Life of Research · Published 17 November 2026 · DOI 10.5281/zenodo.24294850

Calibrated Against Which Label? What Temperature Scaling Certifies When Annotators Disagree, Why a Majority Vote and an Annotator Distribution Each Make the Other's Calibrated Model Look Wrong, and the Target a Calibration Claim Has to Declare

Pranay Mahendrakar
SONYTECH · ORCID 0009-0003-7224-029X
pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Post-hoc calibration is usually fitted and scored against one label per item, and where annotators disagree that label is a majority vote. A 2026 study argues that this is a structural error: temperature scaling fitted to votes settles on temperatures that are too low, looks calibrated against the votes, and stays overconfident against the distribution of annotator labels, with errors that grow with annotation entropy; the best calibrator it fits to the annotator distribution, Dirichlet-Soft, cuts true-label calibration error by 55 to 87 percent. In this analysis the author reviews and synthesises published results to show that the finding is correct and that its framing is one-sided. Calibration is a relation between a forecast and a named random variable, and under disagreement there are at least four candidates: the vote, the label of a randomly drawn annotator, the label of a member of a target population, and an outcome fixed by later evidence. Each result in this literature is calibrated against a different one. A model that reproduces the human distribution exactly is badly miscalibrated against the majority, by the same logic that makes vote-fitted temperature scaling miscalibrated against annotators. The annotator distribution is not a clean target either: re-collecting 100 labels per item changed the majority label on a quarter to a third of low-agreement NLI items; adjudication traced most disagreements in one retinal grading study to missed lesions and artifacts; and the two medical benchmarks in the 2026 study use synthetic annotators whose confusion depends only on the consensus class. Re-reading that study's tables, the author finds that when temperature scaling is refitted to annotator labels, ECE falls by 48 to 87 percent while the soft-target Brier score falls by 2 to 11 percent. The paper consolidates the measured values in Table 1 and Figure 1, states a target-declaration procedure as Algorithm 1, and specifies six tests that would say which target a deployment should be calibrated to and how much of the measured gap survives it.

Keywords: confidence calibration; temperature scaling; annotator disagreement; human label variation; proper scoring rules; soft labels; reference standard; uncertainty quantification

1. Introduction

A classifier is calibrated when the probability it attaches to an outcome matches how often that outcome occurs among the inputs to which it attaches that probability. The definition has two parts, and the second is easy to forget: the outcome is a random variable, and calibration is a property of a forecast relative to that variable. In most machine learning practice the variable is never named, because there is assumed to be only one. Each test item carries one label, and the label is taken as the true class.

Neural classifiers are typically overconfident against that label, and a one-parameter rescaling of the logits, temperature scaling, fixes much of the problem on standard image and text benchmarks (Guo et al., 2017). Later work found the picture more varied across architectures, with some recent non-convolutional models among the best calibrated out of the box (Minderer et al., 2021), proposed richer post-hoc maps (Kull et al., 2019), and documented that the standard binned estimator of expected calibration error (ECE) is biased and sensitive to its own design choices (Nixon et al., 2019; Kumar et al., 2019; Roelofs et al., 2022). A survey of classifier calibration consolidates this machinery (Silva Filho et al., 2023). Throughout, the target is one label per item.

That assumption fails wherever careful people disagree about the label. Disagreement on natural language inference items persists as more ratings are collected and as more context is given, and is not well modelled as noise (Pavlick & Kwiatkowski, 2019). When 100 labels per item were collected for SNLI and MNLI items selected for low agreement, the majority changed on 24.97 and 31.77 percent of them (Nie et al., 2020). Human labels on the CIFAR-10 test set form a distribution, not a point (Peterson et al., 2019). Expert readers of medical images disagree in ways that are partly resolvable and partly not (Krause et al., 2018; Stutz et al., 2023b). An entire research programme now argues that this variation should be preserved in data, modelling and evaluation rather than aggregated away (Aroyo & Welty, 2015; Uma et al., 2021; Basile et al., 2021; Plank, 2022).

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

Calibration is where the two literatures meet, and the meeting has produced a sharp 2026 result. Tao et al. (2026), the anchor of this analysis, argue that calibrators fitted on majority-voted labels "can appear well-calibrated under conventional evaluation yet remain substantially miscalibrated against the underlying annotator distribution". They give a simplified model in which temperature scaling fitted to votes is pulled toward temperatures that are too low, with error against the annotator distribution rising with annotation entropy, and they propose calibrators that fit the full annotator distribution with a proper scoring rule. On four benchmarks the best of these, Dirichlet calibration fitted to the soft target (Dirichlet-Soft), cuts true-label ECE by 55 to 87 percent relative to temperature scaling. A single-author study the same year reaches the same qualitative conclusion on CIFAR-10H and ChaosNLI across three model scales per modality (Dogah, 2026).

This paper agrees with the arithmetic and disputes the framing. The anchor treats a label drawn from the annotator distribution as the true label, and calls error against it true-label calibration error. That choice settles by vocabulary the question the evidence leaves open: which random variable a deployed model's confidence should forecast. In this analysis the author reviews and synthesises published results to show six things.

First, calibration under disagreement has at least four candidate targets, which this paper names MAJ (the label an aggregation rule assigns), ANN (the label of one annotator drawn from the labelling pool), POP (the label of one member of the population the system serves) and REF (an outcome fixed by evidence the annotators did not have). A forecast calibrated to one is in general miscalibrated to the others. The anchor's result is a statement about MAJ versus ANN (Section 2).

Second, the pathology runs both ways. Baan et al. (2022) showed that a classifier that outputs the human judgement distribution exactly is severely miscalibrated by ECE against the majority, worse than the uncalibrated model it is compared with. Vote-fitted temperature scaling failing against annotators and the annotator distribution failing against votes are the same theorem read from opposite ends (Sections 3 and 4).

Third, the annotator distribution mixes valid variation with error, noise and the composition of the labelling pool. Published re-annotations, error audits and an intra-annotator deconvolution put the non-variation share well above zero on the tasks where it has been measured (Section 5).

Fourth, in medicine a reference outcome often exists, and there the annotator distribution measures how hard the image is for readers, not how uncertain the diagnosis is. The anchor's two medical benchmarks use synthetic annotators whose confusion depends only on the consensus class, so on those benchmarks the annotator target is, by construction, a noise channel applied to the reference label (Section 6).

Fifth, for subjective tasks and for language models the annotator pool is itself a choice, and the evidence that one output distribution can serve as both a confidence score and a population forecast is weak (Sections 7 and 8).

Sixth, the size of the reported gain depends on the metric. Re-reading the anchor's tables, the large ECE reductions sit beside small changes in the strictly proper scores the anchor also reports (Section 9).

The paper consolidates the measured values in Table 1 and Figure 1 (Section 11), states a target-declaration procedure as Algorithm 1 (Section 12), sets out the strongest case against this reading (Section 13), and specifies six tests (Section 14). Section 15 states what is not known.

The scope is narrow. The paper does not dispute any number the anchor reports, and it agrees that calibration fitted to votes misstates confidence against annotator labels. Its claim concerns what that result licenses: which target a deployment should calibrate to, and how much of the measured gap is ambiguity rather than a change of target.

2. Calibration is a relation to a named variable

2.1 Four targets

Let x be an input and let p(x) be a model's predicted distribution over K classes. Calibration in its standard form says that among inputs where the model's top-class probability is c, the top class occurs with frequency c. "Occurs" requires a variable Y whose value is observed or drawn for each input. Under disagreement, the candidates for Y are these.

MAJ. Y is the label assigned by a fixed aggregation rule, most often majority or plurality vote over the annotations an item received. The variable is deterministic given the annotations. It depends on how many annotators were asked, which ones, and how ties are broken.

ANN. Y is the label of one annotator drawn at random from the pool that labelled the data. Its distribution given x is the annotator distribution q(x), estimated by the empirical label frequencies. This is the target the anchor calls the true label.

POP. Y is the label that one member of a specified population would give: the users of a moderation system, the members of a group a statement targets, the clinicians who will read the model's output. POP equals ANN only when the labelling pool is a representative sample of that population.

