Reducing Supervision Uncertainty Induces Model Miscalibration

Leixin Zhang [1] **Çağrı Çöltekin** [1]

Abstract

Model calibration measures how well the model predicted probabilities align with their empirical accuracy. Recent studies show that deep learning models are often overconfident, and prior works primarily attribute this phenomenon to architectural and optimization-related factors. In this work, we argue that miscalibration also originates from the supervision target itself. To construct ground-truth labels, standard supervised learning pipelines commonly aggregate annotations from multiple annotators into a single label or simplify fine-grained judgments. We show that these operations discard uncertainty inherent in human judgments and lead to model miscalibration. In this study, we investigate the impact of preserving versus collapsing annotation uncertainty during training. Our results show that preserving annotation uncertainty substantially improves model calibration, achieves stronger predictive performance, and better reflects human judgments. These findings suggest that calibration depends not only on model architecture and optimization, but also on how uncertainty in human judgments is represented in the training signal. Notably, our temperature scaling experiments show that preserving annotation uncertainty during training largely eliminates the need for post-hoc calibration.

1. Introduction

Model calibration quantifies how well a model's predicted probabilities align with the empirical correctness of its outputs. A model is considered well calibrated if predictions assigned a confidence of 80% are correct approximately 80% of the time (Guo et al., 2017; Mukhoti et al., 2020; Minderer et al., 2021; Wang, 2023). Calibration is particularly important in real-world applications when predicted probabilities serve as a basis for downstream decision making. For instance, users may choose to adopt a model's prediction only when its confidence exceeds a predefined threshold, or abstain the prediction when confidence is low. Poorly calibrated models may exhibit unwarranted confidence in erroneous outputs or underestimate their own competence, making it difficult for users to appropriately assess uncertainty and trust model predictions.

A body of research has shown that modern deep learning models are often miscalibrated and tend to be overconfident in their predictions (Guo et al., 2017; Minderer et al., 2021). Prior work has identified several factors contributing to this phenomenon. At the optimization level, commonly used training objectives such as cross-entropy and negative log-likelihood (NLL) encourage models to concentrate probability mass on the target label (Mukhoti et al., 2020). At the architectural level and training procedure, Guo et al. (2017) show that modern neural networks' calibration can be influenced by factors such as network depth, width, batch normalization, weight decay, and training duration. Minderer et al. (2021) demonstrate that model architecture itself is a major determinant of calibration behavior. Beyond model design, calibration can further deteriorate under distribution shift and confidence estimates often fail to generalize reliably to out-of-distribution data (Ovadia et al., 2019).

In this study, we argue that model miscalibration can also arise from the construction of supervision targets. Specifically, we hypothesize that collapsing inherently uncertain human annotations can induce miscalibration. Modern supervised learning typically assumes a single ground-truth label for each instance during both training and evaluation. Within this paradigm, labels that deviate from the majority annotation are often treated as annotation noise or indicators of poor data quality. This implicitly assumes that category boundaries are well-defined and that every instance admits a unique correct label. However, uncertainty is pervasive in human concepts, language use, and annotation processes (Plank, 2022; Sorensen et al., 2024; Jiang & de Marneffe, 2022; Sap et al., 2020; Sang & Stanton, 2022; Huang & Yang, 2023). As noted by Wiebe et al. (2005), many human judgments contain both clear cases and borderline cases that are inherently difficult to resolve. Such uncertainty arises even in seemingly straightforward classification tasks. For example, people may show uncertainty and disagreement

[1]University of Tübingen, Germany. Correspondence to: Leixin Zhang <leixin.zhang@uni-tuebingen.de>, Çağrı Çöltekin <cagri.coeltekin@uni-tuebingen.de>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea.

on whether a color should be classified as "red" or "brown", or whether an animal should be categorized as a "dog" or a "wolf". More generally, category membership often exhibits a graded structure rather than sharp boundaries, such that non-prototypical instances naturally elicit varying judgments that lie along a continuum rather than fitting neatly into discrete categories (Rosch, 1973a;b).

While disagreement modeling in categorical labels has received increasing attention, uncertainty in graded judgments and Likert-scale annotations remains underexplored. In practice, Likert-scale ratings collected from multiple annotators are commonly aggregated through simple averaging, producing a single scalar supervision target. In some cases, ratings are collapsed into binary labels when downstream applications require a yes/no decision (Orlikowski et al., 2023; Zhang & Coltekin, 2026). Although these approaches simplify supervision representation, they systematically reduce uncertainty present in human judgments. In this study, we show that common aggregation operations, whether across annotators or across the ordinal structure of rating scales, induce model miscalibration. In contrast, preserving supervision uncertainty yields better-calibrated models and require considerably less post-hoc correction through techniques such as temperature scaling.

2. Related Work

Model Calibration Prior work has focused on improving the calibration of machine learning models through calibration techniques. Early approaches include Platt scaling (Platt et al., 1999), which fits a sigmoid function to transform model outputs into calibrated probabilities. Isotonic regression (Zadrozny & Elkan, 2002) is a non-parametric calibration method that learns a monotonic mapping from model scores to calibrated probabilities without assuming a specific functional form. More recently, temperature scaling has become one of the most widely adopted calibration methods for deep neural networks due to its simplicity and effectiveness (Guo et al., 2017). Subsequent work has explored more calibration strategies, such as adaptive temperature scaling (Uma et al., 2022; Balanya et al., 2024), and batch-level calibration techniques (Zhou et al., 2024). Li et al. (2025) proposed a lightweight corrector module that refines uncertainty estimates produced by the base model.

Calibration in Large Language Models. The study of calibration has recently extended to large language models (LLMs), where calibration is more challenging because predictions often free-form text generation rather than a single output of class labels. Existing approaches estimate uncertainty from token probabilities, the consistency of generated responses across multiple forward passes, verbalized confidence statements, or internal model representations (Li et al., 2025). Prior work has shown that LLMs frequently exhibit overconfidence when explicitly expressing confidence in their responses (Xiong et al., 2024; Groot & Valdenegro Toro, 2024).

Disagreement Modeling In traditional annotation practice, low inter-annotator agreement is commonly interpreted as poor annotation quality, and labels that deviate from the majority are treated as noise (Klie et al., 2023). However, a growing body of research, particularly in subjective NLP tasks, argues that annotation disagreement can encode meaningful signals about ambiguity, subjectivity, and differing human perspectives (Beigman & Beigman Klebanov, 2009; Plank, 2022). This perspective has motivated work on modeling annotation variation explicitly (Sorensen et al., 2024; Mokhberian et al., 2024; Cabitza et al., 2023). Several approaches incorporate disagreement during training. Early work proposed cost-sensitive learning, assigning lower weights to instances with higher disagreement (Sheng et al., 2008), or leveraging annotator confusion matrices to reweight losses (Plank et al., 2014). More recent studies adopt soft-label or label distribution learning, where targets are represented as the proportion of annotators selecting each class (Peterson et al., 2019; Uma et al., 2021; Lalor et al., 2017; Fornaciari et al., 2021; Glockner et al., 2024). Another direction of research models individual annotator perspectives explicitly. Some approaches introduce annotator-specific heads to predict each annotator's label (Rodrigues & Pereira, 2018; Mostafazadeh Davani et al., 2022), leverage annotator histories (Kanclerz et al., 2021), annotator ID (Mokhberian et al., 2024), and social attributes (Gordon et al., 2022; Zhang & Coltekin, 2026) to infer individual or group-level judgments. Recent work (Baan et al., 2022) also consider calibration evaluation in the context of human annotation disagreement and subjective tasks. Despite growing interest in modeling annotation variations, most existing methods are designed for classification tasks with discrete labels, overlooking the ordinal structure of subjective annotations (Zhang, 2025).

3. Methodology

We investigate this supervision problem with subjective, multi-annotator annotations, where each input instance x_i is labeled by multiple annotators using an ordinal Likert scale with K ordered categories. For each instance x_i, we denote the set of annotations as $\{y_i^{(a)}\}_{a=1}^{A_i}$, where $y_i^{(a)} \in \{1, \dots, K\}$ is the Likert-scale label provided by annotator a, and A_i is the number of annotators for instance i. We represent the probability of each Likert class as:

$$p_i(k) = \frac{1}{A_i} \sum_{a=1}^{A_i} \mathbb{I}\left[y_i^{(a)} = k\right], \tag{1}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function. Let $\mathcal{Y} = \{1 \prec 2 \prec \cdots \prec K\}$ denote an ordered label space. The empirical

annotation distribution obtained from multiple annotators for data item i is:

$$\mathbf{y}_i = \big(p_i(1), \ldots, p_i(K)\big).$$

3.1. Supervision Representations

To systematically study the effects of annotation aggregation on model calibration, we represent supervision targets using four different formats, reflecting different design choices in target construction.

