$0.1619_{\pm 0.0023}$	$0.1534_{\pm 0.0022}$
TS	$0.2052_{\pm 0.0017}$	$0.4790_{\pm 0.0055}$	$0.0871_{\pm 0.0013}$	$0.1538_{\pm 0.0020}$	$0.1534_{\pm 0.0022}$
Iso	$0.2065_{\pm 0.0018}$	$0.4113_{\pm 0.0034}$	$0.0867_{\pm 0.0012}$	$0.1478_{\pm 0.0018}$	$0.1532_{\pm 0.0022}$
BBQ	$0.2080_{\pm 0.0018}$	$0.4420_{\pm 0.0035}$	$0.0955_{\pm 0.0014}$	$0.1611_{\pm 0.0022}$	$0.1536_{\pm 0.0022}$
TvA-Iso	$0.2074_{\pm 0.0014}$	$0.4551_{\pm 0.0034}$	$0.0948_{\pm 0.0013}$	$0.1648_{\pm 0.0019}$	$0.1536_{\pm 0.0021}$
QAC[Dir] ours, 2-step	$\mathbf{0.1956}_{\pm 0.0015}$	$\mathbf{0.3800}_{\pm 0.0034}$	$0.0870_{\pm 0.0013}$	$\mathbf{0.1406}_{\pm 0.0016}$	$\mathbf{0.1531}_{\pm 0.0022}$
QAC[C] CV-selected (per-dataset)	$\mathbf{0.1956}_{\pm 0.0016}^{\text{TS}}$	$\mathbf{0.3800}_{\pm 0.0034}^{\text{Dir}}$	$\mathbf{0.0853}_{\pm 0.0012}^{\text{Spline}}$	$\mathbf{0.1406}_{\pm 0.0016}^{\text{Dir}}$	$\mathbf{0.1531}_{\pm 0.0022}^{\text{Dir}}$

$C(X)$ (Corollary 8): in the population it equals $\mathbb{E}[\mathrm{Var}(\mathbb{E}[\mathrm{err} \mid \hat{p}, Q] \mid \hat{p})]/\mathrm{Var}(\mathrm{err})$, and $\Delta R^2 > 0$ implies $I(Q; Y \mid C(X)) > 0$. ρ_q ignores $\hat{p}$ and serves only as a screen.

On E_toy, ChaosNLI, CIFAR-10N and CIFAR-10H, q adds information in 9 or 10 of 10 seeds. On Audio MUSAN it adds none; the audio model already encodes the noise level. On Camelyon17, whose quality vector is computed from the image, adding q lowers the R^2.

We use $\rho_q \geq 0.05$ as a screening threshold. We chose it after seeing these datasets, between the input-derived ones (≤ 0.026) and the informative ones (≥ 0.055); it is a convention, like a significance level, and it applies only when calibration and test data share a distribution.

4.3 Does quality improve calibration, and is the gain due to quality?

On the four informative datasets, QAC[Dir] has a lower AURC and a higher error-detection AUROC than each of the four standard calibrators shown (Fig. 2). Its Brier score (Table 2) is lower than theirs on three of the four datasets. On CIFAR-10H it ties with isotonic regression (0.0870 vs. 0.0867).

On these four datasets, the improvement is due to the quality score and not to calibrating twice. Compared with a constant q, the real q lowers both the Brier score and the AURC for each of the nine base calibrators and in every seed, on all four informative datasets. Averaged over bases, the Brier score decreases by 0.011 on E_toy, 0.016 on ChaosNLI, 0.004 on CIFAR-10N and 0.001 on CIFAR-10H (paired t-test, $p < 10^{-6}$). Shuffled q gives the same conclusion (Fig. 15). Three points limit these results. Three of the four informative datasets use a q derived from several annotations of the test item, and among the measurement-quality datasets only the simulated E_toy benefits. On top-label ECE and NLL (Appendix F.1), the results are mixed: for example, QAC[Dir] has a higher NLL than temperature scaling on both CIFAR datasets.

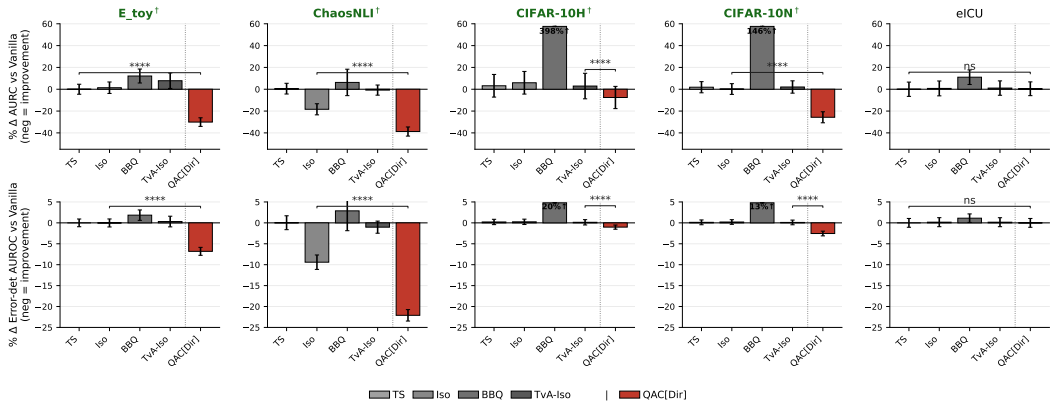

Figure 2: AURC (top) and error-detection AUROC (bottom) on the five main datasets: bars show the percentage change from the uncalibrated model (Vanilla), signed so that lower is better; clipped bars are labelled with their value. Stars: paired bootstrap test of QAC[Dir] against the best of the four standard calibrators shown; ns, not significant. †: informative ($\rho_q \geq 0.05$). Other metrics: Appendix F.1.

Table 3: Real q minus constant q, averaged over the nine base calibrators: mean over ten seeds, and the number of seeds with a positive difference.

dataset	error-detection ΔAUROC	class-discrimination ΔAUROC
E_toy	$+0.053$ $(10/10)$	$+0.013$ $(10/10)$
ChaosNLI	$+0.059$ $(10/10)$	$+0.015$ $(10/10)$
CIFAR-10H	$+0.021$ $(10/10)$	-0.001 $(1/10)$
CIFAR-10N	$+0.032$ $(10/10)$	-0.000 $(4/10)$
eICU	$+0.002$ $(10/10)$	$+0.002$ $(9/10)$
Audio MUSAN	-0.004 $(2/10)$	-0.001 $(0/10)$
Camelyon17	$+0.001$ $(4/10)$	-0.003 $(0/10)$

4.4 Which metrics improve?

Section 3 predicts that QAC changes the confidence of samples that share a score but differ in quality. This reorders predictions by their chance of being correct, and changes the predicted class only when the correction is large. Table 3 confirms this. Compared with a constant q, the real q raises error-detection AUROC by 0.021 to 0.059 on the four informative datasets, for every base and in every seed. Class discrimination rises by about a quarter as much on E_toy and ChaosNLI and does not change on the two CIFAR datasets.

4.5 When does quality not help?

Section 2 predicts that a q computed from the input helps only if the model has not already learned it, and Section 4.2 noted that ρ_q screens only data from the same distribution. We test both.

Quality computed from the input. On Audio MUSAN, q is the noise level of the clip, which the audio model receives as part of its input. On Camelyon17, q is computed from the image that the classifier receives. In both cases q adds no information beyond the model's probabilities (Table 1). Compared with a constant q, the real q gives no gain in error detection, and the Brier score is slightly worse for all nine bases (Audio $+0.0006$, Camelyon17 $+0.002$), and QAC[Dir] is 16% (Audio) and 30% (Camelyon17) worse than the best standard calibrator. The diagnostic ρ_q flags both datasets before QAC is fitted. Camelyon17 is also calibrated on its training hospitals and tested on others, so its failure can come from q being computed from the image, from the shift, or from both.

Shift between hospitals: eICU. On eICU, when calibration and test patients come from the same hospitals, q adds almost no information ($\Delta R^2 = +0.003$) and all methods have a Brier score of about 0.153. We also fit the calibrators on the training hospitals and test on the 30% of hospitals held out in each seed, measuring the largest ECE over three tiers of patient quality (Table 15, Appendix H.6). Relative to temperature scaling, Dirichlet calibration alone lowers this worst-tier ECE by 18% and QAC[Dir] by 22%. The difference between real and constant q (-0.0018) is not significant at ten seeds ($p = 0.17$), and because the tiers are defined by q, this metric favours methods that use q. We therefore cannot attribute the reduction to the quality score.

Finally, when noise is added to q, the gain decreases steadily; half of it is lost when the noise is as large as the variation of q (Appendix H.5).

5 Limitations and conclusion

Limitations. The absolute effects are small (Brier -0.001 to -0.016 against a constant q). q must be available at calibration and inference; for label ambiguity this requires several annotators per new sample, and on ChaosNLI and the CIFAR datasets q comes from the annotations that define the test label. The threshold $\rho_q \geq 0.05$ was chosen on these datasets, the eICU shift result cannot be attributed to q, and the sign analysis rests on small correlations on one dataset.

Conclusion. QAC adds a second calibration stage that uses per-sample data-quality scores. Where the score predicted errors beyond the model's output, it lowered the Brier score relative to the same stage with a constant score, mainly by identifying which predictions are likely to be wrong.

AI use statement

We used generative AI tools to aid and polish the writing of this paper. The authors reviewed all AI-assisted content and take responsibility for the final manuscript.

Ethics statement

This work uses only previously collected, publicly available or credentialed datasets and involves no new data collection and no human subjects research by the authors. The clinical dataset (eICU) is de-identified and was accessed under its PhysioNet data use agreement; no data derived from it is redistributed with this submission, and the accompanying code includes only its processing pipeline. The remaining datasets (ChaosNLI, CIFAR-10H, CIFAR-10N, Camelyon17, MUSAN) are public research benchmarks used as distributed. The main ethical risk we can foresee is specific to the method: a quality signal that is confounded with a protected or clinically meaningful attribute would let a calibrator vary its confidence along that attribute. The pre-deployment diagnostic and the negative results we report are intended to make that failure visible before deployment rather than after, and Section 5 states the conditions under which we do not recommend using the method.

Broader impact. QAC uses data-quality scores that practitioners already collect. If the direction of a quality score's effect is assumed rather than estimated, a model can become more confident on the records it gets wrong. Estimating the direction on held-out data, and screening q with the diagnostic before deployment, reduces this risk.

Reproducibility statement

All reported results are averaged over ten seeds under a paired protocol with the multiplicity correction described in Appendix D, which also records data splits, base-classifier training, calibrator hyperparameters and the stopping rules used. Appendix C gives the provenance of every dataset and the exact definition of the quality signal q used for each. Full statements and proofs of all theoretical results are in Appendix B, with the assumptions stated inline. Extended per-dataset and per-calibrator results, including the cells where the method does not help, are in Appendix F. An anonymised implementation that regenerates the main results table from cached model outputs is provided as supplementary material; the clinical datasets are excluded from it for the data-use reasons given above, and the code to rebuild them from the source databases is included instead.

REFERENCES

Ilia Azizi, Juraj Bodik, Jakob M Heiss, and Bin Yu. Clear: Calibrated learning for epistemic and aleatoric risk. In *International Conference on Learning Representations*, volume 2026, pp. 157741–157813, 2026.

Arindam Banerjee, Xin Guo, and Hui Wang. On the optimality of conditional expectation as a bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669, 2005.

Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 135(643):1512–1519, 2009.

Jingya Cheng, Jiazi Tian, Federica Spoto, Alaleh Azhir, Daniel Mork, and Hossein Estiri. Signal fidelity index-aware calibration for addressing distributional shift in predictive modeling across heterogeneous real-world data. *Scientific Reports*, 16(1):2807, 2025.

Thomas M Cover, Joy A Thomas, Jean-Yves Chouinard, and Stefan M Moser. *Elements of Information Theory*. Wiley, Year, 2006.

Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*, pp. 1184–1193. PMLR, 2018.

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

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

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

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. 2021.

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

Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. Calibration of neural networks using splines. 2020.

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

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. 2021.

Jakob Heiss, Sören Lambrecht, Jakob Weissteiner, Hanna Wutte, Žan Žurič, Josef Teichmann, and Bin Yu. Jucal: Jointly calibrating aleatoric and epistemic uncertainty in classification tasks. *arXiv preprint arXiv:2602.20153*, 2026.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, et al. Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. *Studies in health technology and informatics*, 216:574, 2015.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

Jana Janková, Rajen D Shah, Peter Bühlmann, and Richard J Samworth. Goodness-of-fit testing in high dimensional generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3):773–795, 2020.

Michael G Kahn, Tiffany J Callahan, Juliana Barnard, Alan E Bauck, Jeff Brown, Bruce N Davidson, Hossein Estiri, Carsten Goerg, Erin Holve, Steven G Johnson, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *Egems*, 4(1):1244, 2016.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Meelis Kull and Peter Flach. Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 68–85. Springer, 2015.

Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial intelligence and statistics*, pp. 623–631. PMLR, 2017.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32, 2019.

Adrien Le Coz, Stéphane Herbin, and Faouzi Adjed. Confidence calibration of classifiers with many classes. *Advances in Neural Information Processing Systems*, 37:77686–77725, 2024.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632, 2005.

Yixin Nie, Xiang Zhou, and Mohit Bansal. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9131–9143, 2020.

Feiyang Pan, Xiang Ao, Pingzhong Tang, Min Lu, Dapeng Liu, Lei Xiao, and Qing He. Field-aware calibration: a simple and empirically strong method for reliable probabilistic predictions. In *Proceedings of The Web Conference 2020*, pp. 729–739, 2020.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9617–9626, 2019.

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

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):180178, 2018.

Leonard J Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Freddie Bickford Smith, Jannik Kossen, Eleanor Trollope, Mark Van Der Wilk, Adam Foster, and Tom Rainforth. Rethinking aleatoric and epistemic uncertainty. *arXiv preprint arXiv:2412.20892*, 2024.

David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015.

Dirk Tasche. Calibrating sufficiently. *Statistics*, 55(6):1356–1386, 2021.

Jenthe Thienpondt, Brecht Desplanques, and Kris Demuynck. The idlab voxsrc-20 submission: Large margin fine-tuning and quality-aware score calibration in dnn based speaker verification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5814–5818. IEEE, 2021.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. *arXiv preprint arXiv:2110.12088*, 2021.

Penghui Wei, Weimin Zhang, Ruijie Hou, Jinquan Liu, Shaoguo Liu, Liang Wang, and Bo Zheng. Posterior probability matters: Doubly-adaptive calibration for neural predictions in online advertising. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2645–2649, 2022.

Nicole Gray Weiskopf and Chunhua Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144–151, 2013.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pp. 2001, 2001.

OVERVIEW OF THE APPENDIX

The appendix follows the order of the paper. Appendix A lists the notation. Appendix B gives the full statements and proofs behind Section 2, including the closed-form case. Appendix C describes each dataset, its quality score and its trained classifier, and Appendix D the splits, the Stage-2 settings and the three control arms (real, constant and shuffled q). Appendix E supports Section 3: the E_toy example, the direction of each eICU quality dimension, the Stage-2 design comparison, user-provided quality and the comparison with JUCAL and SFI. Appendix F supports Sections 4.3 and 4.4 with all metrics and per-base results. Appendix G covers the two datasets whose quality is computed from the input, and Appendix H the ablations on the form and noise of q, the hospital-shift analysis and training-time alternatives. The last sections give a checklist for new datasets, a comparison with routing on q, and supporting figures.

A NOTATION

$\mathcal{X}$, $\mathcal{Q}$	input space, quality-score space
$X \in \mathcal{X}$, x	classifier input (random variable, observed value)
$Y \in \{1, \dots, K\}$, y	class label (K classes)
$Q \in \mathcal{Q}$, q	data-quality score, observed at calibration and inference
L	strictly proper scoring rule (loss)
C	post-hoc base calibrator (e.g. isotonic regression, temperature scaling)
$\mathrm{QAC}[C]$	QAC applied on top of C, taking $(C(X), Q)$
$\mathrm{QAC}[C; q^{\text{real}}]$, $\mathrm{QAC}[C; q^{\text{shuf}}]$, $\mathrm{QAC}[C; q^{\text{const}}]$	QAC fitted with the real, shuffled or constant quality score
ρ_q	five-fold cross-validated R^2 of a gradient-boosted regressor that predicts the $0/1$ error of C from q
d_φ	Bregman divergence associated with the convex generator φ of L
$\sigma(C(X))$	σ-algebra generated by $C(X)$

B THEORY: FULL STATEMENTS AND PROOFS

B.1 SETUP

Let $(X, Y, Q) \sim P$ be a joint random triple where $X \in \mathcal{X}$ is the classifier input, $Y \in \{1, \dots, K\}$ a class label in $\{1, \dots, K\}$, and $Q \in \mathcal{Q}$ an auxiliary data-quality signal that may be observed at calibration and inference time but not necessarily at training time. Predictors are taken to be arbitrary measurable functions into the $(K-1)$-simplex,

$$f : \mathcal{X} \to \Delta^{K-1}, \qquad g : \mathcal{X} \times \mathcal{Q} \to \Delta^{K-1},$$

so that realisability of the regular conditional probabilities $P(Y \mid X)$ and $P(Y \mid X, Q)$ is automatic. Let $L : \Delta^{K-1} \times \{1, \dots, K\} \to \mathbb{R}$ be a strictly proper scoring rule. By the Savage–Bregman representation of proper scoring rules (Savage, 1971; Gneiting & Raftery, 2007; Banerjee et al., 2005), there exists a strictly convex generator $\varphi : \Delta^{K-1} \to \mathbb{R}$ such that for all $p, q \in \Delta^{K-1}$ the excess-risk identity

$$\mathbb{E}_{Y \sim q}[L(p, Y)] - \mathbb{E}_{Y \sim q}[L(q, Y)] \;=\; d_\varphi(p, q)$$

holds, where $d_\varphi(p, q) := \varphi(p) - \varphi(q) - \langle \nabla\varphi(q), p - q \rangle$ is the Bregman divergence with q as the truth and p as the prediction; d_φ is non-negative and vanishes iff $p = q$.

Definition 4 (Assumptions)**. (A1)** L is strictly proper. **(A2)** $I(Q; Y \mid X) > 0$, i.e. Q provides additional information about Y given X.

(Realisability of the Bayes optima $P(Y \mid X), P(Y \mid X, Q)$ in the predictor classes is automatic by our choice of f, g as arbitrary measurable functions; no separate assumption is needed.)

B.2 Proof of the Observable Aleatoric Uncertainty (OAU) theorem

Proof of Theorem 1. By (A1) and the realisability of the Bayes optima, the minima are attained at $f^* = P(Y \mid X)$ and $g^* = P(Y \mid X, Q)$ respectively, with values $\mathbb{E}[L(P(Y \mid X), Y)]$ and $\mathbb{E}[L(P(Y \mid X, Q), Y)]$. Conditioning on (X, Q) in the excess-risk identity with truth $P(Y \mid X, Q)$ and prediction $p = P(Y \mid X)$,

$$\mathbb{E}\big[L(P(Y \mid X), Y) - L(P(Y \mid X, Q), Y) \,\big|\, X, Q\big] \;=\; d_\varphi(P(Y \mid X), P(Y \mid X, Q)).$$

Taking outer expectation over $(X, Q) \sim P$,

$$\mathbb{E}[L(P(Y \mid X), Y)] - \mathbb{E}[L(P(Y \mid X, Q), Y)] \;=\; \mathbb{E}_{(X,Q)}\big[d_\varphi(P(Y \mid X), P(Y \mid X, Q))\big] \;\geq\; 0.$$

By (A2), $P(Y \mid X, Q) \neq P(Y \mid X)$ on a set of positive P-measure, and strict convexity of φ gives $d_\varphi(P(Y \mid X), P(Y \mid X, Q)) > 0$ on that set, so the inequality is strict. □

B.3 Corollaries with proofs

Corollary 5 (Necessity of post-hoc when Q is unavailable at training)**.** *When Q is observed only at inference time, any training-time predictor takes the form $f : \mathcal{X} \to \Delta^{K-1}$ and therefore cannot depend on Q. Its optimum loss is $\mathbb{E}[L(P(Y \mid X), Y)]$, which by Theorem 1 is strictly larger than the optimum of a post-hoc predictor that uses both X and Q, $\mathbb{E}[L(P(Y \mid X, Q), Y)]$. Hence post-hoc Quality-Aware Uncertainty Calibration (*QAC*) is the* only available route *to use Q in this regime.*

Proof. Direct from Theorem 1: the LHS of the strict inequality is the post-hoc optimum, the RHS is the training-time optimum. □

What Corollary 5 does *not* claim. Corollary 5 does not say post-hoc QAC outperforms training-time joint $C_{\text{joint}}(X, Q)$ when both are feasible. Training-time joint is a valid alternative route to the OAU optimum when Q is observable at training; the empirical comparison between routes in that regime depends on the sufficiency of C for X w.r.t. (Y, Q) (the joint optimum is attained by post-hoc iff C is sufficient, a multi-class extension of Tasche (2021)). The empirical setting motivating this paper—data-quality signals from downstream pipelines that produce Q post-collection—is the deployment-only-Q regime where Corollary 5 applies; renaming this corollary the "necessity-of-post-hoc" condition rather than "post-hoc dominance" emphasises the structural-availability claim.

Proposition 6 (Three-regime classification)**.** *The OAU gap $\Delta_L := \min_f \mathbb{E}[L(f(X), Y)] - \min_g \mathbb{E}[L(g(X, Q), Y)]$ partitions (X, Y, Q) joint distributions into three exclusive and exhaustive regimes:*

- Active: $I(Q; Y \mid X) > 0$. *Then* $\Delta_L > 0$.
- Determined: $Q \in \sigma(X)$ *(Q measurable function of X). Then* $\Delta_L = 0$.
- Independent: $Q \perp Y \mid X$ *but* $Q \notin \sigma(X)$. *Then* $\Delta_L = 0$.

Proof. Exhaustiveness: every (X, Y, Q) falls in exactly one regime. If $I(Q; Y \mid X) > 0$, Active. If $I(Q; Y \mid X) = 0$, then $Q \perp Y \mid X$ (standard (Cover et al., 2006)); $Q \in \sigma(X)$ (D) or not (I). Mutual exclusivity is by the $I = 0$ vs $I > 0$ partition and the $\sigma(X)$-membership partition. For the gap implications: Active uses Theorem 1 strict positivity. Regimes D and I have $P(Y \mid X, Q) = P(Y \mid X)$ a.s. (equivalence of conditional independence and equality of conditional posteriors), so $d_\varphi(P(Y \mid X, Q), P(Y \mid X)) = 0$ a.s. for any strictly proper Bregman divergence and $\Delta_L = 0$. □

Corollary 2 (Reliability–resolution bound; stated in Section 2.2). By the Bröcker decomposition (Bröcker, 2009; DeGroot & Fienberg, 1983), $\mathrm{Brier}(\cdot) = \mathrm{Reliability}(\cdot) - \mathrm{Resolution}(\cdot) + \mathrm{Uncertainty}$, where $\mathrm{Uncertainty}$ depends only on the marginal of Y and is identical for C and $\mathrm{QAC}[C]$. For multiclass Brier the decomposition is applied class-wise.

Proof. Subtracting the Brocker decomposition for $\mathrm{QAC}[C]$ from that for C and using the cancellation of $\mathrm{Uncertainty}$ gives $\mathrm{Brier}(C) - \mathrm{Brier}(\mathrm{QAC}[C]) = \big(\mathrm{Reliability}(C) - \mathrm{Reliability}(\mathrm{QAC}[C])\big) + \big(\mathrm{Resolution}(\mathrm{QAC}[C]) - \mathrm{Resolution}(C)\big)$. The bound follows from $\mathrm{Reliability}(\mathrm{QAC}[C]) \geq 0$. □

Corollary 3 (quality determined by the calibrated score; stated in Section 2.2). If $Q = h(C(X))$, then $\sigma(C(X), Q) = \sigma(C(X))$.

Proof. If $Q = h(C(X))$, then $(C(X), Q) = (C(X), h(C(X)))$ is a measurable function of $C(X)$. Hence any measurable function of $(C(X), Q)$ can be re-expressed as a measurable function of $C(X)$, and the law of the prediction $\mathrm{QAC}[C](C(X), Q)$ is determined by $C(X)$ alone. □

Corollary 7 (Quantified OAU: log-loss specialisation)**.** *Under the assumptions of Theorem 1, the OAU gap admits the representation*

$$\Delta_L \;=\; \mathbb{E}_{(X,Q)}\big[d_\varphi(P(Y \mid X), P(Y \mid X, Q))\big],$$

which for the log-loss $L(p, y) = -\log p_y$ specialises to the conditional mutual information

$$\Delta_{\log} \;=\; I(Q; Y \mid X).$$

Consequently $\Delta_{\log} \le \min\{H(Y \mid X), H(Q \mid X)\}$, and applied class-wise Pinsker gives $\Delta_{\mathrm{Brier}} \le 2\Delta_{\log}$. The normalised quantity $\rho_Q^{\mathrm{theory}} := I(Q; Y \mid X)/H(Y \mid X) \in [0, 1]$ is the fraction of X-conditional uncertainty Q resolves.