REF. Y is an outcome determined by evidence the annotators did not have at labelling time: a biopsy, a follow-up visit, an adjudication by a panel with more time and information. When REF exists, disagreement

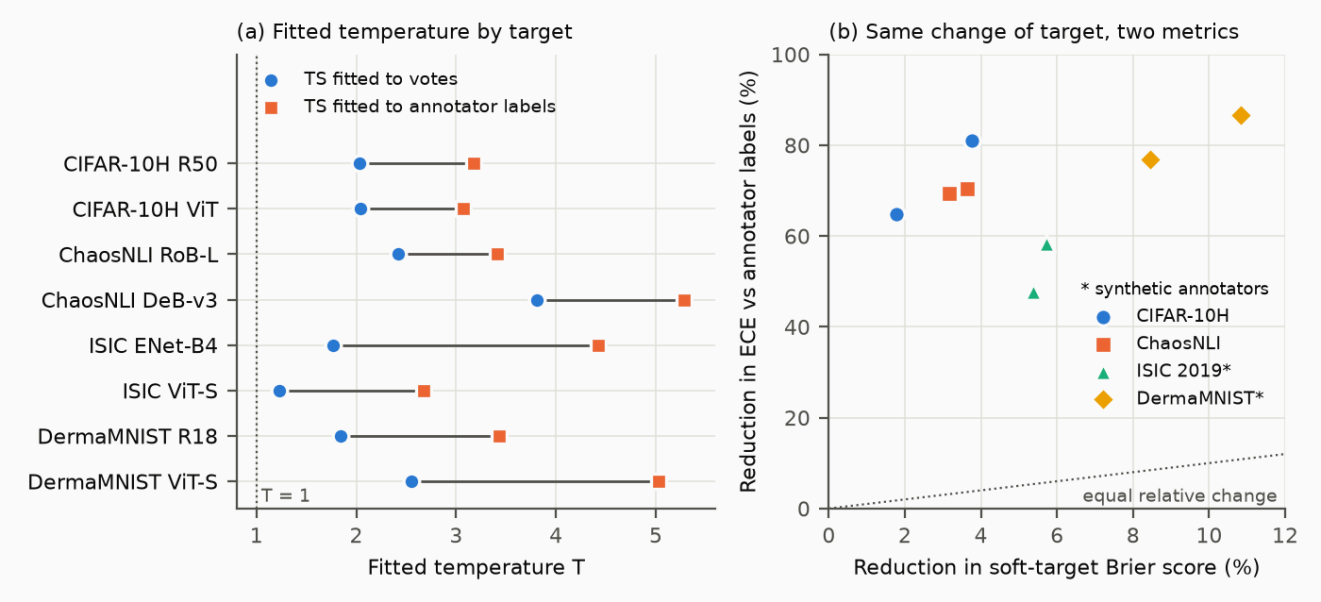


Figure 1. Values from Tao et al. (2026), Tables II and III. (a) Temperature fitted by vote-target temperature scaling (TS, circles) and by annotator-target temperature scaling (SLTS, squares) on each benchmark and backbone; every fitted value exceeds 1. (b) For the same eight settings, the relative reduction in ECE against annotator labels when the target changes from votes to annotators, plotted against the relative reduction in the soft-target Brier score, computed by the author from the published values. Starred benchmarks use synthetic, class-conditional annotators.

among annotators reflects their uncertainty about it, not uncertainty in it.

These are not four estimates of one quantity. They are four quantities. A deployment that scores the model against a reference standard needs REF calibration; a deployment that reports the probability that a randomly chosen user will find a comment offensive needs POP calibration; a benchmark leaderboard scored by accuracy against the majority rewards MAJ calibration. The question "is this model calibrated?" has no answer until one of them is named.

2.2 What proper scores do and do not choose

The choice of loss makes this concrete. A scoring rule is strictly proper if its expected value under a distribution q is uniquely minimised by forecasting q (Gneiting & Raftery, 2007). Negative log-likelihood and the Brier score are strictly proper. When a calibrator is fitted by minimising log-loss against labels, the labels define q. Fitted against votes, the minimiser is the forecast that best predicts the vote. Fitted against single annotator labels drawn at random, the minimiser is the forecast that best predicts a random annotator. Neither is wrong; they answer different questions. Proper scoring guarantees that the calibrator chases the target it is shown. It does not choose the target.

This point also explains one of the anchor's most practically useful results. Its Monte Carlo temperature scaling, fitted with a single annotation per calibration example drawn from the item's annotations, matches the calibrator fitted to the full distribution within 0.6 points of ECE on all eight dataset-architecture settings (Tao et al., 2026). The anchor reads this as showing that "pre-aggregated label distributions are unnecessary". It follows from linearity: the expected log-loss of a forecast against one label drawn from q equals the cross-entropy between q and the forecast. With one free parameter and thousands of calibration items, the sampling noise in single labels averages out across items. The consequence deserves to be stated plainly. Any dataset in which each item carries one label from one randomly assigned annotator is already ANN-calibration data. What makes a dataset MAJ-calibration data is aggregation, not the number of annotators. This also reconciles the anchor's finding with the benchmark result that one annotation is not enough for estimating each item's class distribution (Schmarje et al., 2022): a per-item estimate needs many labels per item, and a one-parameter calibrator needs many items.

2.3 Why vote-fitted temperatures are low

The anchor's direction-of-bias argument has a short intuition. On an item where 60 percent of annotators say A and 40 percent say B, the vote is A every time. A calibrator fitted to votes is rewarded for raising the probability of A toward 1 on such items, which means lowering the temperature. A calibrator fitted to annotator draws is rewarded for raising it only to 0.6. Across a calibration set with many such items, the vote-fitted temperature is lower. The anchor's tables bear this out on every setting: the temperature fitted to votes is 2.03 and 2.04 on the two CIFAR-10H backbones against 3.18 and 3.07 when fitted to the annotator distribution, and 2.42 and 3.81 against 3.41 and 5.28 on ChaosNLI (Tao et al., 2026, Table II). Two features of these numbers matter later. Every fitted temperature is above

1, so the base models are overconfident against the votes as well as against annotators; the disagreement is about how much to soften, not whether. And the gap is a direct readout of how far the two targets diverge on the calibration set, which makes it a property of the data before it is a property of the method.

3. The anchor's result, read as a design

3.1 What was compared

Tao et al. (2026) evaluate on four benchmarks with two backbones each: CIFAR-10H with ResNet-50 and ViT-B/16, ChaosNLI with RoBERTa-Large and DeBERTa-v3, ISIC 2019 with EfficientNet-B4 and ViT-S/16, and DermaMNIST with ResNet-18 and ViT-S/16. CIFAR-10H supplies an empirical distribution of crowd labels for each CIFAR-10 test image (Peterson et al., 2019). ChaosNLI supplies 100 crowd labels for each of several thousand SNLI, MNLI and abductive NLI items (Nie et al., 2020). For the two dermatology sets no multi-annotator labels are public, and the anchor generates annotations from a confusion matrix described as clinically informed (Section 6.2 below returns to this).

The primary metric is what the anchor calls true-label ECE: ECE computed against a label drawn from the annotator distribution, averaged over 100 draws. The baselines are calibrators fitted on voted labels: temperature scaling, adaptive temperature scaling, Platt scaling and Dirichlet calibration. The proposed methods are fitted on the annotator distribution or on draws from it, plus one variant, label-smoothed temperature scaling, that constructs soft targets from the model's own confidence when only votes are available.

3.2 What was found

On CIFAR-10H with ResNet-50, temperature scaling fitted to votes leaves a true-label ECE of 4.29 percent; the soft-label temperature fit leaves 1.51, and Dirichlet calibration fitted to the soft target leaves 1.25. On ChaosNLI with RoBERTa-Large the corresponding numbers are 10.55, 3.22 and 2.57 (Tao et al., 2026, Table II). On the synthetic dermatology benchmarks the vote-fitted calibrators leave the largest errors, 17 to 25 percent, and the soft-fitted ones bring them down to between about 2 and 10 percent depending on method and backbone (Tao et al., 2026, Table III).

Two side results support the anchor's diagnosis that the target, not the calibrator's capacity, is the binding constraint. Dirichlet calibration fitted to votes, the most expressive of the vote-fitted maps, does slightly worse than temperature scaling against the annotator target: 4.46 against 4.29 on CIFAR-10H with ResNet-50 and 11.55 against 10.55 on ChaosNLI with RoBERTa-Large. And temperature scaling's ECE against the annotator target does not improve as the calibration set grows, while the soft-fitted methods converge once a small fraction of the calibration data is used (Tao et al., 2026, Figures 10 and 11). A more flexible map, or more data, fitted to the wrong target, converges more precisely on the wrong target.

The anchor also reports the result in a form that matters for practice: the label-smoothed variant, which needs only votes, reduces ECE by 9 to 77 percent relative to temperature scaling. The range is wide. On ISIC 2019 with ViT-S/16 it moves ECE from 17.04 to 15.49, and the anchor notes that the gain is smaller there because the model's confidence is a weaker proxy for annotation ambiguity. In earlier histopathology work, label smoothing applied during training with model confidence standing in for annotator agreement cut ECE by 64 percent, from 8.9 to 3.2, on colorectal polyp classification, and smoothing driven by real annotator agreement cut calibration error by almost 70 percent (Wei et al., 2022a).

3.3 Independent confirmation

Dogah (2026) runs a related comparison on CIFAR-10H and ChaosNLI with three model scales per modality and both temperature scaling and isotonic regression. In all nine configurations, a calibrator fitted to hard labels has a worse soft-label Brier score than an oracle fitted to soft labels, with gaps from 0.002 to 0.134. The mean gap is 0.079 in language against 0.003 in vision. The gap grows with model scale on CIFAR-10H and on the SNLI portion of ChaosNLI, reverses on the MNLI portion, and the author attributes the reversal mainly to near-chance accuracy on that split. The direction of the effect is therefore replicated. Its size is small on the vision benchmark, where most CIFAR-10H images are unambiguous, and larger on the language benchmark, which was built by selecting low-agreement items.