Likert Soft Labels. We propose using the full empirical distribution over Likert categories as the training target:

$$\mathbf{y}_i = \big(p_i(1), \ldots, p_i(K)\big), \quad p_i(k) \in [0,1], \quad \sum_{k=1}^{K} p_i(k) = 1.$$

Values in the distribution represent the fraction of annotators selecting each Likert class, ordered from the first class to class K. This representation preserves both (i) the ordinal structure of the scale and (ii) disagreement among human judgments, thereby capturing both uncertainty in human perception and inter-annotator variation.

Likert Hard Labels. A common approach to supervision target construction in machine learning is to aggregate annotations from multiple annotators into a single label, typically the mean or median of the Likert responses. We rounded the mean value to the nearest category as the Likert Hard Label. While this preserves the ordinal or continuous nature of label structure, it removes information about disagreement among annotators.

Binary Soft Labels. Many downstream applications require a binary decision, such as whether a comment should be flagged as toxic and prevented from further dissemination. We thus transform Likert-scale ratings into binary labels using a midpoint threshold, assigning positive label to ratings $\geq (K-1)/2$ for K class Likert. We then represent each instance by the soft binary label distribution $\mathbf{y}_i = [1-p_i, p_i]$, where $p_i \in [0,1]$ denotes the fraction of annotators assigning the positive label and $1-p_i$ negative label. This approach preserves some uncertainty but collapses the ordinal structure of the original Likert scale.

Binary Hard Labels. This aggregation converts human responses into a single binary label, $y_i \in \{0,1\}$, for each instance (e.g., hate speech: *yes* or *no*). This discards both ordinal information and annotator disagreement.

3.2. Training Objectives

We consider several loss functions suitable for training models on these four different target formats.

Binary Cross-Entropy (BCE). For binary targets, we use standard binary cross-entropy. For hard labels, the target is either one or zero, representing the positive class; while for soft labels, $y_i \in [0,1]$ represents the fraction of annotators assigning the positive label. The BCE loss for a single instance is:

$$\mathcal{L}_{\text{BCE}} = -\left[y_i \log p_\theta(1 \mid x_i) + (1-y_i)\log p_\theta(0 \mid x_i)\right]. \tag{2}$$

The following loss functions can be applied to both Likert hard labels and Likert distributions.

Cramér Distance. The Cramér distance penalizes prediction errors quadratically to the distance between ordinal categories. For example, on a 5-point Likert scale (0,1,2,3,4), predicting category 4 when the true label is 1 should incur a larger penalty than predicting category 2. The Cramér distance captures it by measuring the (L_2) discrepancy between cumulative target and predicted distributions.[1]

Let

$$C_i(k) = \sum_{j=1}^{k} y_i(j), \qquad \hat{C}_i(k) = \sum_{j=1}^{k} \hat{p}_\theta(j \mid x_i), \tag{3}$$

denote the cumulative target and predicted distributions, respectively. The Cramér distance is defined as

$$\mathcal{L}_{\text{Cramér}} = \sum_{k=1}^{K} \Big(C_i(k) - \hat{C}_i(k)\Big)^2. \tag{4}$$

Given that distance-based loss functions such as EMD and Cramér distance assume equal spacing between adjacent Likert scale points, an assumption that may not hold in practice, we further propose two loss functions based on the concept of ordinal cumulative distributions. These loss functions preserve the ordinal structure of Likert ratings while allowing for more flexible modeling of perceived differences between rating levels.

Ordinal Cumulative Cross-Entropy (Cum-CE). To capture the ordinal structure of Likert-scale labels, we customize cross-entropy to measure the cumulative probability of the distribution. In this approach, a K-level Likert-scale problem is transformed into $K-1$ binary decisions with positive class as $y > k$, which preserves the natural ordering of the categories. The total loss is then computed as the sum of the losses over all $K-1$ thresholds.

[1]The formulation $\sum_{k=1}^{K} \Big|C_i(k) - \hat{C}_i(k)\Big|$ corresponds to the Earth Mover's Distance (EMD), also known as the Wasserstein distance, which measures the amount of probability mass that must be transported across ordinal categories. However, prior work has shown that EMD is often less effective as a training objective (Bellemare et al., 2017).

For each instance x_i, we define the cumulative target as

$$c_i(k) = \sum_{j=k+1}^{K} p_i(j) \tag{5}$$

representing the probability that the true label exceeds category k under the empirical annotation distribution.

Suppose the model outputs a K-class probability distribution:

$$\begin{aligned} \hat{\mathbf{y}}_\theta(x_i) &= \big(\hat{p}_\theta(1 \mid x_i), \ldots, \hat{p}_\theta(K \mid x_i)\big) \\ &= \text{softmax}\Big(g_1(x_i), \ldots, g_K(x_i)\Big), \end{aligned} \tag{6}$$

where $g_k(\cdot)$ denotes the logit associated with category k. The cumulative prediction is represented as:

$$\hat{c}_\theta(k \mid x_i) = \sum_{j=k+1}^{K} \hat{p}_\theta(j \mid x_i), \quad k = 1, \ldots, K-1. \tag{7}$$

The ordinal cumulative BCE loss measures the discrepancy between the cumulative target and the cumulative prediction across all thresholds:

$$\begin{aligned} \mathcal{L}_{\text{Cum-CE}} = -\sum_{k=1}^{K-1} \Big[& c_i(k) \log \hat{c}_\theta(k \mid x_i) \\ & + \big(1 - c_i(k)\big) \log \big(1 - \hat{c}_\theta(k \mid x_i)\big) \Big]. \end{aligned} \tag{8}$$

Ordinal Cumulative Kullback–Leibler Divergence (Cum-KL) For Likert targets, we further adapt the Kullback–Leibler (KL) divergence (Kullback & Leibler, 1951) using the same cumulative ordinal representation. The cumulative KL loss is defined as:

$$\begin{aligned} \mathcal{L}_{\text{Cum-KL}} = \sum_{k=1}^{K-1} \Bigg[& c_i(k) \log \frac{c_i(k)}{\hat{c}_\theta(k \mid x_i)} \\ & + \big(1 - c_i(k)\big) \log \frac{1 - c_i(k)}{1 - \hat{c}_\theta(k \mid x_i)} \Bigg]. \end{aligned} \tag{9}$$

where $(c_i(k))$ and $(\hat{c}_\theta(k \mid x_i))$ denote the target and predicted cumulative probabilities at threshold (k), respectively. By minimizing the divergence between cumulative distributions, the loss explicitly preserves ordinal structure while penalizing deviations across all cumulative decision boundaries.

4. Experiments

This section introduces the datasets used for the experiments in this study and experiment implementation details.

4.1. Datasets

We conduct experiments on four subjective datasets annotated by multiple annotators. Table 1 reports overall dataset statistics.

Dataset	Classes	Unique Items			Annot.
		Train	Val	Test	
Offensive	3	21,487	7,162	7,162	∼4
HateSpeech	3	3,594	1,198	1,198	∼11
Toxic	5	63,621	21,207	21,207	∼5
Aggression	7	13,789	4,596	4,596	∼12

Table 1. Overview of datasets used in our experiments. *Annot.* indicates the average number of annotations per item.

Offensive Language. The Offensive Language dataset (Sap et al., 2020) contains approximately 150k annotated items. To ensure reliable empirical annotation distributions, we filter out items annotated by fewer than three annotators. After filtering, the dataset comprises approximately 128.6k annotations over 35.8k unique items, with an average of four annotators per item. Each item is annotated with three categories: *no*, *maybe*, and *yes*. We map these categories to a three-point Likert scale (0 - 2).

Hate Speech. The hate speech dataset by Kennedy et al. (2020) provides graded annotations of hatefulness. Items annotated by fewer than four annotators are removed, resulting in approximately 67k annotations over 5,990 unique items. Labels are provided on a three-level Likert scale, where 0 denotes non-hateful content and higher values indicate increasing hate speech severity.