Proof. The general representation is the gap formula of Theorem 1. For the log-loss specialisation, write the gap in expectation form,

$$\Delta_{\log} = \mathbb{E}\big[L(P(Y \mid X), Y) - L(P(Y \mid X, Q), Y)\big] = \mathbb{E}\left[\log \frac{P(Y \mid X, Q)(Y)}{P(Y \mid X)(Y)}\right] = I(Q; Y \mid X),$$

the last equality by the definition of conditional mutual information applied to $(X, Q, Y) \sim P$ (Cover et al., 2006, Theorem 2.5.2). The bounds $I(Q; Y \mid X) \le H(Y \mid X)$ and $I(Q; Y \mid X) \le H(Q \mid X)$ are standard conditional mutual information (CMI) inequalities. The Brier bound follows from Pinsker's inequality applied class-wise. □

Corollary 8 (Sufficiency of C for post-hoc to reach the OAU optimum)**.** *Let $C : \mathcal{X} \to \Delta^{K-1}$ be a base calibrator. The post-hoc framework $\mathrm{QAC}[C]$ taking $(C(X), Q)$ as input attains the OAU optimum $\min_g \mathbb{E}[L(g(X, Q), Y)]$ if and only if $C(X)$ is sufficient for X with respect to (Y, Q), i.e. $X \perp Y \mid (C(X), Q)$. When sufficiency fails, the post-hoc gap to the joint optimum is the* grouping loss *(Kull & Flach, 2015) specialised to the post-hoc σ-field hierarchy $\sigma(C(X), Q) \subseteq \sigma(X, Q)$:*

$$\mathrm{GL}(C; Q) \;:=\; \mathbb{E}\big[d_\varphi\big(\, P(Y \mid X, Q),\; \mathbb{P}(Y \mid C(X), Q)\big)\big] \;\ge\; 0.$$

This generalises Tasche (2021, Prop. 3.8) to multi-class, explicit external Q, and any strictly proper L.

Proof. ($\Leftarrow$) Suppose $X \perp Y \mid (C(X), Q)$. By the equivalence between conditional independence and equality of conditional posteriors, $P(Y \mid X, Q) = \mathbb{P}(Y \mid C(X), Q)$ a.s. Define $r^*(c, q) := \mathbb{P}(Y \mid C(X){=}c, Q{=}q)$; then $r^*(C(X), Q) = P(Y \mid X, Q)$ a.s., so r^* is a measurable function of $(C(X), Q)$ that achieves the joint Bayes risk and the OAU optimum.

($\Rightarrow$) Suppose there exists $r^* : \mathrm{range}(C) \times \mathcal{Q} \to \Delta^{K-1}$ with $\mathbb{E}[L(r^*(C(X), Q), Y)] = \mathbb{E}[L(P(Y \mid X, Q), Y)]$. By the conditional excess-risk identity applied with truth $P(Y \mid X, Q)$ and prediction $r^*(C(X), Q)$,

$$\mathbb{E}_{(X,Q)}\big[d_\varphi\big(r^*(C(X), Q),\; P(Y \mid X, Q)\big)\big] \;=\; 0,$$

so $d_\varphi(r^*(C(X), Q), P(Y \mid X, Q)) = 0$ a.s. Strict convexity of φ gives $r^*(C(X), Q) = P(Y \mid X, Q)$ a.s., i.e. $P(Y \mid X, Q)$ is $\sigma(C(X), Q)$-measurable. By the tower property, $\mathbb{P}(Y \mid C(X), Q) = \mathbb{E}[P(Y \mid X, Q) \mid C(X), Q] = P(Y \mid X, Q)$ a.s., which is the conditional independence statement $X \perp Y \mid (C(X), Q)$. □

Empirical example of Corollary 8. On the synthetic E_toy data-generating process (DGP) with redundant cross-source q, $C(s_1)$ is the X-only Bayes posterior — sufficient for Y alone but not for the joint (Y, Q) inference, since the X-only posterior collapses two situations (large $|s_1|$ from a true large z vs from small z with corruption) into the same value. Q disentangles them, but the post-hoc layer sees only the already-collapsed $C(s_1)$. This produces a small but paired-CI-detectable post-hoc-vs-joint-training gap on E_toy ($\sim$0.006–0.010 Brier); on real datasets the gap is typically below the empirical detectability floor and the choice between routes is operational (Appendix H.7).

B.4 HETEROSCEDASTIC-CHANNEL CEILING: CLOSED-FORM OAU GAP

Theorem 9 (Heteroscedastic-Channel ceiling). *Let $V \sim P_V$ on $\mathbb{R}$, $\sigma \sim P_\sigma$ a positive random variable independent of V, and $X = V + \sigma\xi$ where $\xi \sim \mathcal{N}(0,1)$ is independent of (V, σ). Set $Q := 1/\sigma^2$ (the precision of the input channel). Then under squared-loss with $Y = V$, the OAU gap is*

$$\Delta_{\mathrm{sq}}(Q \mid X) = \mathbb{E}_X\big[\,\mathrm{Var}_{Q|X}\big(\mathbb{E}[V \mid X, Q]\big)\,\big] \;\geq\; 0,$$

with equality iff $\mathbb{E}[V \mid X, Q]$ is $\sigma(X)$-measurable a.s. When $V \sim \mathcal{N}(\mu_V, \tau^2)$, the Q-aware posterior is conjugate (precision-weighted) for any P_σ on $\sigma > 0$: $\mathbb{E}[V \mid X, Q] = ((1/\tau^2)\mu_V + QX)/((1/\tau^2) + Q)$ and $\mathrm{Var}(V \mid X, Q) = 1/((1/\tau^2) + Q)$.

Proof. The Bayes-optimal squared-loss predictor of V given a sub-σ-algebra $\mathcal{G}$ is the conditional expectation $\mathbb{E}[V \mid \mathcal{G}]$ with risk $\mathbb{E}[\mathrm{Var}(V \mid \mathcal{G})]$. Apply the law of total variance conditional on X: $\mathrm{Var}(V \mid X) = \mathbb{E}[\mathrm{Var}(V \mid X, Q) \mid X] + \mathrm{Var}(\mathbb{E}[V \mid X, Q] \mid X)$. Take expectation in X and subtract $\mathbb{E}[\mathrm{Var}(V \mid X, Q)]$ from both sides: $\Delta_{\mathrm{sq}} = \mathbb{E}[\mathrm{Var}(V \mid X)] - \mathbb{E}[\mathrm{Var}(V \mid X, Q)] = \mathbb{E}_X[\mathrm{Var}_{Q|X}(\mathbb{E}[V \mid X, Q])]$, which is non-negative pointwise and zero iff $\mathbb{E}[V \mid X, Q]$ is constant in Q at fixed X a.s. The Gaussian-conjugacy posterior follows from standard normal-normal-precision Bayesian arithmetic (the posterior of a Gaussian prior conditioned on a Gaussian observation with known precision is Gaussian, with precision equal to the sum of prior and likelihood precisions and mean equal to the precision-weighted average); the conjugacy applies pointwise in Q, hence holds for any P_σ. □

Binary-classification analogue. For $Y = \mathbf{1}\{V > 0\}$ with the same heteroscedastic Gaussian channel and Gaussian prior, the conditional posterior is $\pi(X, Q) := \mathbb{P}(V > 0 \mid X, Q) = \Phi\big(((1/\tau^2)\mu_V + QX)/\sqrt{(1/\tau^2) + Q}\big)$, and the log-loss OAU gap admits the form $\Delta_{\log} = I(Q; Y \mid X) = \mathbb{E}_X[H(\bar{\pi}(X)) - \mathbb{E}_{Q|X}[H(\pi(X, Q))]]$ where $\bar{\pi}(X) = \mathbb{E}_{Q|X}[\pi(X, Q)] = \mathbb{P}(V > 0 \mid X)$ and H is the binary entropy. Equality $\Delta_{\log} = 0$ holds iff $\pi(X, Q) = \bar{\pi}(X)$ a.s., by Jensen's inequality and strict concavity of H.

Numerical validation (Section 4). On synthetic data drawn from this DGP at $n \in \{500, 10^3, 5 \cdot 10^3, 2 \cdot 10^4, 10^5\}$ with 10 seeds, the empirical $\hat{\Delta}_{\mathrm{sq}}$ converges to the theoretical $\Delta_{\mathrm{sq}} = 0.1289$ to four decimal places at $n = 10^5$, with finite-sample variance scaling as $\sim 1/\sqrt{n}$ (Fig. 3). This is the cleanest possible validation of Theorem 1: a closed-form theoretical prediction matched by empirical estimation, with variance scaling at the predicted rate.

C DATASETS, QUALITY SCORES AND MODELS

We evaluate on seven datasets, each with its own quality score q.

E_toy. A synthetic two-class Gaussian mixture measured by two sensors; sensor 1 is corrupted for 20% of samples. The classifier receives sensor 1 only, and q is the agreement between the two sensors (one dimension). The classifier is a multilayer perceptron.

ChaosNLI (Nie et al., 2020). Three-way natural-language inference with a frozen pretrained encoder. q is the entropy of 100 annotator labels per item (one dimension). The classifier is a DeBERTa-v3-base natural-language-inference model.

CIFAR-10H (Peterson et al., 2019). Ten-class CIFAR-10 test images with about 50 annotator labels each. q is the fraction of annotators who chose the most frequent label (one dimension). The classifier is a ResNet-18 trained on the CIFAR-10 training set.

CIFAR-10N (Wei et al., 2021). Ten-class CIFAR-10 with noisy human training labels (`aggregate` version). q is the CIFAR-10H annotator entropy of each test image (one dimension). CIFAR-10H and CIFAR-10N share this annotator signal, so they are not independent evidence. The classifier is a ResNet-18 trained on the noisy labels.

eICU (Pollard et al., 2018). In-hospital mortality for intensive-care stays in 41 hospitals. q is a five-dimensional vector of patient-level record checks: completeness, plausibility, agreement between nurse-charted and monitor values, timeliness, and integrity. The classifier is a multilayer perceptron

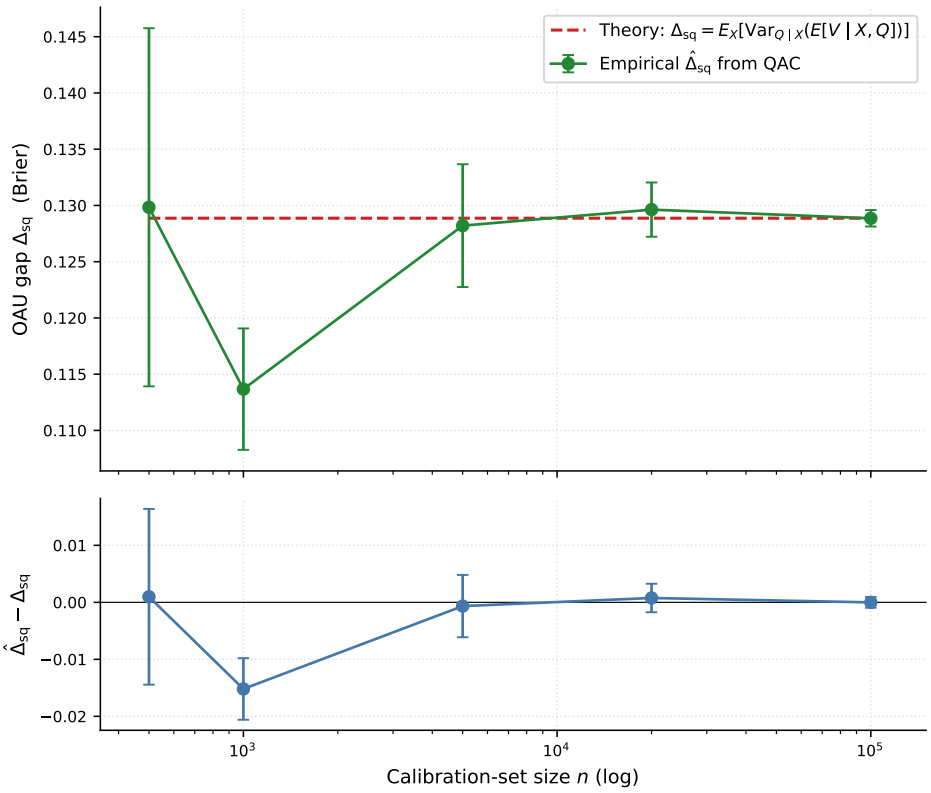

Figure 3: Closed-form Gaussian case (Theorem 9). *Top*: the gap estimated with post-hoc QAC (green) approaches the closed-form value (dashed line) as the calibration set grows. *Bottom*: the difference from the closed-form value, which shrinks roughly as $1/\sqrt{n}$. Ten seeds per n; 95% intervals.

with two hidden layers of 64 units on the stay's clinical features; 30% of the hospitals are held out for testing in each seed.

Audio MUSAN. Speech Commands v0.02 keywords mixed with MUSAN noise and classified by a frozen Audio Spectrogram Transformer. q is the measured signal-to-noise ratio of each clip (one dimension), which is computed from the input. The classifier is the Audio Spectrogram Transformer fine-tuned on Speech Commands v2, used frozen.

Camelyon17 (Koh et al., 2021). Tumour classification of histopathology images from several hospitals. q is a four-dimensional image-quality vector computed from the image. The classifier is a ResNet-50; calibration uses in-distribution hospitals and testing uses held-out ones (WILDS split).