3.4 What the design fixes in advance

Three features of the design determine what the result can mean.

The target is fixed as ANN by definition. The anchor states its reason for keeping ECE against sampled labels: that it "quantifies the overconfidence a user experiences when drawing a single label at deployment time". This is a declaration of a deployment model in which the thing being predicted is a random annotator's label. It is a reasonable declaration for some deployments. It is not a demonstration that ANN is the right target in the deployments the anchor's introduction motivates, which include medical decision support.

The comparison is between two targets, not between a biased and an unbiased estimate of one. Vote-fitted calibrators are scored against annotator draws; annotator-fitted calibrators are scored against annotator draws. The anchor's discussion acknowledges the converse, that fitting to the annotator target can cost calibration against the voted label, and calls the trade-off expected, on the grounds that "voted-label calibration becomes epistemically unreliable when annotators

genuinely disagree". That sentence is true under ANN and false under MAJ, and the anchor does not measure the trade-off on its benchmarks. It reports voted-label ECE only for temperature scaling, Platt scaling and histogram binning on its synthetic two-dimensional Gaussian example, and for no calibrator on CIFAR-10H, ChaosNLI, ISIC 2019 or DermaMNIST, where every ECE column of its Tables II and III is computed against sampled annotator labels.

Half the benchmarks have synthetic annotators. The size of the effect on those two benchmarks is a property of the confusion matrix the anchor chose, not of any population of readers (Section 6.2).

4. The pathology runs both ways

4.1 The oracle that fails majority calibration

Baan et al. (2022) asked what standard calibration metrics say about a classifier that is right about human uncertainty. They state a desideratum: a classifier that outputs, for each input, the human judgement distribution should count as perfectly calibrated. They then show that, under ECE against the majority, it does not. On ChaosNLI-SNLI their oracle classifier, which outputs the empirical human distribution, has a majority-vote ECE of 0.25, against 0.14 for the RoBERTa model it is compared with; on their human-distribution measures the same oracle is perfectly calibrated. Their conclusion is that "human majority accuracy is a bad estimate of correctness to calibrate against" when humans disagree, and they propose instance-level measures based on the full human distribution: one sensitive to the distribution's entropy, one to its class ranking, and one to its distance from the model's prediction.

Read beside the anchor, the two results are symmetric. A forecast that matches q is miscalibrated against MAJ, because on ambiguous items the majority label occurs every time while q gives it 0.6. A forecast fitted to MAJ is miscalibrated against q for the same reason in the other direction. Neither result shows that one target is the truth. Both show that a calibration number without a declared target is not interpretable.

Baan et al. (2022) also report that temperature scaling with an oracle temperature, chosen on the test set to minimise majority-vote ECE, cuts RoBERTa's majority-vote ECE from 0.14 to 0.03 while moving the mean of their instance-level human-calibration error only from 0.26 to 0.22. This is the anchor's finding, seen in 2022 on one benchmark with one backbone and with a different instrument.

4.2 The same shape outside classification

Conformal prediction gives a second view. A prediction set built on calibration data with voted labels has its coverage guarantee with respect to the voted-label distribution. In a dermatology case study, Stutz et al. (2023a) show that sets calibrated for 72 percent coverage on votes fall short of that coverage by 10 percent on average when scored against expert annotations, and they propose Monte Carlo conformal procedures that sample pseudo-labels from an aggregated expert distribution to restore coverage. The structure is the anchor's: the guarantee holds for the target the calibration data encode.

The same shape appears in generation. When next-word predictions from GPT-2, BLOOM and ChatGPT are compared with the spread of continuations a population of English speakers produced for the same contexts, the models match that spread poorly, and ECE does not register the mismatch (Ilia & Aziz, 2024). Measured against multiple human references, generators' output distributions can be probed instance by instance for calibration to human production variability (Giulianelli et al., 2023). In both, the reference variable is a population of human productions, and the standard metric against a single reference does not see the miscalibration.

4.3 Aggregation is itself miscalibrated

If MAJ were a clean, well-defined target, the choice would at least be between two coherent options. It is not. Singer et al. (2026) study the calibration of crowdsourcing aggregators across 39 human crowdsourcing datasets and LLM-based crowds. Majority vote is underconfident; model-based aggregators such as Dawid-Skene (Dawid & Skene, 1979), GLAD and CrowdFM are overconfident. They prove that a Bayesian aggregator is perfectly calibrated under its own data-generating model and that exact calibration of weighted majority votes requires highly restrictive conditions. The labels that most calibration work treats as ground truth are, in this light, the output of a forecaster with its own calibration error, and different aggregators place that error on different sides.

5. What the annotator distribution contains

The anchor's case for ANN rests on reading disagreement as "irreducible aleatoric uncertainty inherent to the task". The literature on disagreement supports that reading for some items and contradicts it for others. The empirical label distribution of an item is the product of at least four things: genuine variation in valid interpretations, annotator error, within-annotator inconsistency, and the composition of the labelling pool.

5.1 Variation is real

The case that much disagreement is valid is well established. On NLI items, disagreements persist as more ratings are gathered and as raters see more context, and current models' uncertainty does not track the kind present in human judgements (Pavlick & Kwiatkowski, 2019). A taxonomy of disagreement sources in NLI separates uncertainty in sentence meaning from

annotator biases and task artifacts and finds all of them, which leads to different interpretations of what a label distribution means (Jiang & de Marneffe, 2022). Collecting 100 labels per item shows that high disagreement is common on low-agreement subsets and that models can barely beat a random guess on items with the lowest human agreement while doing almost perfectly on high-agreement items (Nie et al., 2020). Training on full human label distributions improves generalisation to shifted test sets and robustness to adversarial attack on CIFAR-10 (Peterson et al., 2019). Across six datasets with human label variation, fourteen training methods and six metrics, training on disaggregated annotations or soft labels performs best across metrics (Kurniawan et al., 2025).

5.2 Error is real too

The same literature has measured the other components. VariErr re-annotated 500 MNLI items in two rounds, with annotators writing an explanation for each label and then judging the validity of label-explanation pairs; the resulting 7,732 validity judgements over 1,933 explanations separate labels given for valid reasons from labels given for invalid ones (Weber-Genzel et al., 2024). Automatic error-detection methods fall well short of GPT-4 and of humans at this separation. The point for calibration is simple: if some labels in an item's distribution are errors, a calibrator fitted to that distribution learns to reproduce the error rate.

The disagreement deconvolution makes the noise component explicit. Gordon et al. (2021) estimate each annotator's intra-annotator consistency, separate stable opinions from noise, and compare each prediction to each annotator's stable opinion. Applied to existing social computing datasets, the transformation shows "current metrics dramatically overstate the performance of many human-facing machine learning tasks"; on comment toxicity, ROC AUC falls from .95 to .73. Their correction runs in the direction of disagreement, not against it, but its mechanism is a separation of the empirical label distribution into a stable part and a noise part, which is the separation a calibration target needs.

Real crowd label noise has structure. On CIFAR-10N and CIFAR-100N, human-annotated noisy labels follow an instance-dependent pattern rather than the class-dependent pattern common in synthetic noise studies (Wei et al., 2022b). Instance-dependent error is exactly the kind that looks like ambiguity in an empirical distribution, because it concentrates on the same hard items.

5.3 The estimate is noisy even with 100 labels

ChaosNLI supplies a direct measure of how stable an annotator distribution is. To estimate human performance, Nie et al. (2020) collected a second, disjoint set of 100 labels for 200 sampled items in each subset and compared the two distributions. The Jensen-Shannon distance between them was 0.0614 on the SNLI subset and 0.0695 on the MNLI subset. That is far below the best model's distance from the human distribution, which was above 0.2, so the target is stable enough to show that models miss it. It is not zero, and it sets a floor below which no calibrator can be said to be closer to "the" annotator distribution rather than to one sample of it. With the five labels per item that SNLI and MNLI originally collected, the majority itself changed on 24.97 and 31.77 percent of the low-agreement items when 100 labels were gathered, and on 3.78 and 5.08 percent of the full development sets.

5.4 The pool is a choice

An annotator distribution is a distribution over a particular pool. For subjective tasks the pool's composition changes the target. Annotator disagreement on hate speech may reflect group differences in what counts as offensive; modelling individual annotators and the target group of a statement improved prediction of individual ratings by 22 percent and of the variance among annotators by 33 percent over a baseline (Fleisig et al., 2023). Multi-task models that predict each annotator's labels perform as well as or better than aggregation across seven binary subjective tasks, and their uncertainty tracks annotator disagreement better than standard estimates (Davani et al., 2022). Jury learning makes the choice explicit by letting a practitioner specify which groups, in what proportion, determine the classifier's output (Gordon et al., 2022). Group membership is not a shortcut to the pool, either: adding sociodemographic group layers to multi-annotator models of toxic content did not significantly improve them (Orlikowski et al., 2023). In each case the right distribution is the one of the population whose judgement the deployment serves, which is POP, and the crowd that happened to label the data is a sample of that population only by assumption.