Toxicity. The toxicity dataset (Kumar et al., 2021) contains approximately 107.6k unique text instances, most of which are annotated by five or more annotators. We retain items with at least five annotations, obtaining approximately 106k items. Annotations follow a five-point Likert scale ranging from *not toxic* (0) to *extremely toxic* (4).

Aggression. The Wikipedia Talk dataset (Wulczyn et al., 2016) consists of approximately 100k text instances and agression annotation.[2] The dataset was annotated using a 7-point Likert scale ranging from very friendly to very aggressive, which we map to ordinal scores from 0 to 6. Since more than 80% of the original labels correspond to the neutral class, we randomly subsampled instances from this class, resulting in a more balanced subset of 23k instances for training and evaluation. Each instance was annotated by approximately 12 annotators on average.

4.2. Implementation

Data Split. All datasets are split at the level of unique text instances to prevent annotation leakage across splits. Each dataset is randomly split into training, validation, and test

[2]Dataset: Wikipedia Talk Labels Aggression.
Documentation: Wikimedia Detox Data Release.

sets with a 60% / 20% / 20% proportion, and experiments are repeated across 10 independent runs for robustness.

Model Architecture. All experiments are implemented with PyTorch (Paszke et al., 2019). Text inputs are encoded using the RoBERTa (Zhuang et al., 2021) and the pretrained Sentence-BERT model (Reimers & Gurevych, 2019), which produces fixed-dimensional sentence embeddings and has been shown to perform competitively for sentence-level semantic tasks (Reimers & Gurevych, 2019; Zhang et al., 2024). Following the encoder block, the model consists of two hidden layers with 512 and 256 neurons, respectively, each followed by dropout with a rate of 0.2. To isolate the effect of label representation, we keep the input representation and model architecture fixed across all experiments, varying only the output head and corresponding loss functions. This simple architecture ensures that observed performance differences can be attributed to differences in label supervision rather than model complexity or capacity.

Training Procedure Model parameters are optimized using the Adam optimizer (Kingma, 2014). We apply early stopping based on the performance on the validation set. Training is stopped if no improvement is observed for five consecutive epochs. The model checkpoint with the best validation performance is saved for evaluation on the test set. We repeat each experiment ten times using different random seeds for the train/validation/test splits. All reported results are averaged over these ten independent runs.

4.3. Evaluation

We evaluate models trained with four different label formats from several perspectives: predictive performance, distributional alignment, and model calibration with respect to crowd annotation uncertainty.

Predictive Performance. We evaluate predictive accuracy using the mean squared error (MSE) between the empirical mean of the multiple annotators and the model-predicted mean, where the latter is computed as the expectation of the predicted Likert distribution. For comparability with binary supervision approaches, we also report the Pearson correlation coefficient between the predicted and empirical mean ratings. We evaluate binary classification performance using F1 score and AUC. Predicted Likert distributions are converted into binary predictions using the same threshold employed during label construction.

Distribution Alignment. To evaluate how closely model predictions match the full annotation distribution, we use both KL divergence and Earth Mover's Distance (EMD). Unlike KL divergence, EMD explicitly accounts for the ordinal structure of Likert labels, penalizing predictions in proportion to the distance between rating categories.

Model Calibration. Expected Calibration Error (ECE) and calibration diagrams are widely used tools for evaluating and visualizing model calibration. ECE partitions predictions into confidence bins and measures the discrepancy between the average confidence and empirical accuracy within each bin:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right| \tag{10}$$

However, ECE assumes deterministic ground-truth labels and is therefore less suitable for subjective tasks with soft annotation distributions. To evaluate calibration with soft targets, we primarily use the Brier score:

$$\mathrm{Brier} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \left(\hat{p}_\theta(k \mid x_i) - y_i(k)\right)^2 \tag{11}$$

Lower Brier scores indicate better alignment between predicted probabilities and observed annotation distributions. To enable direct comparison with models trained using binary supervision, we apply a midpoint threshold to the Likert scale to obtain binary labels and compute the Brier score on the resulting binary predictions.

5. Results and Discussion

Across all datasets, models supervised by Likert distribution (Likert Soft) tend to have better performance, The result trained with RoBERTa backbone is shown in Table 2 and Sentence BERT in Table 3 in Section A.

5.1. Comparison of Hard vs. Soft Supervision

In this section, we compare hard label supervision (aggregated labels across multiple annotators) and soft distribution supervision (annotation distributions from multiple annotators), including performance comparisons between Likert hard versus Likert soft, and binary hard versus binary soft.

LIKERT HARD VS. LIKERT SOFT ON MEAN PREDICTION. Although hard Likert labels are directly optimized for mean-rating prediction, they discard information about rating dispersion. By preserving the full annotation distribution, soft supervision provides additional signals about uncertainty and disagreement, allowing the model to infer more accurate expected ratings. This effect is particularly pronounced on Toxic, where the MSE decreases by 33% (0.624 $\rightarrow$ 0.417). These results suggest that preserving the full annotation distribution enables models to better capture the central tendency of human judgments.

LIKERT HARD VS. LIKERT SOFT ON DISTRIBUTION ALIGNMENT. As expected, soft supervision more effectively captures the full annotation distribution. Unlike hard-label supervision, which collapses diverse annotator opinions into