Scope of Q and deployability at inference. The framework's intended Q is a *process-quality measurement*: a signal that audits the data-generation pipeline and flags potential errors or noise in it, rather than an arbitrary deployable covariate. Annotator agreement audits the labelling process; SNR / completeness / schema-conformance flags audit the input-collection pipeline; cross-source agreement audits redundancy in the upstream measurement. Generic demographic or acquisition descriptors (e.g., camera position, patient age) are deployable in the sense that they are observable at inference time, but they are not what the framework is designed to consume; whether they have $I(Q;Y \mid X) > 0$ is dataset-dependent and they are excluded from our Q-design space. Within the kept datasets, two regimes serve different evidential roles. *Input-quality* Q is computed by the input pipeline at no labelling cost and is fully deployable: E_toy (controlled corruption indicator), eICU (5-d Kahn–Weiskopf DQ flags from EHR fields), Audio MUSAN (per-clip measured SNR), and Camelyon17 (4-d image-quality vector). These cases form the deployment-realistic evidence in the paper. *Label-process* Q is computed from multi-rater annotation distributions and serves as principle-

validation: ChaosNLI (annotator-entropy from 100 raters), CIFAR-10H (50-rater max probability), and CIFAR-10N (CIFAR-10H entropy applied to noisy training labels). The benchmark-given Q on these datasets demonstrates that the framework exploits an informative quality signal when one is provided; deployment of label-process Q on new test samples requires either a multi-rater pipeline that emits a per-sample disagreement signal (medical second-opinion review, content-moderation consensus, crowdsourced QA), an annotator/source historical prior, or a self-reported confidence channel decoupled from the actual votes. We discuss this scope in Section 5 (limitations).

D EXPERIMENTAL PROTOCOL DETAILS

Statistical protocol. All comparisons use 10 seeds with paired bootstrap 95% CIs (5000 resamples). For each (dataset, base, method) and each seed, we record the per-test-sample loss-vector $(\ell_i)_{i=1}^{N_{\text{test}}}$ on the held-out test split. For any pair of methods (A, B) the paired statistic at seed s is $\bar{\Delta}_s := \frac{1}{N_{\text{test}}} \sum_i (\ell_i^{(A)} - \ell_i^{(B)})$, and the 95% CI is formed by bootstrapping the seed-level paired differences. Disposition rules: an *improvement* means the CI strictly excludes 0 in the favorable direction; a *tie* means the CI crosses 0; an *under-performance* means the CI strictly excludes 0 in the unfavorable direction.

Calibration / test split. For each dataset and each seed we use a deterministic 50/50 calibration/test split for ChaosNLI, Audio MUSAN, and E_toy. eICU uses the dataset-standard 41-hospital split, CIFAR-10H/N use the test-set 5000/5000 cal/test split, and Camelyon17 uses the WILDS ID-cal / OOD-test split. Splits are identical across all method variants (C alone, QAC$[C; q^{\text{const}}]$, QAC$[C; q^{\text{shuf}}]$, QAC$[C; q^{\text{real}}]$) for paired comparison.

Platt and beta calibrators. Platt scaling is one-vs-rest logistic regression on the logit of each class probability; beta calibration is one-vs-rest logistic regression on $[\log p, -\log(1-p)]$ (Kull et al., 2017). Both are fitted with scikit-learn on the calibration split.

Hyperparameters (Stage-2 regressor). We use sklearn `HistGradientBoosting Regressor` with `monotonic_cst=[1, 0]` (monotone in confidence, free in q), `max_iter=100`, `max_depth=4`, `min_samples_leaf=50`, `learning_rate=0.05`. These are not tuned per-dataset; the same defaults are used everywhere.

Three-method ablation. The decomposition methodology of Section 4.1 requires three parallel runs of the QAC-HGBT layer:

1. QAC$[C; q^{\text{const}}]$: $q_i \equiv 0$. Captures the *stacked-recalibration* component (HGBT flexibility on C's output alone).
2. QAC$[C; q^{\text{shuf}}]$: q_i permuted across calibration samples. Captures stacked-recalibration plus any noise-floor effect of having a second input regardless of its information content.
3. QAC$[C; q^{\text{real}}]$: q_i as observed. Captures the total effect.

The effect of calibrating twice is $\text{loss}(C) - \text{loss}(\text{QAC}[C; q^{\text{const}}])$. The effect of the quality score is $\text{loss}(\text{QAC}[C; q^{\text{const}}]) - \text{loss}(\text{QAC}[C; q^{\text{real}}])$; this is the main comparison in Section 4. The shuffled version is a second control.

Choosing the base calibrator by cross-validation. The main-table 1-step Q and 2-step QAC$[C]$ columns report the variant selected by nested 5-fold cross-validation on the calibration split. For each (dataset, seed) we partition the calibration set into 5 folds, score each candidate variant on the held-out within-calibration Brier, pick the variant with the best held-out Brier, refit on the full calibration set, and evaluate on the test set. The variant most-often picked across the five primary datasets and ten seeds is the paper's deployment recommendation.

Reproducibility. Code is released with the paper. All numerical results in this paper can be re-derived from the cached base-model logits via the SLURM array scripts and aggregator pipeline.

E MECHANISM DETAILS FOR SECTION 3

E.1 E_TOY WORKED EXAMPLE

Source: the E_toy run used for the ensemble comparison, in which Stage 1 is a ten-member ensemble whose logits are pooled and then temperature scaled, and Stage 2 is QAC with the default settings of Section 3. On the test split we keep samples whose Stage-1 top-class confidence lies in $[0.85, 0.95]$ and split them by $q < 0.5$ (sensors disagree) versus $q \geq 0.5$ (agree); on E_toy q is close to binary and 20.3% of samples have $q < 0.5$. Over ten seeds: Stage-1 confidence 0.911 (range 0.908–0.913); accuracy 0.906 overall, 0.953 (0.931–0.965) when the sensors agree and 0.718 (0.677–0.754) when they disagree; QAC output 0.956 (0.937–0.965) and 0.736 (0.696–0.788). On average 356 disagreeing and $1{,}413$ agreeing samples fall in the band per seed.

E.2 DIRECTION OF EACH QUALITY DIMENSION ON EICU

Table 4 gives the result for each dimension.

For each of the ten eICU quality dimensions (scored so that higher means better) we correlate the score with the base model's $0/1$ error on the calibration split while holding the base model's confidence fixed, in two ways: within ten equal-count bins of confidence, and after linear regression of both variables on confidence and logit margin. A dimension is *resolved* if its sign agrees in at least 9 of 10 seeds; a positive correlation means better-scored records are misclassified more often (*inverted*). The seeds are different splits and model initialisations of the same patients, so they show robustness to the split, not independent replication. Without holding confidence fixed, patient-level completeness also appears inverted and hospital-level timeliness appears inverted; both become unresolved once confidence is held fixed.

level	dimension	r (within bins)	r (linear)	seeds > 0
patient	completeness	-0.003	-0.006	5 / 4 (unresolved)
patient	plausibility	-0.025	-0.017	0 / 1 (expected)
patient	patient agreement	$+0.034$	$+0.034$	10 / 10 (inverted)
patient	timeliness	$+0.020$	$+0.015$	10 / 10 (inverted)
patient	integrity	-0.020	-0.017	1 / 1 (expected)
hospital	completeness	-0.007	-0.008	4 / 4 (unresolved)
hospital	plausibility	-0.007	-0.007	3 / 3 (unresolved)
hospital	site agreement	$+0.025$	$+0.025$	9 / 9 (inverted)
hospital	timeliness	$+0.004$	$+0.003$	5 / 5 (unresolved)
hospital	integrity	-0.004	-0.003	3 / 3 (unresolved)

Table 4: Correlation of each eICU quality dimension with base-model error at fixed confidence (10-seed mean) and the number of seeds, out of 10, with a positive correlation under each method.

E.3 STAGE-2 DESIGN: PER-CLASS MONOTONE REGRESSORS

Table 5 gives the results.

We compare the default Stage 2 (one gradient-boosted regressor per class on $(\hat{p}_k, q)$, monotone in $\hat{p}_k$) with two alternatives on the full vector $(\hat{p}, q)$: a single neural network with K outputs, and the same gradient-boosted regressor without the monotonicity constraint. Dirichlet base, scalar q, mean over ten seeds. Brier is similar across designs; the single network has the worst top-label ECE on all five datasets ($1.17\times$–$3.49\times$ ours), and the unconstrained regressor has worse ECE than ours on four of five.

E.4 USER-ELICITED QUALITY (E_TOY)

We treat user-elicited reliability as Q on E_toy. With informative user flags, q_{combined} improves ECE by a ratio of 3.35 ± 1.02 (QAC improves on $8/10$ seeds). With random user flags, ECE drops by 9.5% on average and QAC improves on only $2/10$ seeds; the diagnostic correctly rejects this uninformative Q.

dataset	base C Brier / ECE	ours Brier / ECE	single network Brier / ECE	unconstrained GBT Brier / ECE
E_toy	0.2051 / 0.012	0.1956 / 0.0169	0.1955 / 0.0198	0.1973 / 0.0174
ChaosNLI	0.3945 / 0.032	0.3800 / 0.0335	0.3836 / 0.0414	0.3775 / 0.0296
CIFAR-10H	0.0875 / 0.009	0.0870 / 0.0091	0.0885 / 0.0318	0.0878 / 0.0102
CIFAR-10N	0.1436 / 0.013	0.1406 / 0.0108	0.1413 / 0.0368	0.1413 / 0.0141
eICU	0.1529 / 0.007	0.1531 / 0.0070	0.1534 / 0.0122	0.1533 / 0.0077

Table 5: Stage-2 design comparison (10 seeds).

E.5 Comparison with JUCAL and the Signal Fidelity Index

Table 6 gives the results.

We compare with JUCAL (Heiss et al., 2026) and the Signal Fidelity Index (SFI) calibration of Cheng et al. (2025) where each can be run as published: JUCAL needs a trained ensemble, and SFI scales a single binary probability. All runs use ten-member deep ensembles, ten seeds, and fit every parameter on the calibration split. CLEAR (Azizi et al., 2026) calibrates regression prediction intervals and has no counterpart for class probabilities. On E_toy, where q is external by construction, QAC improves on both; on eICU every method ties with the base and QAC is marginally worse. On CIFAR-10N (not shown), SFI is our own multi-class extension rather than the published method, and JUCAL has the lowest ECE and NLL.

dataset	method	Brier	ECE	NLL	AURC
E_toy	base (pooled ensemble, TS)	0.2046	0.0125	0.3379	0.0547
	JUCAL	0.2047	0.0127	0.3380	0.0546
	SFI (tuned)	0.2017	0.0145	0.3277	0.0450
	QAC	**0.1955**	0.0165	**0.3113**	**0.0390**
eICU	base	0.1516	0.0074	0.2689	0.0370
	JUCAL	0.1516	0.0070	0.2686	0.0368
	SFI (tuned)	0.1516	0.0071	0.2688	0.0369
	QAC	0.1519	0.0082	0.2700	0.0371

Table 6: Head-to-head with JUCAL and SFI (ten seeds).

F Extended results

F.1 Multi-metric comparison

Figure 4 compares all methods on five metrics. Tables 7, 8, 9 and 10 give the values for ECE, NLL, AURC and error-detection AUROC.

The selective-classification master figure (AURC + error-detection AUROC) is in the main paper (Fig. 2). A four-panel multi-metric view (ECE, NLL, AURC, error-detection AUROC) expressed as Δ-from-Vanilla per (dataset, method) is shown below.

Table 7: ECE (lower=better) on the seven kept datasets, 10 seeds; $\pm$ = SE. Locked headline = **QAC[Dir]**; the same 5 SOTA plain calibrators as Tab. 2 are reported. Bold = best method per dataset.