5.5 What this does to the anchor's target

None of this undermines the anchor's demonstration that vote-fitted temperature scaling misstates confidence against annotator labels. It qualifies the word "true". ANN-calibration reproduces the labelling pool's mixture of valid variation, error and noise, weighted by that pool's composition. Where error and noise are a small share of disagreement and the pool matches the population of interest, ANN approximates the target a deployment wants. Where they are not, ANN-calibration is calibration to the labelling process. The published measurements say that the second case is not rare.

6. When a reference outcome exists

6.1 Disagreement about a fact

Medical imaging is the setting where the distinction between ANN and REF bites hardest, because a reference outcome often exists and readers often disagree about it.

Diabetic retinopathy grading is a clean example. Krause et al. (2018) graded images with three retina specialists and adjudicated their disagreements face to face, and separately had three ophthalmologists grade the same validation set. Individual specialists were very specific but their sensitivity for moderate-or-worse disease ranged from 74.4 to 82.1 percent; the majority decision of the three reached 88.1 percent; adjudication found the rest. The weighted kappa between the specialists' pre-adjudication majority and the adjudicated grade was 0.91. When disagreements between the ophthalmologists' majority and the adjudicated consensus were reviewed, the most common causes were missed microaneurysms (36 percent), artifacts (20 percent), and disagreement about whether a lesion was a microaneurysm or a haemorrhage (16 percent). The first two are reading errors, not alternative readings of the same evidence. The authors conclude that "a small set of adjudicated DR grades allows substantial improvements in algorithm performance", and evaluated against the adjudicated reference, a gain from higher image resolution that was barely visible against the majority grade became an improvement in AUC from 0.952 to 0.984 for mild-or-worse disease.

In this setting, calibrating a model to the distribution of individual graders' labels would make it as uncertain as a single grader who sometimes misses microaneurysms. That is calibration to the readers' error profile, which has uses (Section 6.3), but it is not calibration to the patient's disease state. The target that deployment decisions are scored against is the adjudicated grade or a later clinical outcome.

Earlier work on the same disease showed that modelling individual graders and learning how to weight them improves diabetic retinopathy classification over treating the majority as correct or treating the label as a distribution (Guan et al., 2018). Expert disagreement there is informative about the reference, not a replacement for it.

6.2 The anchor's dermatology benchmarks

The two dermatology benchmarks in the anchor carry the largest effects. Temperature scaling fitted to votes leaves true-label ECE of 22.25 and 24.65 percent on DermaMNIST and 18.72 and 17.04 percent on ISIC 2019, against about 2 to 10 percent for the soft-fitted methods (Tao et al., 2026, Table III). The annotations behind these numbers are synthetic. Because no multi-annotator labels are public for either dataset, the anchor generates them from a confusion matrix, and it states the model's limitation directly: "every image with the same consensus class is assigned the same row of the confusion matrix, regardless of its individual visual content".

This design choice changes what the benchmark measures. If the annotator distribution of every melanoma image is the same row, then the annotator target is the consensus label passed through a fixed noise channel. For DermaMNIST, at least, the consensus label is not a crowd vote. DermaMNIST is derived from HAM10000, whose lesions were confirmed by pathology in more than half of cases and otherwise by follow-up, expert consensus or confocal microscopy (Tschandl et al., 2018). On that benchmark, then, the anchor's true label is a synthetic corruption of something close to a REF label, and the same construction applies to ISIC 2019 whatever the provenance of its consensus labels. A calibrator that matches such a target is calibrated to a noise model, and a vote-fitted calibrator that is overconfident against it may be close to correctly confident against the reference the data were built on. The benchmark cannot tell these apart, because the consensus label was used to generate the target.

The anchor acknowledges that instance-level annotator variation would differ from its class-conditional model. The point here is stronger than that the model is simplified. In a class-conditional noise model, the disagreement carries no information about the image beyond its class, so it is not ambiguity in any sense that a model looking at the image could learn. The large reductions on these benchmarks are therefore evidence that soft-fitted calibrators match a known noise channel, not evidence about how they behave on real dermatological ambiguity.

6.3 What real expert disagreement shows

Real multi-reader dermatology data exist. Stutz et al. (2023b) study skin condition classification where each case carries several dermatologists' differential diagnoses, and they treat the ground truth itself as uncertain. They aggregate the annotations into a distribution over condition probabilities, sample from it, and evaluate metrics such as top-k accuracy under each sample. They find that a large share of the dataset carries significant ground-truth uncertainty and that standard evaluation, which aggregates to a single label, overestimates performance, with metrics moving by several percentage points across samples. They also conclude that "assuming a crisp ground truth can be acceptable for many AI applications", while medical diagnosis needs a more careful protocol. This is the most careful published treatment of the ANN-REF distinction in a clinical setting, and its design differs from the anchor's in a way that matters: the uncertainty is modelled at the level of each case's annotations, not imposed as a class-conditional channel.

Expert disagreement also has value as its own prediction target. Models trained to predict directly whether a case will produce high disagreement among doctors identify cases that would benefit most from a second opinion, and do this better than post-processing a classifier's output distribution (Raghu et al., 2019). That result separates two quantities the anchor's framework merges: the probability that a case is disease A, and the

probability that readers will disagree about it. A single calibrated distribution cannot represent both when they come apart, which they do whenever an easy-to-misread case has an unambiguous reference outcome.

Object detection in medical images has recently been calibrated directly to annotators' annotation distribution without access to ground truth, in an approach designed for objects that are ambiguous to annotators (Tan et al., 2026). The same question applies: whether the annotator distribution or an outcome is the quantity the downstream decision is scored against.

7. Subjective tasks: the population is the target

Where no reference outcome exists, the case for annotator-based targets is strongest. Rottger et al. (2022) distinguish two annotation paradigms: "Descriptive annotation allows for the surveying and modelling of different beliefs, whereas prescriptive annotation enables the training of models that consistently apply one belief." The distinction maps onto the targets here. A descriptive dataset is built for POP or ANN calibration. A prescriptive dataset, in which annotators are instructed to apply one guideline, is built for something closer to MAJ, and its disagreement is closer to error against the guideline than to valid variation.

The LeWiDi shared tasks have institutionalised the descriptive view for subjective NLP. The second edition concentrated on subjective tasks and soft evaluation (Leonardelli et al., 2023). The third edition evaluates two paradigms side by side: a soft-label paradigm in which models predict population-level distributions, and a perspectivist paradigm in which they predict individual annotators' interpretations (Leonardelli et al., 2025). These are ANN and a per-annotator refinement of it; neither is MAJ.

Calibration can also be built on disagreement for decision purposes. Crowd-Calibrator uses the distance between the crowd's label distribution and the model's to decide when to abstain, and matches or beats selective-prediction baselines on hate speech detection and NLI (Khurana et al., 2024). In legal case outcome classification, split votes among judges of the European Court of Human Rights provide naturally occurring disagreement, and models show limited alignment with the judges' vote distribution (Xu et al., 2024). The legal case is instructive because it mixes the targets: a split vote is a population of expert opinions, and the court's decision is also an outcome with legal force. A model predicting case outcomes is scored against the decision, while a model meant to anticipate how contested a case is would be scored against the distribution.

8. Language models: one distribution, two jobs

8.1 Confidence versus variation

Large language models bring the conflict into one output. Baan et al. (2024) identify two perspectives on a predictive distribution: as model confidence, in which probabilities indicate the chance of being correct, and as human label variation, in which they indicate the spread of human judgements. They argue that both matter and that exploiting a single predictive distribution for both is limiting, and they recommend disentangled representations of uncertainty about predictions and uncertainty about human labels. The anchor is a post-hoc instance of choosing the second perspective. The calibration literature on language models has mostly chosen the first.

On the confidence side, larger pretrained models are well calibrated on diverse multiple-choice and true-false questions in the right format (Kadavath et al., 2022), and post-training can damage this: the GPT-4 report shows a calibration plot on an MMLU subset and states, "The post-training hurts calibration significantly" (OpenAI, 2023). For models tuned with human feedback, verbalised confidence is often better calibrated than token probabilities, often cutting ECE by about half on TriviaQA, SciQ and TruthfulQA (Tian et al., 2023). All of these score against a single correct answer per question, which is MAJ or REF.

On the variation side, the results are weaker. Instruction-tuned models evaluated on NLI fail to capture the human disagreement distribution, and alignment worsens on items with high human disagreement (Lee et al., 2023). Opinions reflected by language models are substantially misaligned with those of US demographic groups, by an amount the authors compare to the Democrat-Republican divide on climate change, and steering toward a group does not remove it (Santurkar et al., 2023). A benchmark of distributional alignment finds that models "can more accurately describe the opinion distribution than simulate such distributions" (Meister et al., 2024): what a model says the distribution is and what its samples look like are different quantities.