Model	KL↓	EMD↓	Brier↓	Precision↑	Recall↑	F1↑	AUC↑	Mean-MSE↓	Mean-Pearson↑
Offensive									
Binary $_{\text{HARD}}$	-	-	0.131 ± 0.0021	0.805 ± 0.0076	0.815 ± 0.0138	0.810 ± 0.0048	0.897 ± 0.0030	-	0.754 ± 0.0069
Binary $_{\text{SOFT}}$	-	-	0.081 ± 0.0016	0.805 ± 0.0168	0.820 ± 0.0240	0.812 ± 0.0053	0.900 ± 0.0035	-	0.762 ± 0.0055
Cum_CE $_{\text{HARD}}$	0.432 ± 0.0183	0.413 ± 0.0051	0.086 ± 0.0040	0.833 ± 0.0192	0.768 ± 0.0346	0.798 ± 0.0109	0.898 ± 0.0023	0.311 ± 0.0100	0.760 ± 0.0042
Cum_KL $_{\text{HARD}}$	0.430 ± 0.0149	0.411 ± 0.0082	0.086 ± 0.0044	0.832 ± 0.0224	0.769 ± 0.0401	0.798 ± 0.0147	0.899 ± 0.0040	0.310 ± 0.0133	0.760 ± 0.0089
Cramér $_{\text{SOFT}}$	0.402 ± 0.0061	**0.403** ± 0.0075	**0.079** ± 0.0015	0.815 ± 0.0106	0.815 ± 0.0135	0.815 ± 0.0042	0.902 ± 0.0033	0.293 ± 0.0052	0.770 ± 0.0050
Cum_CE $_{\text{SOFT}}$	0.397 ± 0.0067	0.405 ± 0.0097	**0.079** ± 0.0015	0.808 ± 0.0203	0.824 ± 0.0238	0.815 ± 0.0039	**0.903** ± 0.0029	0.292 ± 0.0056	0.771 ± 0.0060
Cum_KL $_{\text{SOFT}}$	**0.395** ± 0.0051	0.410 ± 0.0077	**0.079** ± 0.0010	0.807 ± 0.0160	0.827 ± 0.0221	**0.816** ± 0.0046	**0.903** ± 0.0027	**0.291** ± 0.0039	**0.771** ± 0.0045
Hate Speech									
Binary $_{\text{HARD}}$	-	-	0.137 ± 0.0025	0.788 ± 0.0296	0.669 ± 0.0362	**0.722** ± 0.0123	0.883 ± 0.0048	-	0.742 ± 0.0125
Binary $_{\text{SOFT}}$	-	-	0.059 ± 0.0013	0.847 ± 0.0227	0.595 ± 0.0444	0.697 ± 0.0243	0.888 ± 0.0030	-	0.758 ± 0.0090
Cum_CE $_{\text{HARD}}$	0.353 ± 0.0272	**0.343** ± 0.0086	0.068 ± 0.0056	0.874 ± 0.0235	0.526 ± 0.0326	0.656 ± 0.0215	0.883 ± 0.0071	0.235 ± 0.0168	0.746 ± 0.0168
Cum_KL $_{\text{HARD}}$	0.366 ± 0.0371	0.348 ± 0.0066	0.071 ± 0.0060	0.890 ± 0.0177	0.504 ± 0.0402	0.642 ± 0.0286	0.883 ± 0.0054	0.243 ± 0.0198	0.743 ± 0.0184
Cramér $_{\text{SOFT}}$	0.283 ± 0.0076	0.345 ± 0.0061	**0.057** ± 0.0020	0.846 ± 0.0265	0.592 ± 0.0334	0.696 ± 0.0177	0.891 ± 0.0049	0.206 ± 0.0079	0.764 ± 0.0127
Cum_CE $_{\text{SOFT}}$	**0.281** ± 0.0049	0.345 ± 0.0087	**0.057** ± 0.0019	0.834 ± 0.0321	0.618 ± 0.0476	0.708 ± 0.0212	**0.892** ± 0.0038	**0.205** ± 0.0079	**0.766** ± 0.0128
Cum_KL $_{\text{SOFT}}$	**0.281** ± 0.0039	0.348 ± 0.0058	0.058 ± 0.0011	0.831 ± 0.0314	0.614 ± 0.0488	0.704 ± 0.0215	0.891 ± 0.0042	0.206 ± 0.0061	0.766 ± 0.0125
Toxic									
Binary $_{\text{HARD}}$	-	-	0.135 ± 0.0014	0.523 ± 0.0152	0.519 ± 0.0303	0.520 ± 0.0095	0.797 ± 0.0019	-	0.578 ± 0.0047
Binary $_{\text{SOFT}}$	-	-	0.052 ± 0.0006	0.467 ± 0.0181	0.643 ± 0.0261	0.540 ± 0.0062	0.802 ± 0.0033	-	0.594 ± 0.0057
Cum_CE $_{\text{HARD}}$	0.609 ± 0.0221	0.647 ± 0.0105	0.071 ± 0.0026	0.637 ± 0.0178	0.330 ± 0.0306	0.433 ± 0.0233	0.799 ± 0.0039	0.618 ± 0.0269	0.585 ± 0.0068
Cum_KL $_{\text{HARD}}$	0.615 ± 0.0248	0.649 ± 0.0095	0.073 ± 0.0023	0.644 ± 0.0270	0.312 ± 0.0298	0.418 ± 0.0216	0.799 ± 0.0035	0.624 ± 0.0241	0.585 ± 0.0062
Cramér $_{\text{SOFT}}$	0.463 ± 0.0012	**0.603** ± 0.0047	**0.051** ± 0.0005	0.477 ± 0.0203	0.640 ± 0.0278	**0.546** ± 0.0043	**0.806** ± 0.0035	**0.417** ± 0.0053	0.605 ± 0.0033
Cum_CE $_{\text{SOFT}}$	**0.461** ± 0.0011	0.609 ± 0.0054	**0.051** ± 0.0004	0.461 ± 0.0215	0.666 ± 0.0324	0.543 ± 0.0054	**0.806** ± 0.0024	**0.417** ± 0.0054	**0.606** ± 0.0042
Cum_KL $_{\text{SOFT}}$	0.462 ± 0.0015	0.608 ± 0.0053	**0.051** ± 0.0006	0.473 ± 0.0137	0.642 ± 0.0244	0.544 ± 0.0053	**0.806** ± 0.0034	**0.417** ± 0.0062	0.605 ± 0.0043
Aggression									
Binary $_{\text{HARD}}$	-	-	0.085 ± 0.0016	0.877 ± 0.0090	0.834 ± 0.0135	0.855 ± 0.0049	0.953 ± 0.0022	-	0.835 ± 0.0057
Binary $_{\text{SOFT}}$	-	-	0.033 ± 0.0011	0.863 ± 0.0163	0.853 ± 0.0159	0.858 ± 0.0035	0.953 ± 0.0015	-	0.856 ± 0.0049
Cum_CE $_{\text{HARD}}$	0.493 ± 0.0375	0.477 ± 0.0218	0.046 ± 0.0040	0.914 ± 0.0231	0.782 ± 0.0296	0.842 ± 0.0076	0.955 ± 0.0025	0.308 ± 0.0338	0.891 ± 0.0036
Cum_KL $_{\text{HARD}}$	0.522 ± 0.0199	0.475 ± 0.0197	0.046 ± 0.0035	0.907 ± 0.0134	0.798 ± 0.0172	0.849 ± 0.0068	0.955 ± 0.0029	0.310 ± 0.0338	0.891 ± 0.0048
Cramér $_{\text{SOFT}}$	0.293 ± 0.0037	**0.410** ± 0.0096	0.031 ± 0.0005	0.873 ± 0.0177	0.846 ± 0.0262	0.859 ± 0.0081	**0.957** ± 0.0022	0.202 ± 0.0047	0.899 ± 0.0033
Cum_CE $_{\text{SOFT}}$	0.288 ± 0.0030	0.415 ± 0.0081	**0.030** ± 0.0002	0.868 ± 0.0148	0.856 ± 0.0188	**0.862** ± 0.0053	**0.957** ± 0.0022	**0.201** ± 0.0035	**0.899** ± 0.0025
Cum_KL $_{\text{SOFT}}$	**0.287** ± 0.0040	0.413 ± 0.0043	**0.030** ± 0.0003	0.876 ± 0.0144	0.842 ± 0.0204	0.858 ± 0.0050	**0.957** ± 0.0022	0.202 ± 0.0049	0.898 ± 0.0031

Table 2. Model performance of under different supervision representations (binary vs. Likert-scale; hard vs. soft targets) using RoBERTa backbone. **Bold** denotes the best performance within each column.

a single target label, soft supervision preserves information about disagreement, enabling the model to better reflect the uncertainty and variability present in crowd annotations. The gains are especially large on datasets with richer label spaces.

BINARY HARD VS. BINARY SOFT ON CLASSIFICATION PERFORMANCE. An important observation is that preserving annotation uncertainty does not introduce a trade-off between uncertainty modeling and binary classification performance. Despite being trained with softer supervision targets, the resulting models achieve classification accuracy comparable to or better than models trained on binary hard labels derived from majority voting, even though the latter are directly optimized for the majority-vote F1 objective used in evaluation. For instance, on the Toxic dataset, soft supervision improves the F1 score from 0.520 to 0.540 compared with hard binary supervision. The results suggest that uncertainty-aware targets may act as a regularizer, encouraging more robust representations and improving generalization.

HARD SUPERVISION VS. SOFT SUPERVISION ON CALIBRATION. The advantage of soft supervision also extends to model calibration. Across all datasets, models trained with soft targets consistently achieve lower Brier scores, suggest-

ing that their predicted probabilities are better aligned with observed annotation distributions. This trend holds for both binary and Likert supervision settings.

Taken together, these results suggest that the primary limitation of hard-label supervision is the loss of information about annotator disagreement. Preserving the full annotation distribution enables models to better capture uncertainty and variability in human judgments, resulting in improved calibration and distributional alignment while maintaining competitive predictive performance.

5.2. Comparison of Binary vs. Likert Supervision

In this subsection, we compare models trained with Likert-scale distribution and binary distribution.

LIKERT SOFT VS. BINARY SOFT ON PREDICTIVE PERFORMANCE. We convert predicted Likert distributions into binary predictions using a midpoint threshold for binary classification comparison. Compared with models trained on binary distributions, models trained on full Likert distributions eventually achieve comparable performance on F1 scores. Moreover, across all four datasets, models trained with Likert distributions consistently achieve the highest AUC scores, suggesting that preserving finer-grained supervision provide more robust classification performance and fine-grained supervision preserves information that remains useful even when evaluation is reduced to a binary decision.

LIKERT SOFT VS. BINARY SOFT ON MEAN PREDICTION. Binary supervision provides a substantially compressed representation of annotator judgments, making direct recovery of the underlying mean rating more challenging. We instead evaluate the Pearson correlation between model predictions and the empirical mean annotation score from multiple annotators. For binary supervision, we use the predicted probability of the positive class as an indicator of the underlying mean rating. For Likert supervision, we compute the expected value of the predicted distribution. The results show that Likert supervision consistently achieves higher Pearson correlations than binary supervision. For example, the correlation increases from 0.856 under binary supervision to 0.898 when training with Likert distributions.