Dataset	Vanilla	TS	Iso	BBQ	TvA-Iso	1-step Q	QAC[Dir]
E_toy†	**0.0113**$_{\pm 0.0009}$	0.0153$_{\pm 0.0013}$	0.0155$_{\pm 0.0017}$	0.0189$_{\pm 0.0016}$	0.0157$_{\pm 0.0009}$	0.0197$_{\pm 0.0012}$	0.0169$_{\pm 0.0024}$
ChaosNLI†	0.2612$_{\pm 0.0037}$	0.1711$_{\pm 0.0039}$	0.0485$_{\pm 0.0028}$	0.0427$_{\pm 0.0069}$	**0.0298**$_{\pm 0.0036}$	0.0336$_{\pm 0.0036}$	0.0335$_{\pm 0.0030}$
CIFAR-10H†	0.0312$_{\pm 0.0008}$	0.0161$_{\pm 0.0007}$	0.0107$_{\pm 0.0010}$	0.0104$_{\pm 0.0006}$	**0.0081**$_{\pm 0.0006}$	0.0093$_{\pm 0.0007}$	0.0091$_{\pm 0.0010}$
CIFAR-10N†	0.0561$_{\pm 0.0014}$	0.0253$_{\pm 0.0016}$	0.0146$_{\pm 0.0010}$	0.0267$_{\pm 0.0018}$	0.0124$_{\pm 0.0006}$	0.0125$_{\pm 0.0007}$	**0.0109**$_{\pm 0.0007}$
eICU	0.0096$_{\pm 0.0011}$	0.0114$_{\pm 0.0009}$	0.0071$_{\pm 0.0009}$	0.0071$_{\pm 0.0010}$	**0.0069**$_{\pm 0.0010}$	0.0089$_{\pm 0.0013}$	0.0072$_{\pm 0.0010}$
Audio MUSAN	0.0552$_{\pm 0.0008}$	0.0136$_{\pm 0.0006}$	0.0102$_{\pm 0.0009}$	0.0131$_{\pm 0.0011}$	**0.0062**$_{\pm 0.0006}$	0.0081$_{\pm 0.0008}$	0.0087$_{\pm 0.0007}$
Camelyon17	0.0984$_{\pm 0.0031}$	0.0708$_{\pm 0.0037}$	0.1129$_{\pm 0.0033}$	0.1068$_{\pm 0.0023}$	0.0470$_{\pm 0.0044}$	**0.0416**$_{\pm 0.0045}$	0.1042$_{\pm 0.0030}$

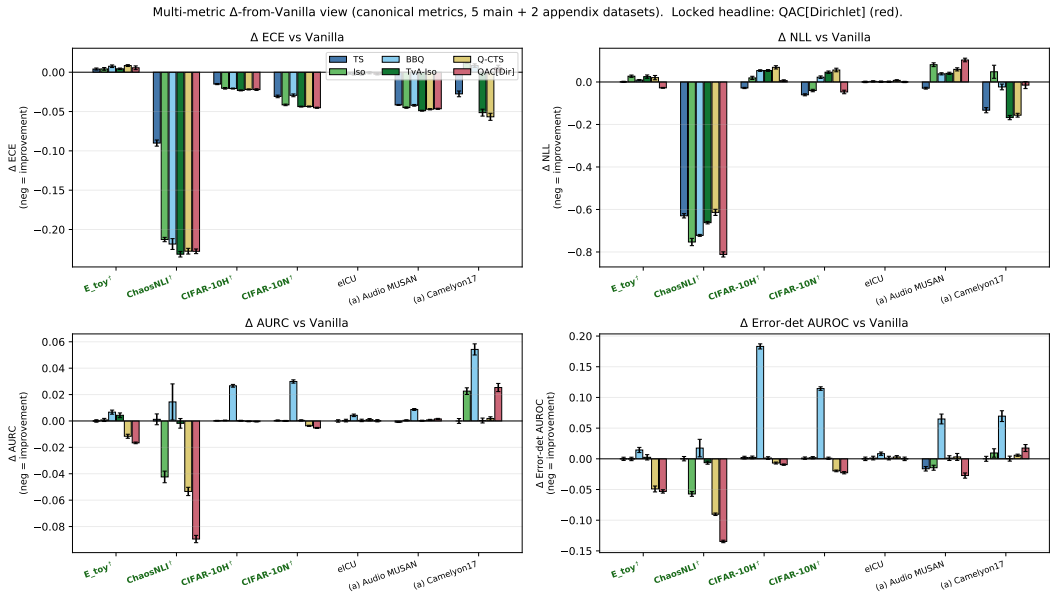

Figure 4: Multi-metric Δ-from-Vanilla view across the seven kept datasets, 10 seeds. Negative bars = improvement (axes flipped for AUROC so all metrics share a "lower=better" sign). On the four informative datasets, QAC[Dir] (red) improves on the uncalibrated model (Vanilla) in AURC and error-detection AUROC. On Brier, ECE and NLL it is the best method on ChaosNLI (ΔBrier ≈ -0.17, ΔNLL ≈ -0.81) but ties or slightly trails isotonic regression on some metrics elsewhere. It gives no gain on Audio MUSAN and is worse on Camelyon17, the two datasets whose quality score is computed from the input.

Table 8: NLL (lower=better) on the seven kept datasets, 10 seeds; $\pm$ = SE. Locked headline = **QAC[Dir]**; the same 5 SOTA plain calibrators as Tab. 2 are reported. Bold = best method per dataset.

Dataset	Vanilla	TS	Iso	BBQ	TvA-Iso	1-step Q	QAC[Dir]
E_toy†	$0.3384_{\pm 0.0024}$	$0.3392_{\pm 0.0025}$	$0.3655_{\pm 0.0054}$	$0.3473_{\pm 0.0027}$	$0.3636_{\pm 0.0078}$	$0.3589_{\pm 0.0101}$	$\mathbf{0.3108}_{\pm 0.0020}$
ChaosNLI†	$1.4740_{\pm 0.0190}$	$0.8443_{\pm 0.0101}$	$0.7209_{\pm 0.0171}$	$0.7521_{\pm 0.0045}$	$0.8119_{\pm 0.0058}$	$0.8603_{\pm 0.0145}$	$\mathbf{0.6625}_{\pm 0.0121}$
CIFAR-10H†	$0.2139_{\pm 0.0039}$	$\mathbf{0.1856}_{\pm 0.0029}$	$0.2331_{\pm 0.0075}$	$0.2675_{\pm 0.0038}$	$0.2678_{\pm 0.0043}$	$0.2827_{\pm 0.0077}$	$0.2200_{\pm 0.0041}$
CIFAR-10N†	$0.4083_{\pm 0.0065}$	$\mathbf{0.3481}_{\pm 0.0048}$	$0.3681_{\pm 0.0057}$	$0.4309_{\pm 0.0060}$	$0.4543_{\pm 0.0056}$	$0.4647_{\pm 0.0084}$	$0.3613_{\pm 0.0083}$
eICU	$\mathbf{0.2728}_{\pm 0.0034}$	$0.2731_{\pm 0.0034}$	$0.2757_{\pm 0.0029}$	$0.2742_{\pm 0.0031}$	$0.2741_{\pm 0.0029}$	$0.2803_{\pm 0.0032}$	$\mathbf{0.2728}_{\pm 0.0029}$
Audio MUSAN	$0.2194_{\pm 0.0033}$	$\mathbf{0.1902}_{\pm 0.0043}$	$0.3007_{\pm 0.0090}$	$0.2578_{\pm 0.0051}$	$0.2600_{\pm 0.0052}$	$0.2788_{\pm 0.0074}$	$0.3229_{\pm 0.0085}$
Camelyon17	$0.5188_{\pm 0.0181}$	$0.3860_{\pm 0.0121}$	$0.5668_{\pm 0.0307}$	$0.4955_{\pm 0.0135}$	$\mathbf{0.3513}_{\pm 0.0101}$	$0.3624_{\pm 0.0090}$	$0.5040_{\pm 0.0164}$

F.2 Per-cell quality effect against constant q

Tables 11 and 12 give the effect of the real quality score for each dataset and base calibrator.

G Datasets whose quality is computed from the input

On two datasets the quality score is computed from the model's input. On **Audio MUSAN** ($\rho_q = 0.026$), q is the measured noise level of the clip, and QAC[Dir] has the same Brier score as the uncalibrated model. On **Camelyon17** ($\rho_q = -0.078$), q is computed from the image, and QAC has a higher Brier score than the standard calibrators. The diagnostic flags both datasets. Table 13 and Figure 5 report the per-method Brier on both.

Table 9: AURC (lower=better) on the seven kept datasets, 10 seeds; $\pm$ = SE. Locked headline = **QAC[Dir]**; the same 5 SOTA plain calibrators as Tab. 2 are reported. Bold = best method per dataset.

Dataset	Vanilla	TS	Iso	BBQ	TvA-Iso	1-step Q	QAC[Dir]
E_toy†	$0.0551_{\pm 0.0009}$	$0.0551_{\pm 0.0009}$	$0.0559_{\pm 0.0012}$	$0.0618_{\pm 0.0016}$	$0.0594_{\pm 0.0018}$	$0.0435_{\pm 0.0015}$	$\mathbf{0.0385}_{\pm 0.0006}$
ChaosNLI†	$0.2308_{\pm 0.0041}$	$0.2321_{\pm 0.0041}$	$0.1884_{\pm 0.0044}$	$0.2453_{\pm 0.0137}$	$0.2290_{\pm 0.0037}$	$0.1774_{\pm 0.0031}$	$\mathbf{0.1414}_{\pm 0.0028}$
CIFAR-10H†	$0.0067_{\pm 0.0003}$	$0.0069_{\pm 0.0002}$	$0.0071_{\pm 0.0002}$	$0.0333_{\pm 0.0010}$	$0.0069_{\pm 0.0003}$	$0.0064_{\pm 0.0003}$	$\mathbf{0.0062}_{\pm 0.0002}$
CIFAR-10N†	$0.0205_{\pm 0.0004}$	$0.0209_{\pm 0.0004}$	$0.0206_{\pm 0.0004}$	$0.0505_{\pm 0.0013}$	$0.0210_{\pm 0.0005}$	$0.0169_{\pm 0.0004}$	$\mathbf{0.0153}_{\pm 0.0004}$
eICU	$\mathbf{0.0387}_{\pm 0.0009}$	$\mathbf{0.0387}_{\pm 0.0009}$	$0.0390_{\pm 0.0010}$	$0.0429_{\pm 0.0010}$	$0.0391_{\pm 0.0009}$	$0.0397_{\pm 0.0008}$	$0.0388_{\pm 0.0008}$
Audio MUSAN	$0.0074_{\pm 0.0003}$	$\mathbf{0.0065}_{\pm 0.0003}$	$0.0079_{\pm 0.0004}$	$0.0161_{\pm 0.0007}$	$0.0076_{\pm 0.0004}$	$0.0081_{\pm 0.0004}$	$0.0089_{\pm 0.0004}$
Camelyon17	$\mathbf{0.0459}_{\pm 0.0019}$	$\mathbf{0.0459}_{\pm 0.0019}$	$0.0685_{\pm 0.0025}$	$0.1001_{\pm 0.0042}$	$0.0462_{\pm 0.0019}$	$0.0478_{\pm 0.0013}$	$0.0712_{\pm 0.0031}$

Table 10: Error-detection AUROC (higher=better) on the seven kept datasets, 10 seeds; $\pm$ = SE. Locked headline = **QAC[Dir]**; the same 5 SOTA plain calibrators as Tab. 2 are reported. Bold = best method per dataset.

Dataset	Vanilla	TS	Iso	BBQ	TvA-Iso	1-step Q	QAC[Dir]
E_toy†	$0.7796_{\pm 0.0026}$	$0.7796_{\pm 0.0026}$	$0.7797_{\pm 0.0028}$	$0.7652_{\pm 0.0042}$	$0.7771_{\pm 0.0043}$	$0.8287_{\pm 0.0048}$	$\mathbf{0.8327}_{\pm 0.0028}$
ChaosNLI†	$0.6088_{\pm 0.0037}$	$0.6086_{\pm 0.0035}$	$0.6661_{\pm 0.0038}$	$0.5914_{\pm 0.0143}$	$0.6151_{\pm 0.0025}$	$0.6991_{\pm 0.0020}$	$\mathbf{0.7434}_{\pm 0.0020}$
CIFAR-10H†	$0.9295_{\pm 0.0021}$	$0.9275_{\pm 0.0021}$	$0.9272_{\pm 0.0022}$	$0.7464_{\pm 0.0042}$	$0.9283_{\pm 0.0022}$	$0.9364_{\pm 0.0016}$	$\mathbf{0.9390}_{\pm 0.0011}$
CIFAR-10N†	$0.8889_{\pm 0.0018}$	$0.8877_{\pm 0.0019}$	$0.8871_{\pm 0.0018}$	$0.7744_{\pm 0.0029}$	$0.8879_{\pm 0.0018}$	$0.9085_{\pm 0.0014}$	$\mathbf{0.9115}_{\pm 0.0018}$
eICU	$0.7497_{\pm 0.0028}$	$0.7497_{\pm 0.0028}$	$0.7485_{\pm 0.0030}$	$0.7413_{\pm 0.0027}$	$0.7485_{\pm 0.0030}$	$0.7467_{\pm 0.0022}$	$\mathbf{0.7498}_{\pm 0.0029}$
Audio MUSAN	$0.8734_{\pm 0.0039}$	$0.8897_{\pm 0.0036}$	$0.8880_{\pm 0.0041}$	$0.8085_{\pm 0.0080}$	$0.8723_{\pm 0.0039}$	$0.8703_{\pm 0.0057}$	$\mathbf{0.9007}_{\pm 0.0042}$
Camelyon17	$\mathbf{0.7994}_{\pm 0.0040}$	$\mathbf{0.7994}_{\pm 0.0040}$	$0.7899_{\pm 0.0069}$	$0.7299_{\pm 0.0088}$	$0.7991_{\pm 0.0041}$	$0.7936_{\pm 0.0018}$	$0.7818_{\pm 0.0057}$

Table 11: Per (dataset $\times$ base) quality effect on Brier, $\mathrm{Brier}(\mathrm{QAC}[C; q^{\text{const}}]) - \mathrm{Brier}(\mathrm{QAC}[C; q^{\text{real}}])$, on the seven datasets (positive favours real q). * / †: 95% bootstrap interval over ten seeds strictly above / below 0; bold = largest per row. Bases: Isot=isotonic OvR, BBQ, Beta, TvA=TvA-isotonic (Le Coz et al., 2024), Dir=Dirichlet, Spl=spline, Plt=Platt, HBin=histogram binning, GTS=global temperature scaling. Real q improves on all nine bases on the four informative datasets, ties on all nine on eICU, and is worse on all nine on Audio MUSAN and Camelyon17, whose quality is computed from the input.