8.2 Reasoning and explanations

Two recent lines of work bear directly on whether a language model's output distribution can be made into an annotator forecast. A few expert explanations substantially improve an LLM's approximation of NLI judgement distributions, but fine-tuning small soft-label models on the LLM-generated distributions gives partly inconsistent results, with similar distances hiding substantially different distributions (Chen et al., 2024). LLM-generated explanations can stand in for human ones when estimating judgement distributions from a few human labels (Chen et al., 2025). And in reasoning-tuned models, long chains of thought improve distributional alignment, but final accuracy is dictated

by chain-of-thought content (99 percent of variance in the authors' decomposition) while the ranking of the distribution is governed by model priors (over 80 percent); the authors conclude that chain of thought "fails to function as a granular distribution calibrator for ambiguous tasks" (Chen et al., 2026). Evidence-based alignment with retrieved survey data improves country-conditioned answer distributions under shift, with average relative reductions in Jensen-Shannon divergence of up to 44 percent (Pham et al., 2026).

The combined picture is that a language model's top-answer confidence and its population forecast are separable and are currently trained by different procedures toward different targets. A single post-hoc temperature cannot make one output serve both. If a deployment needs both, it needs two outputs, which is Baan et al.'s (2024) recommendation and Raghu et al.'s (2019) design.

9. What the size of the gain means

9.1 ECE and proper scores tell different stories

The anchor reports five metrics for every method: three variants of ECE against annotator draws and two strictly proper scores, Brier and NLL, computed against the soft target. The headline figures are ECE reductions. The proper scores move much less. Re-reading the anchor's Tables II and III, the author computes, for each of the eight settings, the relative change from vote-fitted temperature scaling to the soft-label temperature fit, which differs from it only in the target.

On CIFAR-10H with ResNet-50, ECE falls from 4.29 to 1.51 percent, a 65 percent reduction, and the soft-target Brier score falls from 0.112 to 0.110, a 2 percent reduction. On ChaosNLI with RoBERTa-Large, ECE falls by 69 percent, from 10.55 to 3.22, and Brier by 3 percent, from 0.537 to 0.520. On DermaMNIST with ViT-S/16, the largest case, ECE falls by 87 percent, from 24.65 to 3.29, and Brier by 11 percent, from 0.683 to 0.609. Across all eight settings the ECE reduction ranges from 48 to 87 percent and the Brier reduction from 2 to 11 percent. Figure 1(b) plots the two against each other.

Both are honest measurements, and they measure different things. ECE isolates the reliability component of forecast quality, the gap between stated confidence and observed frequency (DeGroot & Fienberg, 1983). A strictly proper score combines reliability with resolution, the forecast's ability to separate items with different outcome frequencies, and with the irreducible uncertainty of the target itself. When the target is a random draw from a high-entropy distribution, much of the expected proper score is set by that entropy and no forecast can remove it. A large relative change in ECE beside a small relative change in Brier says that the soft-fitted calibrators fix most of a reliability error that was a small part of the total.

That is not a criticism of the reliability gain. It does correct one natural reading of the headline, which is that soft-target calibration makes predictions 55 to 87 percent better. For Dirichlet-Soft, the method behind that range, ECE falls by 55 to 87 percent and the soft-target Brier score by 3 to 12 percent; for the soft-label temperature fit, ECE falls by 48 to 87 percent and Brier by 2 to 11 percent. Either way, the model's stated confidence becomes much more honest about a quantity whose total predictability changes by a few percent.

9.2 The instrument's own uncertainty

ECE carries known estimation problems. Its binned estimator is biased, with equal-width bins more biased than equal-mass bins, and debiased estimators exist (Roelofs et al., 2022; Kumar et al., 2019). Scaling methods such as temperature scaling can be less calibrated than reported, and their remaining miscalibration is hard to estimate (Kumar et al., 2019). The ranking of recalibration methods can change with the choice of calibration measure (Nixon et al., 2019). The anchor reports adaptive-bin and classwise variants alongside standard ECE. Vote-fitted temperature scaling against the soft-label temperature fit ranks the same way on all three variants in every setting, but the ranking among soft-fitted calibrators does not: IR-Soft, which has the lowest ECE of any method on ISIC 2019, has a higher classwise ECE than vote-fitted temperature scaling with ViT-S/16 (5.65 against 5.49), and with EfficientNet-B4 it has the worst classwise ECE of the soft-fitted methods (Tao et al., 2026, Table III). Under ANN, a further source of variance enters: each evaluation label is a random draw, and the anchor averages over 100 draws to reduce it. Section 5.3's disjoint-sample distances on ChaosNLI give a sense of how far the target itself moves between samples of annotators, which no number of draws from one sample can remove.

9.3 Within-target gains are not in dispute

Within the ANN target, the anchor's comparisons are like-for-like and its rankings are stable across seeds, calibration-set sizes and annotation counts (Tao et al., 2026, Figure 9). A practitioner who has decided that ANN is the right target should use a soft-fitted calibrator, and the evidence that a single random annotation per item suffices is a genuine reduction in cost. The disputes in this paper are about the choice of target and the description of the gain, not about the ranking of calibrators once the target is fixed.

10. Training-time alternatives and the post-hoc ceiling

The anchor's methods are post-hoc and require no retraining. That is a practical strength and a ceiling. A single temperature, or a Dirichlet map, applied to fixed logits can change how confident a model is on average and how that confidence is distributed across classes. It cannot teach the model which inputs are ambiguous if

the model's logits do not already encode that.

Training with distributional targets attacks that limitation. Peterson et al. (2019) showed that training on CIFAR-10H's human distributions makes classifiers exhibit human-like uncertainty they otherwise lack, with gains in out-of-distribution generalisation and adversarial robustness. Eliciting soft labels from each annotator directly, as in CIFAR-10S with 248 crowd participants, achieves comparable model performance to prior soft-label approaches with far fewer annotators, at a cost in time per elicitation (Collins et al., 2022). On NLI, distribution-estimation methods such as Monte Carlo dropout, ensembles, re-calibration and distribution distillation capture the human judgement distribution better than the softmax baseline, but the best results remain far below the estimated human upper bound (Zhou et al., 2022). Label smoothing, which softens targets uniformly rather than per instance, improves calibration (Muller et al., 2019) without any information about which items are ambiguous.

The anchor's label-smoothed variant sits between these: it uses the model's own confidence as a proxy for ambiguity. Its effectiveness depends on whether the model's confidence already tracks annotator disagreement. Where it does, as on CIFAR-10H with ResNet-50, the variant nearly matches the full-distribution methods; where it does not, as on ISIC with ViT-S/16, it barely improves on temperature scaling. A post-hoc calibrator inherits the base model's resolution. If the deployment target is ANN or POP, training-time methods that change resolution are likely to matter more than post-hoc methods that only rescale; the proper-score figures in Section 9 suggest the post-hoc ceiling on those targets is low.

The distinction between aleatoric and epistemic uncertainty is the standard vocabulary for this ceiling (Kendall & Gal, 2017; Hullermeier & Waegeman, 2021), and the anchor uses it: annotator disagreement is aleatoric. That labelling needs a reference variable. Disagreement is aleatoric with respect to ANN and POP by definition; with respect to REF it is, in part, the readers' epistemic uncertainty about an outcome that exists. The two-dimensional taxonomy of uncertainty sources proposed for natural language generation makes a similar point: the aleatoric-epistemic dichotomy is less informative than a classification of where uncertainty comes from (Baan et al., 2023).

11. The measured values

Table 1 collects the published numbers on which the argument rests, one row per cited result, with the target each row is scored against. Figure 1 plots the anchor's own numbers that carry most of the weight: the temperatures fitted to each target, and the relative change in ECE beside the relative change in a strictly proper score when the target changes.

Three readings of Table 1 and Figure 1 carry the argument. In panel (a), the ratio of the two temperatures is largest on the synthetic dermatology benchmarks, where the annotator target was generated from the consensus label through a fixed confusion matrix; the size of the gap there is a property of that matrix. In panel (b), every point sits far above the line of equal relative change: the change of target removes most of the reliability error measured by ECE and a small fraction of the expected proper score. And across the table, the rows scored against REF (Krause et al., 2018) and the rows that decompose the annotator distribution (Nie et al., 2020; Gordon et al., 2021) are the ones the anchor's design does not reach.

12. A target-declaration procedure

The analysis of Sections 2 to 10 can be stated as a procedure for deciding which calibration target a reported calibration claim should be scored against, and what a change of target can be credited with. Algorithm 1 sets it out. It needs only quantities a study can report or collect: the deployment decision, the labelling pool, per-item annotations, and where possible a reference outcome for a subset.

Applied to the anchor's reported numbers, the procedure credits the soft-fitted calibrators with a large and well-replicated gain under ANN on CIFAR-10H and ChaosNLI, with the caveat for ChaosNLI that the reported ECE gap is not compared with the disjoint-sample floor. On ISIC 2019 and DermaMNIST it credits them with matching a synthetic class-conditional channel; since the channel was generated from labels that are close to REF, step 2 applies and those results say nothing yet about real readers or real outcomes. Under MAJ the procedure credits vote-fitted temperature scaling as the matching calibrator, which is Baan et al.'s (2022) converse. It credits nothing under POP, because no study reviewed here checks whether the labelling pool samples a deployment population.

13. The case against this reading

The strongest objections deserve to be stated at full strength.