LIKERT SOFT VS. BINARY SOFT ON CALIBRATION. The calibration results further confirm that the benefits of Likert distribution supervision. By preserving uncertainty about label assignment, Likert-distribution supervision produces probability estimates that more closely match observed annotation frequencies. They achieves the lowest Brier scores, significantly outperforming binary supervision across datasets (see Sections C.1 and C.2 for significance test results). These results indicate that preserving finer-grained annotation information yields better-calibrated models.

6. Post-Calibration Effect: Temperature Scaling

To investigate whether the observed miscalibration arises from systematic overconfidence or underconfidence, we apply temperature scaling as a diagnostic intervention. Specifically, we focus on calibration for multiclass Likert predictions.[3] Because temperature scaling adjusts prediction confidence without changing class rankings, the optimal temperature provides insight into whether model predictions are overly sharp or overly diffuse relative to empirical annotation distributions. Given logits $g_k(x_i)$ produced by a trained model, temperature scaling adjusts the predictive distribution as:

$$\hat{p}_\theta^{(T)}(k \mid x_i) = \frac{\exp(g_k(x_i)/T)}{\sum_{j=1}^{K} \exp(g_j(x_i)/T)}, \tag{12}$$

where $T > 1$ smooths predicted distributions and $T < 1$ sharpens the distributions. If predicted distributions are systematically sharper than the empirical annotation distributions, increasing the temperature score should improve alignment and calibration scores; conversely, lower temperatures can correct overly flat outputs. We investigate calibration behavior under different temperatures (T $\in \{0.2, 0.5, 1.0, 1.5, 2.0, 4.0\}$).

A pattern emerges from the temperature-scaling analysis: As shown in Figure 1, models trained with soft targets achieve their best calibration at T=1, indicating that the uncertainty encoded in their original predictions is already well aligned with empirical annotation frequencies (indicated by Brier score) and empirical correctness (indicated by ECE score). In contrast, models trained with hard targets consistently benefit from temperatures greater than one, implying that their predictions are systematically overconfident. This observation provides direct evidence that collapsing annotation disagreement into a single target distribution introduces miscalibration (over-confidence) during training. The difference between hard and soft supervision is particularly evident on the Toxic and Aggression datasets, where hard-supervised models require substantial post-hoc smoothing (T=2) to achieve their best calibration, while soft-supervised models are already well calibrated without temperature scaling.

More importantly, temperature scaling cannot fully compensate for the information lost during label aggregation. Even after calibration, hard-supervised models remain less calibrated than their soft-supervised counterparts on several

[3]This differs from the results reported in Table 2, where Brier scores are computed on binarized labels for comparison with binary supervision.

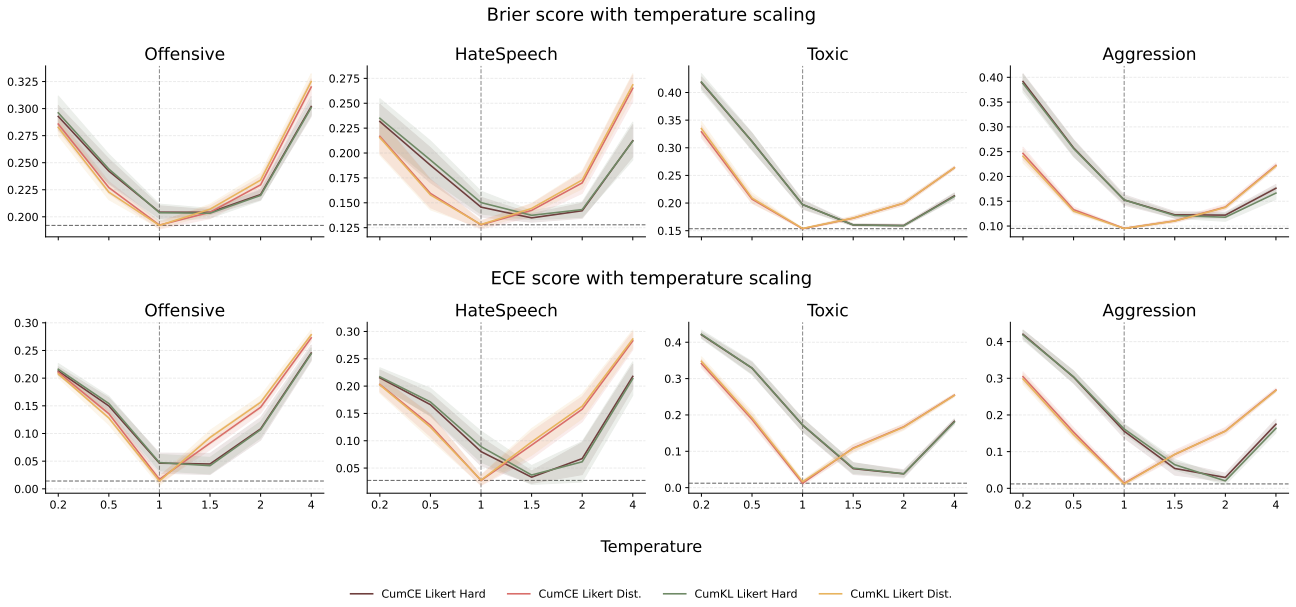

Figure 1. Temperature scaling assessment for RoBERTa models trained with hard and soft Likert supervision. Calibration quality is evaluated using the Brier score, which measures the discrepancy between predicted probability distributions and the empirical annotation distributions, and the ECE score, which quantifies the mismatch between model confidence and observed accuracy in majority-vote classification. Prediction sharpness is adjusted using temperature scaling. Hard label supervision is shown in colors: ■ Purple, ■ Green. Soft label supervision is shown in colors: ■ Orange, ■ Red.

datasets. From a theoretical perspective, temperature scaling applies a global adjustment to the sharpness of predicted distributions. However, uncertainty is inherently instance-dependent: some examples are clear and unambiguous, while others are genuinely uncertain (Zhang & Çöltekin, 2026). The optimal degree of smoothing varies across instances, and a single temperature parameter cannot fully recover the uncertainty information discarded during label aggregation. Consequently, even if an optimal temperature exists in principle, temperature scaling remains an imperfect substitute for soft supervision that explicitly preserves annotation uncertainty. More importantly, the optimal temperature is not known a priori and must be estimated from validation data. In real-world deployment, there is no guarantee that the temperature selected on a validation set will remain optimal for unseen test data. As a result, a perfectly calibrated temperature value is unlikely to be available in practice, limiting the effectiveness of post-hoc temperature scaling as a replacement for uncertainty-aware supervision.

7. Conclusion

This work revisits a common but largely unquestioned practice in supervision construction: collapsing annotation uncertainty from multi-annotator Likert ratings, either by aggregating annotator judgments into a single target value or by reducing fine-grained Likert-scale annotations to coarser supervision such as binary decisions. We show that this reduction of supervision uncertainty systematically induces model miscalibration. By aggregating diverse human judgments into a single "ground-truth" label, hard-label supervision discards information about human disagreement and semantic ambiguity that is essential for learning well-calibrated predictive distributions. Across four subjective NLP datasets, models trained on full Likert annotation distributions consistently achieve better calibration and closer alignment with empirical human annotations than models trained on uncertainty-reduced supervision formats. Uncertainty-preserving supervision also yields competitive predictive performance, demonstrating that preserving annotation uncertainty can both improve the reliability of confidence estimates and maintain strong task performance.

Our temperature-scaling analysis provides further evidence that the source of miscalibration originates in the supervision signal itself. Models trained with hard labels are systematic overconfident, whereas models trained on annotation distributions are often best calibrated without further calibration adjustment. Taken together, our findings suggest that many calibration problems are not solely a consequence of model architecture or optimization, but also arise from the way supervision signals are constructed. Preserving annotation uncertainty during training provides a simple and effective mechanism for improving calibration.

Impact Statement

Reliable uncertainty estimation is essential for the safe deployment of AI system. Our work demonstrates that a common supervision practice: aggregating multiple annotations into simplified supervision formats can systematically induce model miscalibration by removing information about human uncertainty and human disagreement. These findings suggest that calibration should not be viewed solely as a post-hoc modeling problem. The design of supervision signals can have a substantial impact on the reliability of model confidence estimates. By preserving annotation uncertainty during training, models produce probability estimates that are better aligned with observed human judgments and require less post-hoc calibration. More broadly, this work highlights the importance of treating annotator disagreement as a meaningful source of information rather than noise to be eliminated. Better-calibrated models can support more transparent and trustworthy decision-making by communicating uncertainty more faithfully, particularly in applications involving subjective, ambiguous, or contested concepts. We therefore advocate for greater consideration of uncertainty-preserving supervision when developing AI systems intended to operate in human-centered settings.