Dataset	Isot	BBQ	Beta	TvA	Dir	Spl	Plt	HBin	GTS
ChaosNLI	$+0.0139^*$	$\mathbf{+0.0198}^*$	$+0.0148^*$	$+0.0159^*$	$+0.0138^*$	$+0.0145^*$	$+0.0141^*$	$+0.0186^*$	$+0.0149^*$
E_toy	$+0.0108^*$	$+0.0112^*$	$+0.0108^*$	$+0.0112^*$	$+0.0110^*$	$+0.0108^*$	$+0.0108^*$	$\mathbf{+0.0112}^*$	$+0.0108^*$
CIFAR-10H	$+0.0011^*$	$+0.0016^*$	$+0.0012^*$	$+0.0007^*$	$+0.0013^*$	$+0.0012^*$	$+0.0012^*$	$\mathbf{+0.0017}^*$	$+0.0012^*$
CIFAR-10N	$+0.0035^*$	$\mathbf{+0.0059}^*$	$+0.0036^*$	$+0.0043^*$	$+0.0039^*$	$+0.0038^*$	$+0.0036^*$	$+0.0056^*$	$+0.0036^*$
eICU	$+0.0000$	-0.0000	-0.0000	$\mathbf{+0.0001}$	-0.0000	-0.0000	$+0.0000$	-0.0001	-0.0000
Audio MUSAN	$-0.0004^\dagger$	$-0.0003^\dagger$	$-0.0009^\dagger$	$-0.0004^\dagger$	$\mathbf{-0.0002}^\dagger$	$-0.0010^\dagger$	$-0.0009^\dagger$	$-0.0002^\dagger$	$-0.0009^\dagger$
Camelyon17	$-0.0020^\dagger$	$-0.0019^\dagger$	$-0.0019^\dagger$	$-0.0019^\dagger$	$-0.0020^\dagger$	$-0.0023^\dagger$	$-0.0022^\dagger$	$\mathbf{-0.0016}^\dagger$	$-0.0019^\dagger$

Table 12: Per (dataset $\times$ base) quality effect on error-detection AUROC, $\mathrm{AUROC}(\mathrm{QAC}[C; q^{\text{real}}]) - \mathrm{AUROC}(\mathrm{QAC}[C; q^{\text{const}}])$ (positive favours real q). Markers as in Table 11. Real q improves on all nine bases on the four informative datasets. On Audio MUSAN and Camelyon17 the effect is mostly negative or null; the exceptions are the two binning bases, BBQ and histogram binning.

Dataset	Isot	BBQ	Beta	TvA	Dir	Spl	Plt	HBin	GTS
ChaosNLI	$+0.0495^*$	$+0.0810^*$	$+0.0520^*$	$+0.0522^*$	$+0.0504^*$	$+0.0524^*$	$+0.0515^*$	$\mathbf{+0.0876}^*$	$+0.0523^*$
E_toy	$+0.0536^*$	$+0.0553^*$	$+0.0510^*$	$+0.0551^*$	$+0.0536^*$	$+0.0513^*$	$+0.0510^*$	$\mathbf{+0.0558}^*$	$+0.0510^*$
CIFAR-10H	$+0.0128^*$	$\mathbf{+0.0555}^*$	$+0.0110^*$	$+0.0114^*$	$+0.0102^*$	$+0.0121^*$	$+0.0110^*$	$+0.0553^*$	$+0.0111^*$
CIFAR-10N	$+0.0207^*$	$\mathbf{+0.0680}^*$	$+0.0236^*$	$+0.0253^*$	$+0.0205^*$	$+0.0218^*$	$+0.0236^*$	$+0.0630^*$	$+0.0229^*$
eICU	$+0.0016^*$	$+0.0040^*$	$+0.0009^*$	$+0.0003$	$+0.0001$	-0.0001	$+0.0006$	$\mathbf{+0.0083}^*$	$+0.0005$
Audio MUSAN	$-0.0053^\dagger$	-0.0043	-0.0029	$-0.0167^\dagger$	$-0.0054^\dagger$	$-0.0036^\dagger$	-0.0024	$\mathbf{+0.0086}^*$	-0.0036
Camelyon17	$-0.0074^\dagger$	$+0.0176^*$	$-0.0081^\dagger$	-0.0037	-0.0037	$-0.0099^\dagger$	-0.0065	$\mathbf{+0.0355}^*$	$-0.0070^\dagger$

H Q-DESIGN ABLATIONS AND CROSS-SITE ANALYSES

This appendix details the families of Q-design ablations and cross-site analyses behind Section 4.5 of the main text (Section 4.3–Section 4.5). All experiments use the paired QAC-protocol (identical cal/test splits across method variants; 9 base calibrators; 10 seeds; paired bootstrap 95% CIs).

Table 13: Negative-control datasets (Brier). Audio MUSAN: $\rho_q = 0.026$ (uninformative $\Rightarrow$ predicted null; observed: QAC ties Vanilla, no harm). Camelyon17: $\rho_q = -0.078$ (quality computed from the image $\Rightarrow$ predicted failure; observed: QAC under-performs TvA-Iso). 10 seeds; $\pm$ = SE. Bold = best method per dataset.

Method	Audio MUSAN	Camelyon17
Vanilla	$0.0758_{\pm 0.0011}$	$0.2326_{\pm 0.0054}$
TS	$\mathbf{0.0647}_{\pm 0.0012}$	$0.2166_{\pm 0.0052}$
Iso	$0.0660_{\pm 0.0012}$	$0.2717_{\pm 0.0038}$
BBQ	$0.0700_{\pm 0.0014}$	$0.2712_{\pm 0.0038}$
TvA-Iso	$0.0684_{\pm 0.0012}$	$0.2082_{\pm 0.0049}$
Q-CTS	$0.0689_{\pm 0.0013}$	$\mathbf{0.2075}_{\pm 0.0044}$
QAC[Dir]	$0.0751_{\pm 0.0011}$	$0.2700_{\pm 0.0041}$

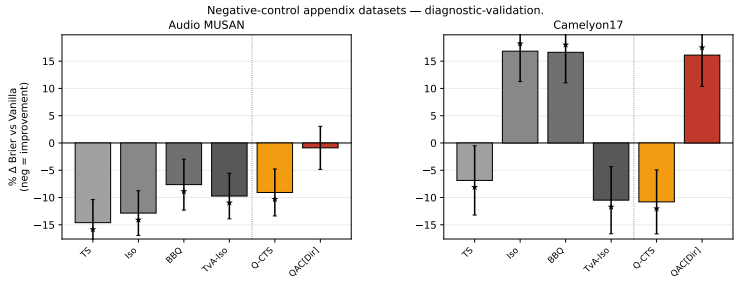

Figure 5: Brier score on the two datasets whose quality score is computed from the input. On Audio MUSAN, QAC[Dir] (red) ties the uncalibrated model (Vanilla); on Camelyon17 it is worse than TvA-Iso. 10 seeds; error bars show the SE.

H.1 Single-source vs. multi-source input (E_toy)

When the data pipeline produces multiple correlated sources of the same underlying signal, the single-vs-multi-source choice matters. Multi-source training incorporates the cross-source information into $\sigma(C(X))$ directly: q then becomes redundant in the sense of Corollary 3. In the single-source regime q is external and reducible (Corollary 2). On E_toy the empirical signature is unambiguous: in unisource mode the controlled corruption serves as an external quality signal and QAC delivers the OAU gap; in multisource mode the base classifier already captures the cross-source information and QAC provides limited additional gain (Corollary 3 regime).

H.2 Datapoint-level vs. cohort-level q (eICU granularity)

DQ ontologies operate at multiple aggregation levels (Kahn et al., 2016). We test which level matters on eICU, where the same DQ vector is available at per-patient and per-hospital granularity. Figure 6 shows the result: deploy at the finest aggregation that varies sample-by-sample.

H.3 Multi-dimensional q reduction (eICU)

A practitioner with a multi-dim q vector faces a choice: pick a single best dimension, reduce to a scalar, or pass the vector through. We report this on eICU (the only multi-dim-q dataset in the kept scope: 5-d Kahn–Weiskopf DQ vector) via the calibration-set-size sweep below, which jointly characterises the trade-off between q-dimensionality and cal-set size. Recommendation: *vector q when the precondition holds; scalar reduction is a robust fallback when cal is small.*

Calibration-set-size dependence. The optimal q-dimensionality also depends on the calibration-set size: orthogonal q-dims add useful capacity only when the calibrator has enough samples to estimate that capacity reliably. Figure 7 sweeps both axes on eICU, holding the nine bases and ten seeds fixed.

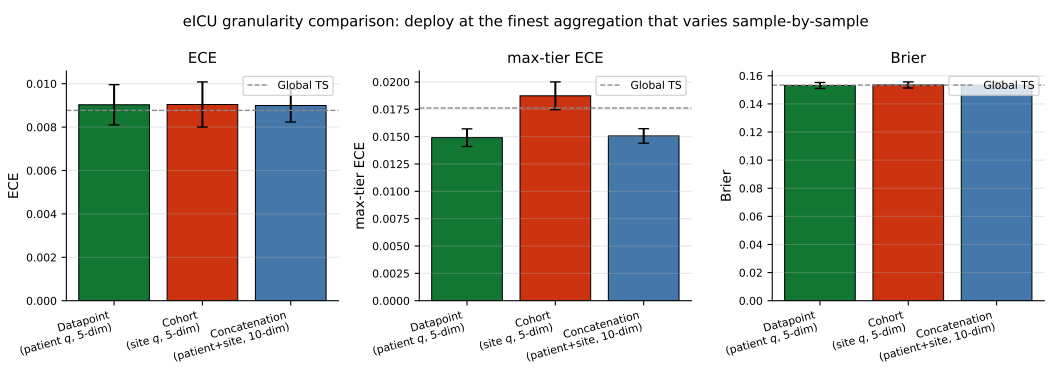

Figure 6: Datapoint-level vs. cohort-level q on eICU 41-hospital held-out cross-site protocol. Three panels: top-label ECE, max-per-tier ECE (the relevant calibration safety metric under site shift), Brier. Datapoint-level patient q (5-dim, green) cuts max-tier ECE 15.4% over Global TS; cohort-level site q (5-dim averaged within hospital, red) under-performs Global TS; the 10-dim concatenation matches datapoint alone. The cohort signal is too coarse to track sample-level error. Mean over 10 seeds; SE error bars.

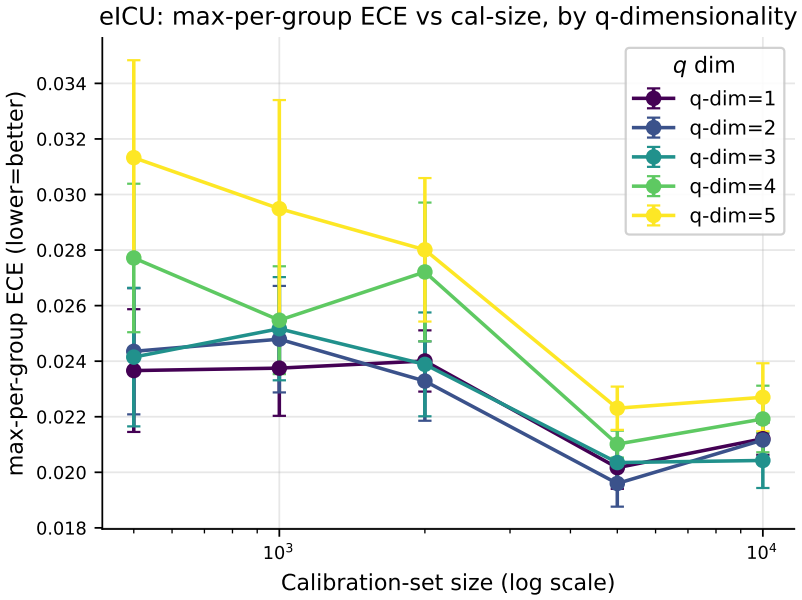

Figure 7: eICU max-per-group ECE as a function of calibration-set size (log axis) and q-dimensionality $\{1, 2, 3, 4, 5\}$. Five-dim q substantially under-performs single-dim q at $n_{\text{cal}} = 500$; the gap closes by $n_{\text{cal}} = 5000$. Practitioner guidance: small cal sets favour low q-dim; add orthogonal dims as cal-set size grows. Mean over 10 seeds, SE error bars.