The anchor already declared its target. It did. It states why it keeps ECE against sampled labels, that the metric captures the overconfidence a user meets when one label is drawn at deployment time, and it reports proper scores alongside. This paper's objection is not that the target was hidden but that it was named "true", which forecloses the comparison with MAJ and REF that the anchor's own medical motivation requires. If the anchor's conclusion is read as a claim about ANN calibration, this paper has no quarrel with it.

For perceptual and inferential crowd tasks there is no reference beyond the crowd. On CIFAR-10H and ChaosNLI, this is largely right. No biopsy exists for

Table 1. Measured values from the cited sources. Target names the random variable the number is scored against (MAJ: aggregated vote; ANN: a randomly drawn annotator's label or the annotator distribution; REF: adjudicated or later-established outcome). TS is temperature scaling fitted to votes; SLTS is temperature scaling fitted to the annotator distribution; "/" separates the two backbones in the order given in the anchor.

Source	Setting	Target	Quantity	Value	Bears on
Tao et al. (2026), Tab. II	CIFAR-10H, R50 / ViT	ANN	ECE, TS vs SLTS (%)	4.29 vs 1.51 / 4.48 vs 0.85	anchor effect, low ambiguity
Tao et al. (2026), Tab. II	ChaosNLI, RoBERTa / DeBERTa	ANN	ECE, TS vs SLTS (%)	10.55 vs 3.22 / 11.63 vs 3.45	anchor effect, NLI
Tao et al. (2026), Tab. III	ISIC 2019, synthetic	ANN	ECE, TS vs SLTS (%)	18.72 vs 9.76 / 17.04 vs 7.10	class-conditional channel
Tao et al. (2026), Tab. III	DermaMNIST, synthetic	ANN	ECE, TS vs SLTS (%)	22.25 vs 5.13 / 24.65 vs 3.29	class-conditional channel
Tao et al. (2026), Tab. II-III	eight settings	ANN	Brier, TS to SLTS	down 2 to 11%	proper score moves little
Tao et al. (2026), Tab. II	CIFAR-10H R50; ChaosNLI RoBERTa	ANN	ECE, Dirichlet on votes vs TS	4.46 vs 4.29; 11.55 vs 10.55	capacity is not the limit
Tao et al. (2026), Sec. VI	eight settings	ANN	one label per item vs full	within 0.6 pp ECE	ANN data need not be aggregated
Dogah (2026)	CIFAR-10H, ChaosNLI, 9 configs	ANN	Brier gap, hard vs soft fit	0.002 to 0.134	replication of direction
Baan et al. (2022)	ChaosNLI-SNLI	MAJ	ECE, oracle vs RoBERTa	0.25 vs 0.14	converse pathology
Baan et al. (2022)	ChaosNLI-SNLI, RoBERTa	MAJ / ANN	oracle TS: majority ECE / human error	0.14 to 0.03 / 0.26 to 0.22	TS barely moves ANN error
Stutz et al. (2023a)	dermatology, conformal	MAJ vs ANN	coverage shortfall at 72%	about 10 points	same shape, sets
Singer et al. (2026)	39 crowd datasets	aggregator	calibration of aggregators	MV under-, model-based over-confident	MAJ is itself miscalibrated
Nie et al. (2020)	SNLI / MNLI low-agreement	MAJ	majority changed, 5 to 100 labels	24.97 / 31.77%	vote instability
Nie et al. (2020)	SNLI / MNLI, 200 items	ANN	JSD between disjoint 100-label sets	0.0614 / 0.0695	floor on ANN target
Gordon et al. (2021)	comment toxicity	ANN, denoised	ROC AUC, before / after	.95 / .73	aggregate metric overstates; noise separable from stable opinion
Krause et al. (2018)	retinal grading	REF	causes of majority-adjudication disagreement	missed MAs 36%, artifacts 20%	disagreement as error
Krause et al. (2018)	retinal grading	REF	sensitivity, single vs majority	74.4-82.1% vs 88.1%	readers vs reference
Wei et al. (2022a)	colorectal polyps	ANN-like	calibration error, label smoothing	almost 70% lower	agreement-aware softening
Fleisig et al. (2023)	hate speech	POP	gain on individual ratings / variance	22% / 33%	pool composition matters
Chen et al. (2026)	reasoning LLMs, HLV tasks	ANN	variance share, CoT / priors	99% accuracy / over 80% ranking	one output, two jobs
Tian et al. (2023)	RLHF LMs, QA	MAJ/REF	ECE, verbalised vs token prob	about 50% lower	confidence target

whether a blurry image is a cat or a dog or whether a hypothesis follows from a premise; human judgement is the phenomenon. The paper concedes that on these benchmarks ANN is a natural target, and that Section 6's objection does not apply to them. Section 5's objection still does: the error and noise components have been measured on exactly these kinds of data (Weber-Genzel et al., 2024; Gordon et al., 2021), so ANN on these benchmarks is calibration to a labelling process that includes some error.

The error share is small enough to ignore. It may be on some datasets. Pavlick & Kwiatkowski (2019) found that much NLI disagreement persists as more ratings are gathered, which is the signature of variation rather than noise, and ChaosNLI's disjoint-sample distance is small relative to model error. The paper accepts that for crowd NLI the non-variation share is probably not the dominant component of disagreement. What it disputes is that the share is known: none of the calibration studies reviewed here measures it, and the one retinal grading study that traced disagreements to causes found most of them were errors.

ECE is the right metric for overconfidence, so the Brier comparison is beside the point. The paper agrees that ECE measures the reliability component, and that a large reduction in reliability error is a real improvement in the honesty of stated confidence. Section 9's point is about description, not validity. A reader told that Dirichlet-Soft reduces calibration error by 55 to 87 percent should also be told that it lowers the soft-target

Algorithm 1. Declaring the target of a calibration claim under disagreement

```
Input: model f; items X; per-item annotations L; labelling pool A;
       deployment decision D; reference outcomes R for a subset (optional)
1  name the variable Y that D is scored against:
     MAJ  label assigned by a stated aggregation rule over L
     ANN  label of one annotator drawn from pool A
     POP  label of one member of the population P that D serves
     REF  outcome fixed by evidence the annotators lacked (biopsy, follow-up,
          adjudication)
2  if Y = REF:
     if R exists: fit and score calibration against R
                  report annotator entropy as a covariate of error against R
     else: state that the claim is about readers, not about the outcome
3  if Y in {ANN, POP}:
     if Y = POP and A is not a sample of P: report the mismatch; reweight or
                                            relabel before scoring
     estimate the non-variation share e of disagreement (validity audit,
       re-annotation, intra-annotator repeat)
     if e is large: clean or model error before using L as the target
     fit with a proper score on single draws from L (one per item suffices)
     report the disjoint-sample floor of the target if it can be estimated
4  if Y = MAJ:
     state the aggregation rule and annotator count; vote-fitted TS is the
     matching calibrator; do not describe the output as world uncertainty
5  for the declared Y report: ECE (equal-mass or debiased), a strictly proper
     score and its irreducible part where computable, across seeds
6  report the cross-target matrix: each calibrator, fitted on each target,
     scored on every other target that the data support
7  credit a calibrator with a gain on Y only if it holds under Y's own score
     and exceeds the target floor of step 3
```

Brier score against the same target by 3 to 12 percent (for the soft-label temperature fit, 48 to 87 percent and 2 to 11 percent), because that is what decides whether the method changes downstream decisions.

The synthetic medical annotations are a stated simplification and do not carry the paper. Partly conceded. The anchor's real-annotation benchmarks show the same direction, and the synthetic ones are described honestly. But they carry the largest numbers in the anchor's headline range, and they were chosen to represent the high-stakes setting the anchor's introduction motivates. If they had been run with real multi-reader data, such as the dermatology data of Stutz et al. (2023b), the effect size there would be a finding about clinical ambiguity. As run, it is a finding about the confusion matrix.

Majority vote is miscalibrated too, so preferring MAJ is no better. This objection supports the paper rather than opposing it. Singer et al. (2026) show that aggregators carry their own calibration errors, in opposite directions for different aggregators. That is a further reason to name the variable and the aggregation rule. The paper does not prefer MAJ; it argues that each target is a choice with consequences, and that REF, where it exists, is usually the one deployment decisions are scored against.

14. Tests that would settle it

Six tests would turn the target question from a matter of declaration into one of evidence. Each names what it isolates and the result that would contradict this paper's reading.

T1. A reference-outcome stratum. On a dataset that carries both multiple readers' labels and a later-established outcome for a subset, such as pathology-confirmed dermatology cases with multi-reader differentials, fit temperature scaling three ways (to votes, to annotator draws, to the reference) and score all three against the reference. Isolates the ANN-REF gap. If the annotator-fitted calibrator is as well calibrated against the reference as the reference-fitted one, disagreement in that domain tracks outcome uncertainty and Section 6's concern does not bite there.

T2. An error-cleaned annotator target. On items with validity judgements of the VariErr kind (Weber-Genzel et al., 2024), compute the TS-versus-SLTS gap twice: with the raw label distribution as target and with labels judged invalid removed. Isolates the error component. If the gap is unchanged, error does not drive the anchor's effect on NLI.