References

Baan, J., Aziz, W., Plank, B., and Fernandez, R. Stop measuring calibration when humans disagree. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1892–1915, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.124. URL https://aclanthology.org/2022.emnlp-main.124/.

Balanya, S. A., Maronas, J., and Ramos, D. Adaptive temperature scaling for robust calibration of deep neural networks. *Neural Computing and Applications*, 36(14): 8073–8095, 2024.

Beigman, E. and Beigman Klebanov, B. Learning with annotation noise. In Su, K.-Y., Su, J., Wiebe, J., and Li, H. (eds.), *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 280–287, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL https://aclanthology.org/P09-1032.

Bellemare, M. G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S., and Munos, R. The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017.

Cabitza, F., Campagner, A., and Basile, V. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 6860–6868, 2023.

Fornaciari, T., Uma, A., Paun, S., Plank, B., Hovy, D., Poesio, M., et al. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2021.

Glockner, M., Staliūnaitė, I., Thorne, J., Vallejo, G., Vlachos, A., and Gurevych, I. Ambifc: Fact-checking ambiguous claims with evidence. *Transactions of the Association for Computational Linguistics*, 12:1–18, 2024.

Gordon, M. L., Lam, M. S., Park, J. S., Patel, K., Hancock, J., Hashimoto, T., and Bernstein, M. S. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2022.

Groot, T. and Valdenegro Toro, M. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In Ovalle, A., Chang, K.-W., Cao, Y. T., Mehrabi, N., Zhao, J., Galstyan, A., Dhamala, J., Kumar, A., and Gupta, R. (eds.), *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pp. 145–171, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.trustnlp-1.13. URL https://aclanthology.org/2024.trustnlp-1.13/.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Huang, J. and Yang, D. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7591–7609, 2023.

Jiang, N.-J. and de Marneffe, M.-C. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10: 1357–1374, 2022. URL https://aclanthology.org/2022.tacl-1.78.

Kanclerz, K., Figas, A., Gruza, M., Kajdanowicz, T., Kocoń, J., Puchalska, D., and Kazienko, P. Controversy and conformity: from generalized to personalized aggressiveness detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5915–5926, 2021.

Kennedy, C. J., Bacon, G., Sahn, A., and von Vacano, C. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*, 2020.

Kingma, D. P. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Klie, J.-C., Webber, B., and Gurevych, I. Annotation error detection: Analyzing the past and present for a more coherent future. *Computational Linguistics*, 49(1):157–198, 2023.

Kullback, S. and Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86, 1951.

Kumar, D., Kelley, P. G., Consolvo, S., Mason, J., Bursztein, E., Durumeric, Z., Thomas, K., and Bailey, M. Designing toxic content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pp. 299–318, 2021.

Lalor, J. P., Wu, H., and Yu, H. Soft label memorization-generalization for natural language inference. *arXiv preprint arXiv:1702.08563*, 2017.

Li, R., Long, J., Qi, M., Xia, H., Sha, L., Wang, P., and Sui, Z. Towards harmonized uncertainty estimation for large language models. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22938–22953, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1118. URL https://aclanthology.org/2025.acl-long.1118/.

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., and Lucic, M. Revisiting the calibration of modern neural networks. *Advances in neural information processing systems*, 34:15682–15694, 2021.

Mokhberian, N., Marmarelis, M., Hopp, F., Basile, V., Morstatter, F., and Lerman, K. Capturing perspectives of crowdsourced annotators in subjective learning tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7337–7349, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.407. URL https://aclanthology.org/2024.naacl-long.407.

Mostafazadeh Davani, A., Díaz, M., and Prabhakaran, V. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022. doi: 10.1162/tacl_a_00449. URL https://aclanthology.org/2022.tacl-1.6.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., and Dokania, P. Calibrating deep neural networks using focal loss. *Advances in neural information processing systems*, 33:15288–15299, 2020.

Orlikowski, M., Röttger, P., Cimiano, P., and Hovy, D. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1017–1029, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.88. URL https://aclanthology.org/2023.acl-short.88.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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

Plank, B. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10671–10682, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.731. URL https://aclanthology.org/2022.emnlp-main.731.

Plank, B., Hovy, D., and Søgaard, A. Learning part-of-speech taggers with inter-annotator agreement loss. In Wintner, S., Goldwater, S., and Riezler, S. (eds.), *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 742–751, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1078. URL https://aclanthology.org/E14-1078.

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

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Rodrigues, F. and Pereira, F. Deep learning from crowds. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Rosch, E. H. On the internal structure of perceptual and semantic categories. In *Cognitive development and acquisition of language*, pp. 111–144. Elsevier, 1973a.

Rosch, E. H. Natural categories. *Cognitive psychology*, 4 (3):328–350, 1973b.

Sang, Y. and Stanton, J. The origin and value of disagreement among data labelers: A case study of individual differences in hate speech annotation. In *International Conference on Information*, pp. 425–444. Springer, 2022.

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., and Choi, Y. Social bias frames: Reasoning about social and power implications of language. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5477–5490, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.486. URL https://aclanthology.org/2020.acl-main.486/.

Sheng, V. S., Provost, F., and Ipeirotis, P. G. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 614–622, 2008.

Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

Uma, A., Almanea, D., and Poesio, M. Scaling and disagreements: Bias, noise, and ambiguity. *Frontiers in Artificial Intelligence*, 5:818451, 2022.

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., and Poesio, M. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, 2021.

Wang, C. Calibration in deep learning: A survey of the state-of-the-art. *arXiv preprint arXiv:2308.01222*, 2023.

Wiebe, J., Wilson, T., and Cardie, C. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210, 2005.

Wulczyn, E., Thain, N., and Dixon, L. Wikipedia detox, 2016. URL https://doi.org/10.6084/m9.figshare.4054689.

Xiong, M., Hu, Z., Lu, X., LI, Y., Fu, J., He, J., and Hooi, B. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In Kim, B., Yue, Y., Chaudhuri, S., Fragkiadaki, K., Khan, M., and Sun, Y. (eds.), *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024.

Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699, 2002.

Zhang, L. Proposal: From one-fit-all to perspective aware modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 1016–1025, 2025.

Zhang, L. and Çöltekin, Ç. Quantifying and predicting disagreement in graded human ratings. In *Proceedings of the Fifth Workshop on Perspectivist Approaches to NLP (NLPerspectives) at Language Resources and Evaluation Conference (LREC) 2026*. ELRA, 2026.

Zhang, L. and Coltekin, C. Modeling human perspectives with socio-demographic representations. *arXiv preprint arXiv:2604.18069*, 2026.

Zhang, L., Burian, D., John, V., and Bojar, O. Unveiling semantic information in sentence embeddings. In *Proceedings of the Fifth International Workshop on Designing Meaning Representations@ LREC-COLING 2024*, pp. 39–47, 2024.

Zhou, H., Wan, X., Proleev, L., Mincu, D., Chen, J., Heller, K., and Roy, S. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In *International Conference on Learning Representations*, volume 2024, pp. 49–70, 2024.

Zhuang, L., Wayne, L., Ya, S., and Jun, Z. A robustly optimized BERT pre-training approach with post-training. In Li, S., Sun, M., Liu, Y., Wu, H., Liu, K., Che, W., He, S., and Rao, G. (eds.), *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pp. 1218–1227, Huhhot, China, August 2021. Chinese Information Processing Society of China. URL https://aclanthology.org/2021.ccl-1.108/.