H.4 Q-CTS: A ONE-STAGE VARIANT FROM OUR EARLIER EXPERIMENTS

Before the two-stage form of Section 3, we studied a one-stage calibrator, Q-conditional temperature scaling (Q-CTS). It divides the logits by a temperature that depends on the quality score, $\text{softmax}(z/T(q))$, where $T(q) > 0$ is a small neural network (one hidden layer) fitted by minimising the log-loss on the calibration set. Q-CTS can only make all classes more or less confident together, and its temperature cannot depend on the model's output. The two-stage form removes both limits and can reuse any existing calibrator.

Table 14 gives the Brier score of Q-CTS on the five main datasets, next to QAC[Dir] from Table 2. Q-CTS is worse than QAC[Dir] on every dataset. Under the eICU hospital shift (Table 15), Q-CTS and QAC reach a similar worst-tier ECE. Figure 8 compares seven Q-CTS parameterisations on eICU.

method	E_toy	ChaosNLI	CIFAR-10H	CIFAR-10N	eICU
Q-CTS	0.1970	0.4399	0.0947	0.1632	0.1542
QAC[Dir]	0.1956	0.3800	0.0870	0.1406	0.1531

Table 14: Brier score (10 seeds) of the one-stage Q-CTS and of QAC[Dir].

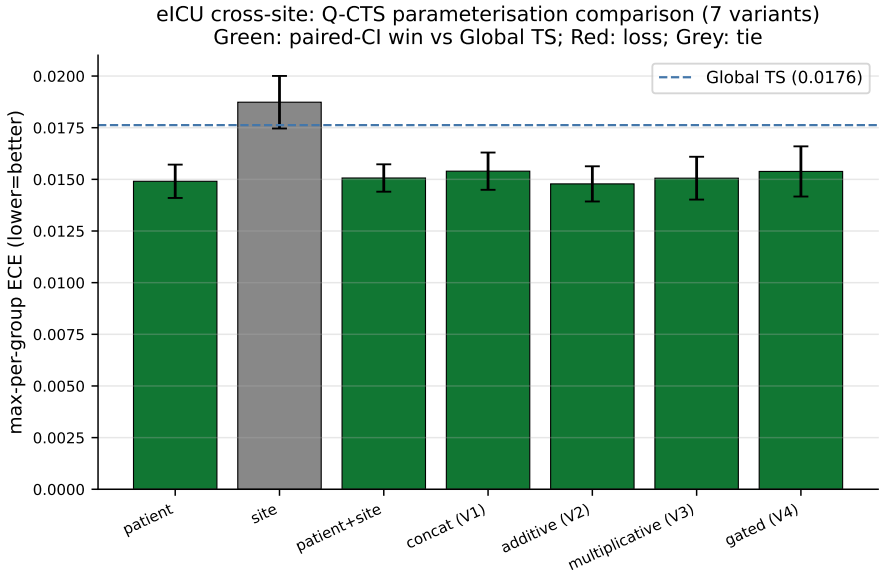

Figure 8: Comparison of seven 1-step Q-CTS parameterisations on eICU 41-hospital cross-site held-out protocol (max-per-group ECE). Six of seven cut ECE 12–16% over Global TS; the exception is Q-CTS-site (cohort identifier alone), which fails to extrapolate to held-out sites and slightly underperforms TS. The patient and additive parameterisations are indistinguishable within random variation across seeds and are the 1-step variants used elsewhere in the paper. The choice of parameterisation matters less than the granularity of the q signal (Appendix H.2, Section H.2). Mean over 10 seeds; SE error bars.

H.5 Q-NOISE ROBUSTNESS

Figure 9 shows the Brier score as noise is added to q.

H.6 CROSS-SITE / CROSS-DOMAIN DETAIL

Figure 10 shows the earlier cross-site results with Q-CTS and the Camelyon17 per-base effects; Table 15 gives the results for QAC.

QAC **under hospital shift.** All methods in Table 15 are fitted post hoc on the same saved logits: the base model is trained on the training hospitals, calibrators on a calibration split of the same hospitals, and all are tested on the 30% of hospitals held out in each seed. The metric is the largest top-label ECE over three tiers of the mean patient-quality score, with tier edges from the calibration split. QAC uses the default scalar q of the main grid; the five-dimensional version is shown for reference.

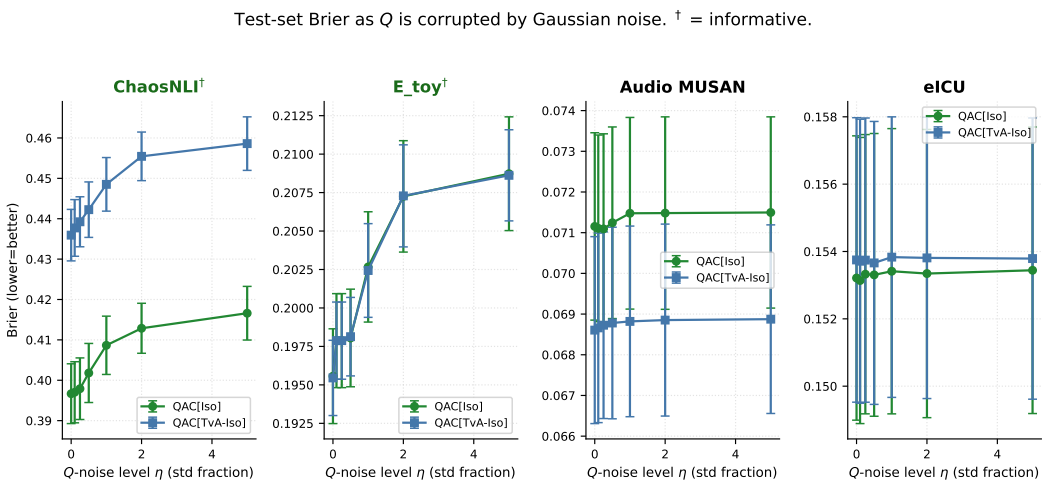

Figure 9: Q-noise robustness sweep on the kept datasets. We add Gaussian noise to Q at $\eta \cdot \sigma_q$ where $\eta \in \{0, 0.10, 0.25, 0.50, 1.00, 2.00, 5.00\}$, then run the full post-hoc QAC pipeline. On informative datasets (†, ChaosNLI and E_toy) the test-set Brier of QAC degrades smoothly with η; half of the gain is lost around $\eta \approx 1$ (amplitude SNR ≈ 1), and at $\eta = 5$ QAC[Iso] is about 1% worse than the no-Q floor (E_toy 0.2087 vs 0.2064; ChaosNLI 0.4166 vs 0.4115). On uninformative / boundary datasets (eICU, Audio MUSAN) the curves are flat — Q noise has nothing to corrupt because the framework's gain on those datasets is already small (Corollary 3 regime). Two QAC variants shown: QAC[Iso] and QAC[TvA−Iso].

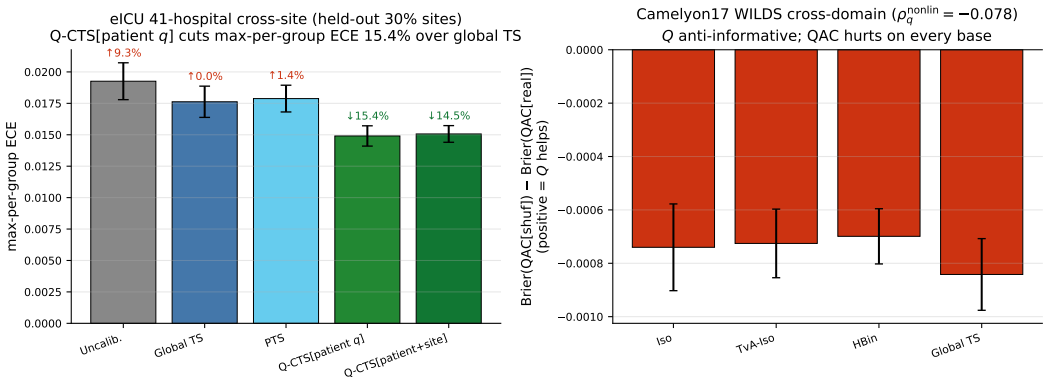

Figure 10: Cross-site detail. *Left*: eICU 41-hospital cross-site protocol — held-out 30% of sites at test, max-per-group ECE on the held-out tier (groups are three quantile tiers of patient quality). Q-CTS with patient-level q and with patient and site q both lower the worst-tier ECE by about 15% relative to global TS in this earlier run, which trained its own base models. *Right*: Camelyon17, effect of real versus shuffled q on the Brier score for four bases, $\mathrm{Brier}(\mathrm{QAC}[\mathrm{shuf}]) - \mathrm{Brier}(\mathrm{QAC}[\mathrm{real}])$. The effect is negative for all four: q is computed from the image and does not help, as flagged by $\rho_q = -0.078$. Mean over 10 seeds, SE error bars.

H.7 Training-time vs. post-hoc integration

Corollary 5 formalises that when Q is observed only at calibration/inference, post-hoc q-conditional calibration is strictly better than training-time methods that cannot use q. Figure 11 shows Q-LS progressively under-performs vanilla as ε increases.

method	worst-tier ECE	vs. global TS	seeds better than TS
global TS	0.0209 ± 0.0012	–	–
Dirichlet	0.0172 ± 0.0012	-18%	$7/10$
QAC[Dir], constant q	0.0182 ± 0.0011	-13%	$7/10$
QAC[Dir], default q	0.0164 ± 0.0009	-22%	$9/10$
QAC[Dir], 5-dim q	0.0157 ± 0.0007	-25%	$8/10$
Q-CTS, 5-dim q	0.0160 ± 0.0008	-24%	$8/10$

Table 15: eICU under hospital shift (mean $\pm$ SE over 10 seeds). Real versus constant q in QAC[Dir]: -0.0018 (default q, $7/10$ seeds, paired t-test $p = 0.17$) and -0.0026 (5-dim q, $7/10$, $p = 0.06$). Brier is 0.153 for every method.

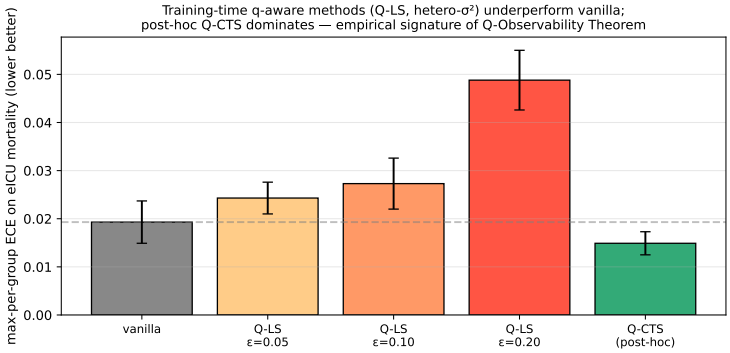

Figure 11: Training-time q-aware methods (Q-Label-Smoothing, hetero-σ^2) under-perform vanilla on eICU max-per-group ECE. Post-hoc methods (the one-stage Q-CTS[patient q] shown) dominate, empirical signature of Corollary 5 (necessity of post-hoc when Q is deployment-only).

H.8 Cross-domain ρ_q diagnostic for Q-shift detection

A complementary diagnostic, useful when calibration and deployment data may come from different distributions, fits $\mathrm{HGBT}(q \to \text{per-sample NLL})$ on the calibration set and evaluates R^2 on a held-out test set drawn from the deployment distribution. The $q \to$ error relationship transfers cleanly on CIFAR-10H ($+0.043 \to +0.050$) and CIFAR-10N ($+0.091 \to +0.105$), where calibration and deployment are i.i.d. samples from the same distribution. On Camelyon17, cal $R^2 = +0.033$ but cross-domain $R^2 \approx 0$ on test, the $q \to$ error mapping does not transfer across the WILDS site boundary, and the framework picks up an anti-correlation. We recommend computing this cross-domain test in deployments where the calibration and inference distributions may differ.

H.9 Self-derived Q as a Corollary 3 boundary test

If Q is fully determined by $C(X)$, the framework is null (Corollary 3). The cleanest empirical test of this boundary is to set $Q := \max_k C(X)_k$ (the model's own confidence) and ask whether QAC[$C; q^{\text{self}}$] still improves over C. We rerun the QAC pipeline with this self-derived Q on the four kept datasets where calibration logits are stored alongside an external q candidate (E_toy, ChaosNLI, Audio MUSAN, eICU; `results/qac_cheap_q/`). The genuine effect collapses to within $\pm 5 \times 10^{-4}$ Brier on every dataset (max over seeds; see `scripts/baselines/aggregate_qac_cheap_q.py`) and the diagnostic ρ_q on the self-derived Q saturates at the dataset-level miscalibration of C, an order of magnitude below the external-Q values reported in the main paper (Fig. 12). This empirically confirms Corollary 3: the framework extracts no value from Q when $Q \subseteq \sigma(C(X))$. The practical implication is that QAC is not a recalibration trick in disguise: the gain reported on the four informative datasets requires Q to carry information X does not already encode.