T3. The cross-target matrix. For every benchmark that supports it, report each calibrator fitted on each available target and scored on every other, including the vote-scored ECE of every calibrator, vote-fitted and soft-fitted alike, which the anchor reports for none of its four benchmarks. Isolates the trade-off the anchor calls expected. If soft-fitted calibrators lose little against MAJ while gaining much against ANN, the trade-off is cheap and the case for ANN as a default strengthens.

T4. Instance-level medical annotations. Re-run the dermatology comparison with real per-case multi-reader annotations rather than a class-conditional channel. Isolates whether the 77 to 87 percent ECE reductions on DermaMNIST survive real ambiguity. If they do, the anchor's largest numbers generalise to clinical readers.

T5. Pool shift. Fit on one annotator pool and score on a disjoint pool of the same size, using ChaosNLI's disjoint-sample design (Nie et al., 2020) or a demographic split of the kind multi-annotator toxicity datasets permit (Fleisig et al., 2023). Isolates how much of an ANN gain is specific to one pool. If the gain transfers fully, ANN calibration approximates POP calibration for that task.

T6. A decision-level outcome. Specify a downstream decision, such as abstention or referral for a second opinion, and measure its utility under each target, in the manner of selective prediction with crowd agreement (Khurana et al., 2024) and second-opinion triage (Raghu et al., 2019). Isolates whether the reliability gain changes decisions. If decisions under the soft-fitted calibrator are no better, the gain is real but immaterial for that deployment.

T2 and T3 can be run on public data with the anchor's released methods and modest compute. T1 and T4 require clinical data with reference outcomes, which some groups already hold (Stutz et al., 2023b; Krause et al., 2018).

15. What is not known

Whether calibrators fitted to annotator labels are better or worse calibrated than vote-fitted ones against a reference outcome is not known. No study reviewed here scores both against a reference outcome on the same data.

What share of disagreement on CIFAR-10H and ChaosNLI is error or within-annotator noise rather than valid variation is not known. VariErr measures it for 500 MNLI items; the deconvolution measures noise on social computing tasks; neither has been applied to the calibration benchmarks.

How the soft-fitted calibrators perform on real multi-reader medical annotations is not known. The anchor's medical numbers come from a class-conditional synthetic channel.

How much vote-scored calibration the soft-fitted methods give up is not known. The anchor reports voted-label ECE only for temperature scaling, Platt scaling and histogram binning on its synthetic two-dimensional example, and for no calibrator on its four benchmarks.

Whether a single post-hoc map can make a language model's output serve as both a confidence score and a population forecast is not known; the evidence reviewed here points against it but no study tests both targets on one model with one calibrator.

Whether any annotation pool used in these benchmarks is a representative sample of a deployment population is not known for any of them.

16. Conclusion

The anchor is right that temperature scaling fitted to majority votes is overconfident against the labels of individual annotators, and right that a calibrator fitted to annotator labels with a proper score fixes most of that, even from one label per item. Its tables are careful and its replication by an independent study holds in direction. What the result shows is narrower than its vocabulary. The anchor compares two targets and names one of them true. Baan et al. (2022) had already shown the mirror image: a model that reproduces the human distribution fails calibration against the vote. The annotator distribution, the target the anchor promotes, is a mixture of valid variation, error, noise and pool composition, and the published re-annotations, error audits and adjudications show that the non-variation part is not negligible where it has been measured. In medicine, where the anchor locates the stakes, a reference outcome often exists, and its two medical benchmarks generate the annotator target from something close to that outcome through a fixed confusion matrix. And the headline reductions in calibration error sit beside reductions of a few percent in the strictly proper scores the anchor also reports.

A calibration number is an answer to the question "how often does the outcome occur when the model says p?" Under disagreement that question has four versions, and Algorithm 1 is the requirement that a paper say which one it answered. The six tests of Section 14 are what it would take to know which version a given deployment should ask.

References

Aroyo, L., & Welty, C. (2015). Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. AI Magazine, 36(1). doi:10.1609/aimag.v36i1.2564

Baan, J., Aziz, W., Plank, B., & Fernandez, R. (2022). Stop Measuring Calibration When Humans Disagree. Proceedings of EMNLP 2022. arXiv:2210.16133. doi:10.48550/arXiv.2210.16133

Baan, J., Daheim, N., Ilia, E., Ulmer, D., Li, H.-S., Fernandez, R., Plank, B., Sennrich, R., Zerva, C., & Aziz, W. (2023). Uncertainty in Natural Language Generation: From Theory to Applications. arXiv:2307.15703. doi:10.48550/arXiv.2307.15703

Baan, J., Fernandez, R., Plank, B., & Aziz, W. (2024). Interpreting Predictive Probabilities: Model Confidence or Human Label Variation? Proceedings of EACL 2024. arXiv:2402.16102. doi:10.48550/arXiv.2402.16102

Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M., & Uma, A. (2021). We Need to Consider Disagreement in Evaluation. Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future. doi:10.18653/v1/2021.bppf-1.3

Chen, B., Wang, X., Peng, S., Litschko, R., Korhonen, A., & Plank, B. (2024). "Seeing the Big through the Small": Can LLMs Approximate Human Judgment Distributions on NLI from a Few Explanations? Findings of EMNLP 2024. arXiv:2406.17600. doi:10.48550/arXiv.2406.17600

Chen, B., Peng, S., Korhonen, A., & Plank, B. (2025). A Rose by Any Other Name: LLM-Generated Explanations Are Good Proxies for Human Explanations to Collect Label Distributions on NLI. Findings of ACL 2025. arXiv:2412.13942. doi:10.48550/arXiv.2412.13942

Chen, B., Hu, T., Zhang, C., Litschko, R., Korhonen, A., & Plank, B. (2026). Decoupling the Effect of Chain-of-Thought Reasoning: A Human Label Variation Perspective. Findings of ACL 2026. arXiv:2601.03154. doi:10.48550/arXiv.2601.03154

Collins, K. M., Bhatt, U., & Weller, A. (2022). Eliciting and Learning with Soft Labels from Every Annotator. Proceedings of HCOMP 2022. arXiv:2207.00810. doi:10.48550/arXiv.2207.00810

Davani, A. M., Diaz, M., & Prabhakaran, V. (2022). Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. arXiv:2110.05719. doi:10.48550/arXiv.2110.05719

Dawid, A. P., & Skene, A. M. (1979). Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. Journal of the Royal Statistical Society Series C, 28(1). doi:10.2307/2346806

DeGroot, M. H., & Fienberg, S. E. (1983). The Comparison and Evaluation of Forecasters. The Statistician, 32(1/2). doi:10.2307/2987588

Dogah, W. (2026). Temperature Scaling Is Not Enough: Calibration Gaps Under Human Label Distributions. arXiv:2607.13423. doi:10.48550/arXiv.2607.13423

Fleisig, E., Abebe, R., & Klein, D. (2023). When the Majority is Wrong: Modeling Annotator Disagreement for Subjective Tasks. arXiv:2305.06626. doi:10.48550/arXiv.2305.06626

Giulianelli, M., Baan, J., Aziz, W., Fernandez, R., & Plank, B. (2023). What Comes Next? Evaluating Uncertainty in Neural Text Generators Against Human Production Variability. Proceedings of EMNLP 2023. arXiv:2305.11707. doi:10.48550/arXiv.2305.11707

Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association, 102(477). doi:10.1198/016214506000001437

Gordon, M. L., Zhou, K., Patel, K., Hashimoto, T., & Bernstein, M. S. (2021). The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. Proceedings of CHI 2021. doi:10.1145/3411764.3445423

Gordon, M. L., Lam, M. S., Park, J. S., Patel, K., Hancock, J. T., Hashimoto, T., & Bernstein, M. S. (2022). Jury Learning: Integrating Dissenting Voices into Machine Learning Models. Proceedings of CHI 2022. arXiv:2202.02950. doi:10.48550/arXiv.2202.02950

Guan, M. Y., Gulshan, V., Dai, A. M., & Hinton, G. E. (2018). Who Said What: Modeling Individual Labelers Improves Classification. Proceedings of AAAI 2018. arXiv:1703.08774. doi:10.48550/arXiv.1703.08774

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. Proceedings of ICML 2017. arXiv:1706.04599. doi:10.48550/arXiv.1706.04599

Hullermeier, E., & Waegeman, W. (2021). Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. Machine Learning, 110. doi:10.1007/s10994-021-05946-3

Ilia, E., & Aziz, W. (2024). Predict the Next Word: Humans exhibit uncertainty in this task and language models ___. Proceedings of EACL 2024. arXiv:2402.17527. doi:10.48550/arXiv.2402.17527

Jiang, N.-J., & de Marneffe, M.-C. (2022). Investigating Reasons for Disagreement in Natural Language Inference. Transactions of the ACL. arXiv:2209.03392. doi:10.48550/arXiv.2209.03392

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., et al. (2022). Language Models (Mostly) Know What They Know. arXiv:2207.05221. doi:10.48550/arXiv.2207.05221

Kendall, A., & Gal, Y. (2017). What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? Proceedings of NeurIPS 2017. arXiv:1703.04977. doi:10.48550/arXiv.1703.04977