A. Overall Results of Model Performance using Sentence-BERT

Dataset	KL↓	EMD↓	Brier↓	Precision↑	Recall↑	F1↑	AUC↑	Mean-MSE↓	Mean-Pearson ↑
Offensive									
Binary $_{\text{HARD}}$	-	-	0.136 ± 0.0022	0.811 ± 0.0082	0.776 ± 0.0196	0.793 ± 0.0080	0.889 ± 0.0033	-	0.741 ± 0.0058
Binary $_{\text{SOFT}}$	-	-	0.086 ± 0.0018	0.800 ± 0.0169	0.799 ± 0.0249	0.799 ± 0.0052	0.891 ± 0.0029	-	0.746 ± 0.0058
Cum_CE $_{\text{HARD}}$	0.448 ± 0.0104	**0.418** ± 0.0087	0.092 ± 0.0022	0.831 ± 0.0167	0.738 ± 0.0292	0.781 ± 0.0098	0.890 ± 0.0026	0.331 ± 0.0063	0.743 ± 0.0043
Cum_KL $_{\text{HARD}}$	0.443 ± 0.0111	0.424 ± 0.0118	0.091 ± 0.0033	0.830 ± 0.0164	0.738 ± 0.0325	0.780 ± 0.0133	0.889 ± 0.0053	0.330 ± 0.0100	0.743 ± 0.0080
Cramér $_{\text{SOFT}}$	0.414 ± 0.0065	0.419 ± 0.0071	0.086 ± 0.0014	0.802 ± 0.0156	0.797 ± 0.0223	0.799 ± 0.0052	0.892 ± 0.0030	0.315 ± 0.0053	0.751 ± 0.0050
Cum_CE $_{\text{SOFT}}$	0.410 ± 0.0045	0.420 ± 0.0065	**0.085** ± 0.0017	0.812 ± 0.0162	0.787 ± 0.0262	0.799 ± 0.0073	**0.893** ± 0.0042	**0.312** ± 0.0059	**0.754** ± 0.0057
Cum_KL $_{\text{SOFT}}$	**0.409** ± 0.0039	0.422 ± 0.0088	**0.085** ± 0.0017	0.792 ± 0.0105	0.810 ± 0.0212	**0.801** ± 0.0066	0.892 ± 0.0035	**0.312** ± 0.0068	0.752 ± 0.0057
Hate Speech									
Binary $_{\text{HARD}}$	-	-	0.142 ± 0.0037	0.766 ± 0.0349	0.667 ± 0.0523	**0.710** ± 0.0160	0.875 ± 0.0058	-	0.733 ± 0.0140
Binary $_{\text{SOFT}}$	-	-	0.061 ± 0.0020	0.843 ± 0.0408	0.549 ± 0.0645	0.661 ± 0.0374	0.879 ± 0.0042	-	0.750 ± 0.0120
Cum_CE $_{\text{HARD}}$	0.379 ± 0.0371	0.350 ± 0.0128	0.073 ± 0.0074	0.856 ± 0.0300	0.503 ± 0.0581	0.631 ± 0.0413	0.873 ± 0.0071	0.251 ± 0.0247	0.731 ± 0.0232
Cum_KL $_{\text{HARD}}$	0.373 ± 0.0254	0.350 ± 0.0113	0.073 ± 0.0049	0.862 ± 0.0285	0.504 ± 0.0499	0.634 ± 0.0332	0.871 ± 0.0067	0.249 ± 0.0156	0.730 ± 0.0158
Cramér $_{\text{SOFT}}$	0.291 ± 0.0041	0.354 ± 0.0110	**0.060** ± 0.0012	0.843 ± 0.0323	0.570 ± 0.0445	0.678 ± 0.0224	0.880 ± 0.0061	0.214 ± 0.0053	0.755 ± 0.0105
Cum_CE $_{\text{SOFT}}$	0.289 ± 0.0061	0.354 ± 0.0066	**0.060** ± 0.0017	0.820 ± 0.0201	0.597 ± 0.0548	0.688 ± 0.0312	0.880 ± 0.0051	0.214 ± 0.0080	0.754 ± 0.0126
Cum_KL $_{\text{SOFT}}$	**0.287** ± 0.0066	**0.349** ± 0.0061	**0.060** ± 0.0018	0.828 ± 0.0121	0.581 ± 0.0379	0.682 ± 0.0227	**0.881** ± 0.0052	**0.212** ± 0.0082	**0.756** ± 0.0123
Toxic									
Binary $_{\text{HARD}}$	-	-	0.138 ± 0.0015	0.516 ± 0.0154	0.482 ± 0.0267	0.497 ± 0.0098	0.782 ± 0.0026	-	0.559 ± 0.0041
Binary $_{\text{SOFT}}$	-	-	0.054 ± 0.0005	0.453 ± 0.0136	0.622 ± 0.0273	0.524 ± 0.0031	0.788 ± 0.0022	-	0.574 ± 0.0044
Cum_CE $_{\text{HARD}}$	0.631 ± 0.0300	0.660 ± 0.0115	0.074 ± 0.0042	0.624 ± 0.0285	0.292 ± 0.0433	0.395 ± 0.0339	0.782 ± 0.0028	0.647 ± 0.0340	0.560 ± 0.0065
Cum_KL $_{\text{HARD}}$	0.617 ± 0.0214	0.657 ± 0.0098	0.073 ± 0.0043	0.626 ± 0.0314	0.295 ± 0.0450	0.397 ± 0.0362	0.782 ± 0.0025	0.635 ± 0.0327	0.561 ± 0.0051
Cramér $_{\text{SOFT}}$	0.471 ± 0.0017	**0.615** ± 0.0038	**0.053** ± 0.0005	0.454 ± 0.0114	0.631 ± 0.0174	**0.528** ± 0.0027	**0.791** ± 0.0011	**0.434** ± 0.0048	**0.584** ± 0.0034
Cum_CE $_{\text{SOFT}}$	**0.470** ± 0.0012	0.621 ± 0.0059	**0.053** ± 0.0004	0.450 ± 0.0171	0.634 ± 0.0356	0.525 ± 0.0044	**0.791** ± 0.0015	0.435 ± 0.0046	0.583 ± 0.0053
Cum_KL $_{\text{SOFT}}$	**0.470** ± 0.0016	0.621 ± 0.0046	**0.053** ± 0.0005	0.452 ± 0.0217	0.629 ± 0.0448	0.525 ± 0.0050	**0.791** ± 0.0015	0.436 ± 0.0051	0.582 ± 0.0053
Aggression									
Binary $_{\text{HARD}}$	-	-	0.095 ± 0.0030	0.862 ± 0.0150	0.815 ± 0.0167	0.838 ± 0.0050	0.942 ± 0.0033	-	0.814 ± 0.0094
Binary $_{\text{SOFT}}$	-	-	0.038 ± 0.0009	0.841 ± 0.0163	0.847 ± 0.0179	0.843 ± 0.0067	0.944 ± 0.0026	-	0.843 ± 0.0033
Cum_CE $_{\text{HARD}}$	0.526 ± 0.0388	0.508 ± 0.0156	0.052 ± 0.0033	0.883 ± 0.0164	0.783 ± 0.0170	0.830 ± 0.0065	0.944 ± 0.0030	0.360 ± 0.0266	0.869 ± 0.0064
KL_Loss $_{\text{HARD}}$	0.527 ± 0.0410	0.511 ± 0.0213	0.054 ± 0.0055	0.892 ± 0.0160	0.772 ± 0.0386	0.827 ± 0.0166	0.944 ± 0.0034	0.360 ± 0.0332	0.870 ± 0.0056
Cramér $_{\text{SOFT}}$	0.322 ± 0.0070	**0.440** ± 0.0077	0.037 ± 0.0015	0.838 ± 0.0159	0.854 ± 0.0182	**0.846** ± 0.0078	**0.947** ± 0.0039	0.238 ± 0.0080	0.879 ± 0.0048
Cum_CE $_{\text{SOFT}}$	**0.316** ± 0.0073	0.451 ± 0.0050	**0.036** ± 0.0016	0.849 ± 0.0191	0.843 ± 0.0254	0.845 ± 0.0084	**0.947** ± 0.0039	**0.237** ± 0.0081	**0.880** ± 0.0053
Cum_KL $_{\text{SOFT}}$	**0.316** ± 0.0066	0.448 ± 0.0109	0.037 ± 0.0016	0.851 ± 0.0265	0.840 ± 0.0335	0.844 ± 0.0104	**0.947** ± 0.0035	**0.237** ± 0.0083	0.879 ± 0.0048

Table 3. Model performance of under different supervision representations (binary vs. Likert-scale; hard vs. soft targets) using Sentence-BERT backbone. **Bold** denotes the best performance within each column.