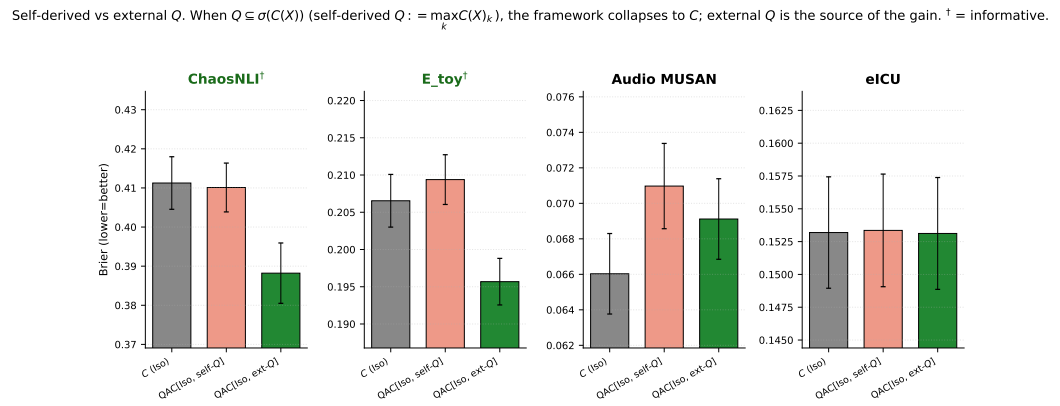

Figure 12: Self-derived Q as Corollary 3 boundary test. Per-dataset Brier of three configurations: C alone (grey), $\mathrm{QAC}[C; q^{\mathrm{self}}]$ with $q^{\mathrm{self}} = \max_k C(X)_k$ (orange, self-derived from C), and $\mathrm{QAC}[C; q^{\mathrm{ext}}]$ with the external Q used elsewhere in the paper (green). On the four informative datasets external Q delivers paired-CI Brier improvements; the self-derived Q collapses to within $\pm 5 \times 10^{-4}$ of C on every dataset, confirming the corollary numerically.

I Operational checklist

The following checks can be run on the calibration set before QAC is used on a new dataset. They assume that a quality score already exists. How to obtain one, by instrumenting the data pipeline or by collecting several annotations per sample, is outside the scope of this checklist. Each check needs one pass over the calibration set; the first takes 0.3 seconds on the 17,436 eICU calibration samples.

First, compute ρ_q, the five-fold cross-validated R^2 of a gradient-boosted regressor that predicts the errors of C from q (Section 4.2). In our datasets, QAC improved calibration when $\rho_q \geq 0.05$ and not below. This threshold was chosen on these datasets, and it applies only when calibration and test data come from the same distribution.

Second, check whether C leaves room for improvement. If $\mathrm{Reliability}(C)$ is close to zero, Corollary 2 limits the possible gain to the change in resolution.

Third, check whether q can be computed from the input or from $C(X)$. If it can, q adds no information (Corollary 3). The increase ΔR^2 of Table 1 measures this. On Audio MUSAN and Camelyon17, where q is computed from the input, it is zero or negative, and QAC with real q was slightly worse than with a constant q.

Fourth, if a shift between calibration and deployment data is expected, check whether the relation between q and the errors of C holds on data from the new setting, with the cross-domain R^2 test of Section H.8.

J Routing / triage interpretation

A natural intuition is that q should also serve directly as a triage signal: route low-q samples to a human reviewer. We tested this on every dataset where both q-routing and confidence-based routing have meaningful coverage. The empirical finding is clarifying rather than supportive: confidence-based routing on QAC-calibrated probabilities strictly dominates direct q-based routing on every informative dataset (e.g., on ChaosNLI[TvA-iso] at 50%-coverage routing, confidence-routing recovers $+0.081$ in retained accuracy versus q-routing's $+0.005$). Interpretation: *q's value is leveraged through the calibration step*; once the framework has produced a quality-aware calibrated confidence, that confidence already incorporates q's information for downstream triage. The framework benefits triage by giving routine confidence-based triage a more trustworthy confidence to act on, not by introducing q as a separate triage axis.

K ADDITIONAL FIGURES REFERENCED FROM SECTION 4

This section holds supporting figures: the realised QAC gain against the diagnostic ρ_q (Figure 13), the Bröcker decomposition (Figure 14), the shuffled-q control (Figure 15) and reliability diagrams (Figure 16). The body retains the master selective-classification figure (Fig. 2) and the headline Brier table (Table 2).

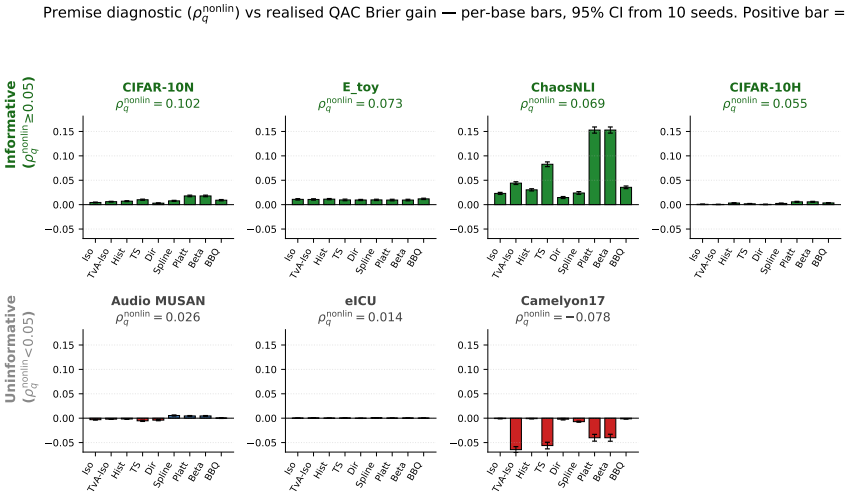

Figure 13: Per-dataset Brier-gain panels grouped by the precondition diagnostic on the seven kept datasets. *Top row*: the four datasets above the screening threshold ($\rho_q \geq 0.05$, *supposed-to-work*) — each panel shows the realised Brier(C) − Brier(QAC[C]) per base calibrator with 95% CIs from 10 seeds; bars are positive across all 9 bases. *Bottom row*: the three remaining datasets — Audio MUSAN (uninformative-null, correct prediction), eICU (low marginal ρ but cross-site QAC improvement, the documented exception), Camelyon17 where the $Q \to Y$ relation fails to transfer cross-domain (red bars; Section 4.5).

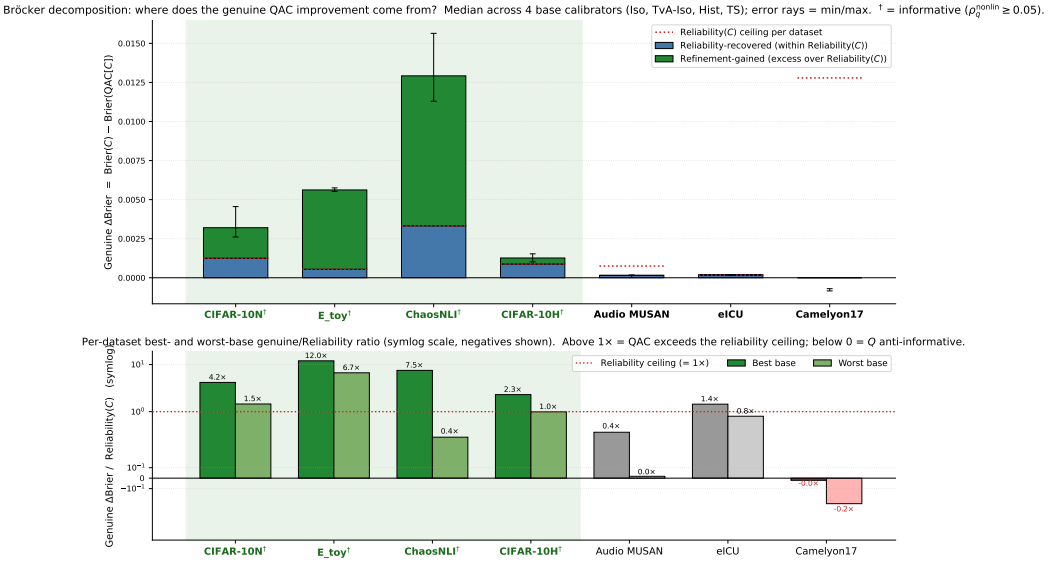

Figure 14: Bröcker decomposition of QAC's Brier improvement. *Top*: per-dataset stacked bars: blue = gain recovered within the reliability ceiling $\mathrm{Reliability}(C)$, green = additional gain in excess of the reliability ceiling. Median across 4 base calibrators (Iso, TvA-Iso, Hist, TS) with min/max error rays. Per-dataset $\mathrm{Reliability}(C)$ marked with red dotted segments. *Bottom*: per-dataset best- and worst-base genuine/Reliability ratios on log scale; ratio $> 1\times$ = framework exceeds the reliability ceiling. † marks the four informative datasets.

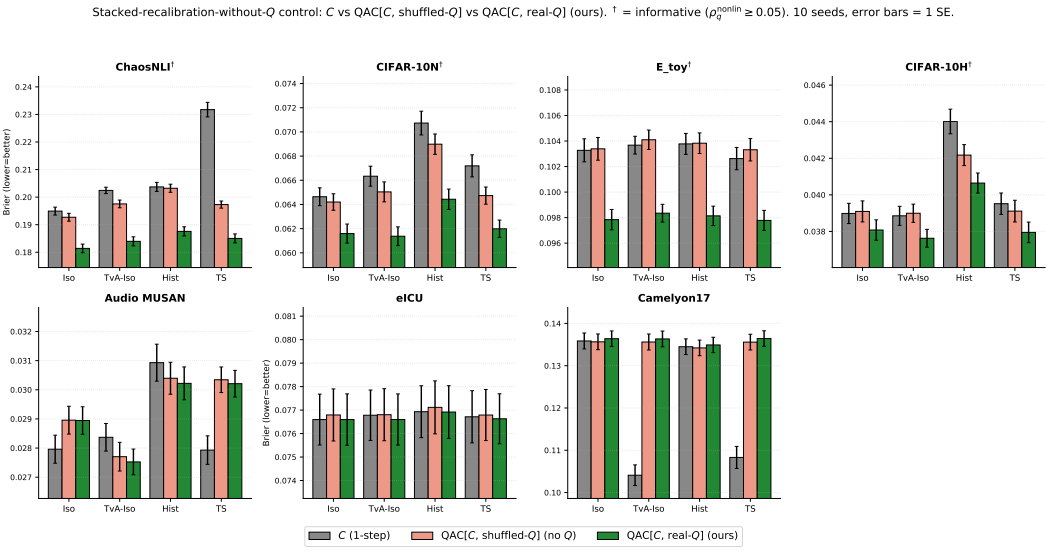

Figure 15: Stacked-no-Q controlled comparison: per (dataset, base) Brier of C vs. QAC[C, shuffled-Q] vs. QAC[C, real-Q] from a single paired protocol so the three are commensurable. Pink (shuffled Q) does not improve over grey (C); green (real Q) does, on the four informative datasets. Mean over 10 seeds; error bars = 1 SE.

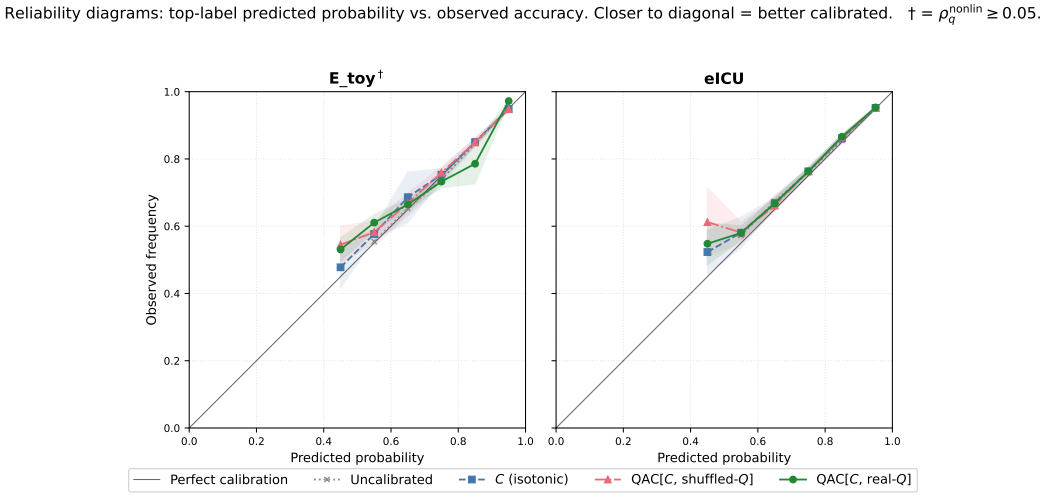

Figure 16: Reliability diagrams: top-label predicted probability vs. observed accuracy for the five datasets with raw logits archived. Closer to the diagonal = better calibrated. The green QAC[C, real-Q] curve dominates on E_toy (informative); on uninformative datasets the methods overlap. Mean over 10 seeds; 95% CI shading.