Khurana, U., Nalisnick, E., Fokkens, A., & Swayamdipta, S. (2024). Crowd-Calibrator: Can Annotator Disagreement Inform Calibration in Subjective Tasks? Proceedings of COLM 2024. arXiv:2408.14141. doi:10.48550/arXiv.2408.14141

Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Corrado, G. S., Peng, L., & Webster, D. R. (2018). Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. Ophthalmology, 125(8). arXiv:1710.01711. doi:10.48550/arXiv.1710.01711

Kull, M., Perello-Nieto, M., Kangsepp, M., Silva Filho, T., Song, H., & Flach, P. (2019). Beyond Temperature Scaling: Obtaining Well-Calibrated Multiclass Probabilities with Dirichlet Calibration. Proceedings of NeurIPS 2019. arXiv:1910.12656. doi:10.48550/arXiv.1910.12656

Kumar, A., Liang, P., & Ma, T. (2019). Verified Uncertainty Calibration. Proceedings of NeurIPS 2019. arXiv:1909.10155. doi:10.48550/arXiv.1909.10155

Kurniawan, K., Mistica, M., Baldwin, T., & Lau, J. H. (2025). Training and Evaluating with Human Label Variation: An Empirical Study. Computational Linguistics. arXiv:2502.01891. doi:10.48550/arXiv.2502.01891

Lee, N., An, N. M., & Thorne, J. (2023). Can Large Language Models Capture Dissenting Human Voices? Proceedings of EMNLP 2023. arXiv:2305.13788. doi:10.48550/arXiv.2305.13788

Leonardelli, E., Uma, A., Abercrombie, G., Almanea, D., Basile, V., Fornaciari, T., Plank, B., Rieser, V., & Poesio, M. (2023). SemEval-2023 Task 11: Learning With Disagreements (LeWiDi). arXiv:2304.14803. doi:10.48550/arXiv.2304.14803

Leonardelli, E., Casola, S., Peng, S., Rizzi, G., Basile, V., Fersini, E., Frassinelli, D., Jang, H., Pavlovic, M., Plank, B., & Poesio, M. (2025). LeWiDi-2025 at NLPerspectives: Third Edition of the Learning with Disagreements Shared Task. arXiv:2510.08460. doi:10.48550/arXiv.2510.08460

Meister, N., Guestrin, C., & Hashimoto, T. (2024). Benchmarking Distributional Alignment of Large Language Models. arXiv:2411.05403. doi:10.48550/arXiv.2411.05403

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., & Lucic, M. (2021). Revisiting the Calibration of Modern Neural Networks. Proceedings of NeurIPS 2021. arXiv:2106.07998. doi:10.48550/arXiv.2106.07998

Muller, R., Kornblith, S., & Hinton, G. (2019). When Does Label Smoothing Help? Proceedings of NeurIPS 2019. arXiv:1906.02629. doi:10.48550/arXiv.1906.02629

Nie, Y., Zhou, X., & Bansal, M. (2020). What Can We Learn from Collective Human Opinions on Natural Language Inference Data? Proceedings of EMNLP 2020. arXiv:2010.03532. doi:10.48550/arXiv.2010.03532

Nixon, J., Dusenberry, M., Jerfel, G., Nguyen, T., Liu, J., Zhang, L., & Tran, D. (2019). Measuring Calibration in Deep Learning. arXiv:1904.01685. doi:10.48550/arXiv.1904.01685

OpenAI (2023). GPT-4 Technical Report. arXiv:2303.08774. doi:10.48550/arXiv.2303.08774

Orlikowski, M., Rottger, P., Cimiano, P., & Hovy, D. (2023). The Ecological Fallacy in Annotation: Modelling Human Label Variation goes beyond Sociodemographics. Proceedings of ACL 2023. arXiv:2306.11559. doi:10.48550/arXiv.2306.11559

Pavlick, E., & Kwiatkowski, T. (2019). Inherent Disagreements in Human Textual Inferences. Transactions of the Association for Computational Linguistics, 7. doi:10.1162/tacl_a_00293

Peterson, J. C., Battleday, R. M., Griffiths, T. L., & Russakovsky, O. (2019). Human Uncertainty Makes Classification More Robust. Proceedings of ICCV 2019. arXiv:1908.07086. doi:10.48550/arXiv.1908.07086

Pham, V.-T., Qu, L., Li, Z., & Haffari, G. (2026). Evidence-based Distributional Alignment for Large Language Models. arXiv:2603.13305. doi:10.48550/arXiv.2603.13305

Plank, B. (2022). The 'Problem' of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. Proceedings of EMNLP 2022. arXiv:2211.02570. doi:10.48550/arXiv.2211.02570

Raghu, M., Blumer, K., Sayres, R., Obermeyer, Z., Kleinberg, R., Mullainathan, S., & Kleinberg, J. (2019). Direct Uncertainty Prediction for Medical Second Opinions. Proceedings of ICML

2019. arXiv:1807.01771. doi:10.48550/arXiv.1807.01771

Roelofs, R., Cain, N., Shlens, J., & Mozer, M. C. (2022). Mitigating Bias in Calibration Error Estimation. Proceedings of AISTATS 2022. arXiv:2012.08668. doi:10.48550/arXiv.2012.08668

Rottger, P., Vidgen, B., Hovy, D., & Pierrehumbert, J. B. (2022). Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. Proceedings of NAACL 2022. arXiv:2112.07475. doi:10.48550/arXiv.2112.07475

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose Opinions Do Language Models Reflect? arXiv:2303.17548. doi:10.48550/arXiv.2303.17548

Schmarje, L., Grossmann, V., Zelenka, C., Dippel, S., Kiko, R., Oszust, M., et al. (2022). Is One Annotation Enough? A Data-Centric Image Classification Benchmark for Noisy and Ambiguous Label Estimation. NeurIPS 2022 Datasets and Benchmarks. arXiv:2207.06214. doi:10.48550/arXiv.2207.06214

Silva Filho, T., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., & Flach, P. (2023). Classifier Calibration: A Survey on How to Assess and Improve Predicted Class Probabilities. Machine Learning, 112. doi:10.1007/s10994-023-06336-7

Singer, G., Arab, R., Frikha, M., Gruffaz, S., Vo Van, O., Vayatis, N., & Kalogeratos, A. (2026). On the Calibration of Crowdsourcing Aggregators. arXiv:2610.16815. doi:10.48550/arXiv.2610.16815

Stutz, D., Roy, A. G., Matejovicova, T., Strachan, P., Cemgil, A. T., & Doucet, A. (2023a). Conformal Prediction under Ambiguous Ground Truth. arXiv:2307.09302. doi:10.48550/arXiv.2307.09302

Stutz, D., Cemgil, A. T., Roy, A. G., Matejovicova, T., Barsbey, M., Strachan, P., et al. (2023b). Evaluating AI Systems under Uncertain Ground Truth: A Case Study in Dermatology. arXiv:2307.02191. doi:10.48550/arXiv.2307.02191

Tan, Z. Q., Addison, O., & Li, Y. (2026). Calibrating Probabilistic Object Detectors with Annotator Disagreement. arXiv:2605.24722. doi:10.48550/arXiv.2605.24722

Tao, L., Luo, H., Dong, M., & Xu, C. (2026). Confidence Calibration under Ambiguous Ground Truth. arXiv:2603.22879. doi:10.48550/arXiv.2603.22879

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., & Manning, C. D. (2023). Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. Proceedings of EMNLP 2023. arXiv:2305.14975. doi:10.48550/arXiv.2305.14975

Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 Dataset, a Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions. Scientific Data, 5. arXiv:1803.10417. doi:10.48550/arXiv.1803.10417

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., & Poesio, M. (2021). Learning from Disagreement: A Survey. Journal of Artificial Intelligence Research, 72. doi:10.1613/jair.1.12752

Weber-Genzel, L., Peng, S., de Marneffe, M.-C., & Plank, B. (2024). VariErr NLI: Separating Annotation Error from Human Label Variation. Proceedings of ACL 2024. arXiv:2403.01931. doi:10.48550/arXiv.2403.01931

Wei, J., Torresani, L., Wei, J., & Hassanpour, S. (2022a). Calibrating Histopathology Image Classifiers using Label Smoothing. arXiv:2201.11866. doi:10.48550/arXiv.2201.11866

Wei, J., Zhu, Z., Cheng, H., Liu, T., Niu, G., & Liu, Y. (2022b). Learning with Noisy Labels Revisited: A Study Using Real-World Human Annotations. Proceedings of ICLR 2022. arXiv:2110.12088. doi:10.48550/arXiv.2110.12088

Xu, S., Santosh, T. Y. S. S., Ichim, O., Plank, B., & Grabmair, M. (2024). Through the Lens of Split Vote: Exploring Disagreement, Difficulty and Calibration in Legal Case Outcome Classification. arXiv:2402.07214. doi:10.48550/arXiv.2402.07214

Zhou, X., Nie, Y., & Bansal, M. (2022). Distributed NLI: Learning to Predict Human Opinion Distributions for Language Reasoning. Findings of ACL 2022. arXiv:2104.08676. doi:10.48550/arXiv.2104.08676

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

DOI: 10.5281/zenodo.24294850