B. Post-Calibration Effect using Sentence-BERT

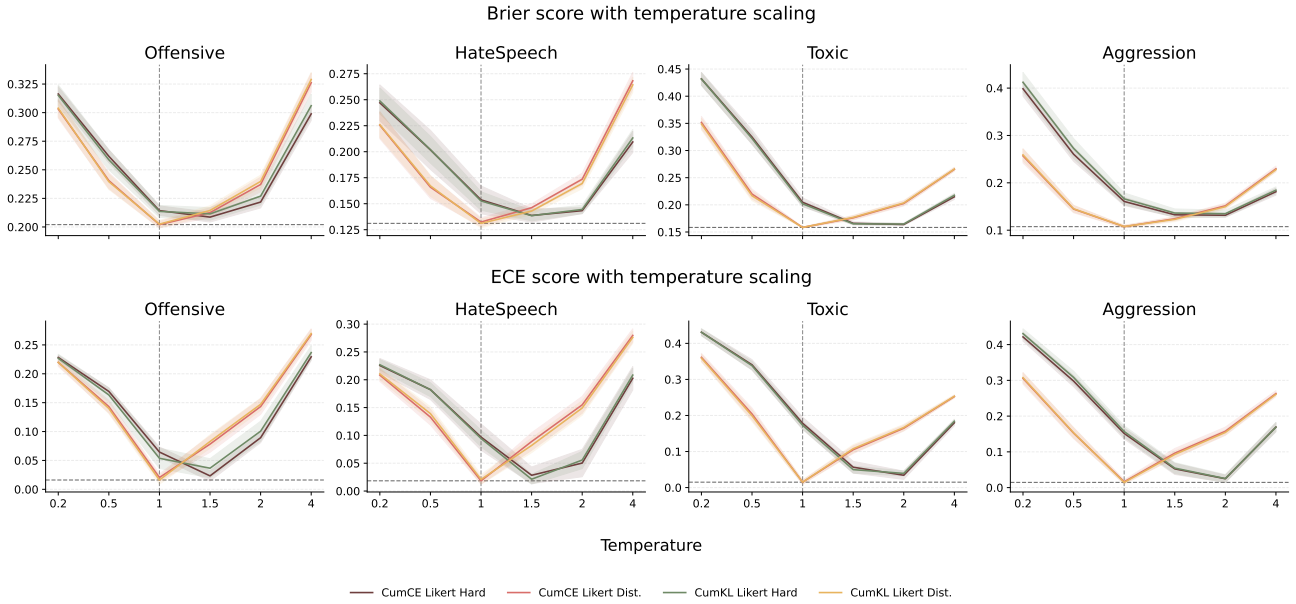

Figure 2. Temperature scaling assessment for Sentence-BERT models trained with hard and soft Likert supervision. Calibration quality is evaluated using the Brier score, which measures the discrepancy between predicted probability distributions and the empirical annotation distributions, and the ECE score, which quantifies the mismatch between model confidence and observed accuracy in majority-vote classification. Prediction sharpness is adjusted using temperature scaling. Hard label supervision is shown in colors: ■ Purple, ■ Green. Soft label supervision is shown in colors: ■ Orange, ■ Red.

C. Significance Test

This section presents statistical significance tests on four datasets, where significance is assessed via bootstrap resampling with 1,000 iterations. In the following tables, each row corresponds to models using label representations with some aggregation, and each column corresponds to models predicting Likert distributions. Each cell shows the pairwise difference (Δ = column method − row method) for the corresponding metric. For Brier, smaller (negative) values indicate better performance of distribution learning; for F1, larger (positive) values indicate better performance of distribution learning. Significance levels are indicated by stars: *** $p \leq 0.01$, ** $0.01 < p \leq 0.05$, * $0.05 < p \leq 0.1$.

C.1. Significance Test for Models Using RoBERTa

		Brier			F_1		
Dataset	**Model**	**Cramér**	**Cum_CE $_{SOFT}$**	**Cum_KL $_{SOFT}$**	**Cramér**	**Cum_CE $_{SOFT}$**	**Cum_KL $_{SOFT}$**
Offensive	Binary $_{HARD}$	-0.004***	-0.004***	-0.004***	0.011***	0.011***	0.009***
	Binary $_{SOFT}$	-0.001	-0.000	-0.001	0.008***	0.009**	0.006*
	Cum_CE $_{HARD}$	-0.014***	-0.014***	-0.014***	0.043***	0.044***	0.041***
	Cum_KL $_{HARD}$	-0.007***	-0.007***	-0.007***	0.016***	0.016***	0.014***
Hate Speech	Binary $_{HARD}$	-0.008***	-0.009***	-0.008***	-0.006	-0.001	-0.002
	Binary $_{SOFT}$	-0.003***	-0.004***	-0.002	0.020*	0.026*	0.025*
	Cum_CE $_{HARD}$	-0.008***	-0.009***	-0.007***	0.028***	0.034***	0.032***
	Cum_KL $_{HARD}$	-0.009***	-0.010***	-0.008***	0.063***	0.070***	0.068***
Toxic	Binary $_{HARD}$	-0.003***	-0.003***	-0.003***	-0.141***	-0.098***	-0.190***
	Binary $_{SOFT}$	-0.002***	-0.003***	-0.003***	-0.015	0.028***	-0.063***
	Cum_CE $_{HARD}$	-0.009***	-0.009***	-0.009***	0.056***	0.099***	0.008
	Cum_KL $_{HARD}$	-0.006***	-0.007***	-0.007***	0.045***	0.087***	-0.003
Aggression	Binary $_{HARD}$	-0.032***	-0.032***	-0.032***	0.032***	0.024***	0.026***
	Binary $_{SOFT}$	-0.027***	-0.027***	-0.027***	0.066***	0.058***	0.061***
	Cum_CE $_{HARD}$	-0.015***	-0.015***	-0.015***	0.014***	0.006	0.008**
	Cum_KL $_{HARD}$	-0.014***	-0.014***	-0.014***	0.015***	0.006*	0.009***

Table 4. Statistical significance tests of the performance differences between full Likert-distribution supervision and uncertainty-reduced supervision using RoBERTa.

C.2. Significance Test for Models Using Sentence-BERT

		Brier			F_1		
Dataset	**Model**	**Cramér**	**Cum_CE $_{SOFT}$**	**Cum_KL $_{SOFT}$**	**Cramér**	**Cum_CE $_{SOFT}$**	**Cum_KL $_{SOFT}$**
Offensive	Binary $_{HARD}$	-0.002**	-0.003***	-0.003**	0.004	0.004	0.011***
	Binary $_{SOFT}$	0.000	-0.001	-0.000	-0.002	-0.002	0.004
	Cum_CE $_{HARD}$	-0.005***	-0.006***	-0.005***	0.015***	0.015***	0.021***
	Cum_KL $_{HARD}$	-0.001*	-0.003***	-0.002*	-0.001	-0.001	0.006
Hate Speech	Binary $_{HARD}$	-0.011***	-0.013***	-0.013***	-0.058***	-0.030**	-0.048***
	Binary $_{SOFT}$	0.002*	0.001	0.001	0.015	0.043***	0.024*
	Cum_CE $_{HARD}$	-0.003**	-0.005***	-0.005***	-0.011	0.017	-0.002
	Cum_KL $_{HARD}$	-0.010***	-0.011***	-0.011***	0.009	0.038***	0.019
Toxic	Binary $_{HARD}$	-0.007***	-0.008***	-0.007***	0.023***	0.017***	0.004
	Binary $_{SOFT}$	-0.001***	-0.001***	-0.000**	0.061***	0.056***	0.042***
	Cum_CE $_{HARD}$	-0.030***	-0.030***	-0.029***	0.174***	0.168***	0.155***
	Cum_KL $_{HARD}$	-0.022***	-0.022***	-0.021***	0.222***	0.216***	0.203***
Aggression	Binary $_{HARD}$	-0.088***	-0.087***	-0.088***	0.156***	0.147***	0.159***
	Binary $_{SOFT}$	-0.053***	-0.052***	-0.053***	0.154***	0.145***	0.156***
	Cum_CE $_{HARD}$	-0.010***	-0.009***	-0.010***	0.014***	0.004	0.016***
	Cum_KL $_{HARD}$	-0.013***	-0.012***	-0.014***	0.025***	0.015***	0.027***

Table 5. Significance tests Statistical significance tests of the performance differences between full Likert-distribution supervision and uncertainty-reduced supervision using Sentence-BERT.