Addressing the Binning Problem in Calibration Assessment through Scalar Annotations

Zhengping Jiang
Johns Hopkins University, USA
zjiang31@jhu.edu

Anqi Liu
Johns Hopkins University, USA
aliu@cs.jhu.edu

Benjamnin Van Durme
Johns Hopkins University, USA
vandurme@jhu.edu

Abstract

Computational linguistics models commonly target the prediction of discrete—*categorical*—labels. When assessing how well-calibrated these model predictions are, popular evaluation schemes require practitioners to manually determine a *binning* scheme: grouping labels into bins to approximate true label posterior. The problem is that these metrics are sensitive to binning decisions. We consider two solutions to the binning problem that apply at the stage of data annotation: collecting either distributed (redundant) labels or direct scalar value assignment.

In this paper, we show that although both approaches address the binning problem by evaluating instance-level calibration, direct scalar assignment is significantly more cost-effective. We provide theoretical analysis and empirical evidence to support our proposal for dataset creators to adopt scalar annotation protocols to enable a higher-quality assessment of model calibration.

1 Introduction

With recently released large-scale language models (LLMs) demonstrating impressive few-shot, zero-shot, and task-agnostic performance (Brown et al., 2020; Kojima et al., 2022; Ouyang et al., 2022), there is a boom of interest in deploying NLP-based systems to aid various human decision making (Chen et al., 2021; Nori et al., 2023). However, the black-box nature of LLMs gives little insight into how the predictions are made by these models (Zhao et al., 2021), risking user trust in model prediction reliability.

A common proposal to address this concern is to explore *model calibration* (Guo et al., 2017; Kull et al., 2019), which requires a model to approximately predict the true label distribution. This evaluation has been adopted by many recent language model benchmarking efforts (Desai and Durrett, 2020; Hendrycks et al., 2020; Jiang et al., 2022; OpenAI, 2023); these works often consider confidence calibration for classification and adopt Expected Calibration Error (ECE) (Guo et al., 2017) as the main empirical evaluation metric. ECE, along with variants like Adaptive Calibration Error (ACE) (Nixon et al., 2019), involve binning in their calculation, which groups hard *categorical labels* into bins to approximate label distributions. This is mainly because many popular NLP tasks are annotated predominantly with categorical labels. However, these empirical evaluations are sensitive to the choice of binning schemes (Nixon et al., 2019), and can severely underestimate calibration error (Ovadia et al., 2019; Kumar et al., 2019; Baan et al., 2022).

Instance-level calibration (Zhao et al., 2020) avoids the binning issue and matches model confidence with human annotations at an individual level, as uncertainty from human annotations is a good surrogate for true label distribution (Nie et al., 2020b; Baan et al., 2022). Following this intuition, recent work, particularly in Natural Language Inference (NLI), has crowdsourced massive number of redundant labels per instance (Pavlick and Kwiatkowski, 2019; Nie et al., 2020b), which we call *distributed labels*. These annotations cater well to the evaluation of instance-level calibration and provide valuable insight into model behavior but are often prohibitively expensive to obtain.

In this work, we propose a theoretically sound method for cost-efficient empirical calibration evaluation that can also be measured on an instance level and does not rely on binning schemes. This is done by eliciting *scalar labels* that score instances along a particular aspect and evaluate whether the predictive distribution is consistent with these scalars. An example that compares categorical, distributed, and scalar annotations can be found in Table 1. We prove that our annotations provide a lower bound for calibration error

Transactions of the Association for Computational Linguistics, vol. 12, pp. 120–136, 2024. https://doi.org/10.1162/tacl_a_00636
Action Editor: Kristina Toutanova. Submission batch: 7/2023; Revision batch: 10/2023; Published 2/2024.

P: *A man is singing into a microphone.*
H: *A man is performing on stage.*

	Cat.	**Dist.**	**Scalar**
Scheme	NEU	ENT = .2 NEU = .7 CON = .1	$p_P^H = .8$
Cost	$	$$$	$
Instance Cal.	×	√	√

Table 1: Comparing the scalar annotations against the categorical (Cat.) and distributed (Dist.) counterparts on an NLI instance. Scalar labels are as cost-efficient as categorical labels and provide instance-level calibration guarantees similar to distributed labels.

and better characterize uncertainty along the specific dimension of interests (Zhao et al., 2021). Our contributions are as follows:

- We propose widespread use of scalar labeling to capture subjective human uncertainty; it can be reliably collected, adds no overhead compared to categorical labels, yet is comparably informative as distributed labels.
- Using scalar labels, the evaluation of model calibration does not depend on the choice of the binning scheme and can be evaluated at an instance level with provable guarantees. In particular, we can use scalar labels to form a lower bound of the calibration error.
- We show on multiple NLP tasks that scalar annotations can be collected with high agreement, discriminate better than fine-grained categorical labels, and evaluate classification models consistently.

2 Motivation and Background

A probabilistic classifier $\hat{\mathsf{p}} : \mathcal{X} \to \Delta^{K-1}$ mapping input $x \in \mathcal{X}$ to a probability distribution $\big(\hat{\mathsf{p}}_1(x), \ldots, \hat{\mathsf{p}}_K(x)\big)$ of K classes is calibrated if

$$\hat{\mathsf{p}}_c(x) = \Pr(Y = y_c|x), \forall c \in [1, K],$$

where $\Pr(Y|x)$ is the true label distribution. However, achieving perfect calibration is usually infeasible, so very often some version of continuous relaxation is used to characterize the numerical error of calibration. Formally, given a critic function s.t.

$$d := \Delta^{K-1} \times \Delta^{K-1} \to \mathbb{R}^+,$$

We define the *expected calibration error* w.r.t. d to be

$$\mathrm{E}^d := \mathbb{E}_{x\sim\mathcal{X}}\Big[d\big(\hat{\mathsf{p}}(x), \Pr(Y|X = x)\big)\Big].$$

Notice that we relax the requirement on d by Vaicenavicius et al. (2019) to allow asymmetric critic function. This is to allow evaluations like confidence-calibration (Guo et al., 2017) to be included in this framework as well.

However, since $p(Y|X = x)$ is usually unknown, evaluating directly against E^d is generally infeasible. One way to approximate $\Pr(Y|X = x)$ is by binning the model's prediction with a predefined partition of the probability simplex Δ^{K-1} (Guo et al., 2017).

Alternatively, a large number of labels can be collected (Nie et al., 2020b) to approximate each instance's conditional label distribution, against which model predictive distribution for each instance is evaluated. Instance-level calibration (Zhao et al., 2020) like this does not rely on predefined binning schemes as it requires the model to predict conditional label distribution perfectly. However, the high cost of obtaining these labels makes it very hard to upscale (Clark et al., 2019). Often, many practical decisions have to be made during data collection to limit annotation cost (Collins et al., 2022), leading to suboptimal evaluation.

We motivate our work by considering a special family of critic functions d that there exists some *scoring rule* $\psi : \Delta^{K-1} \to \mathbb{R}$ against which the critic function can be written as:

$$d(\hat{\mathsf{p}}(x), \Pr(Y|X = x)) = d'(\psi(\hat{\mathsf{p}}(x)), \psi(\Pr(Y|X = x))).$$

That is, the critic function can be calculated by just comparing ''score maps'' of the two distributions. This inspires new annotation potentialities: Can we directly annotate the *transformed scores* $\psi(\hat{\mathsf{p}}(x))$ instead of the original labels Y or the

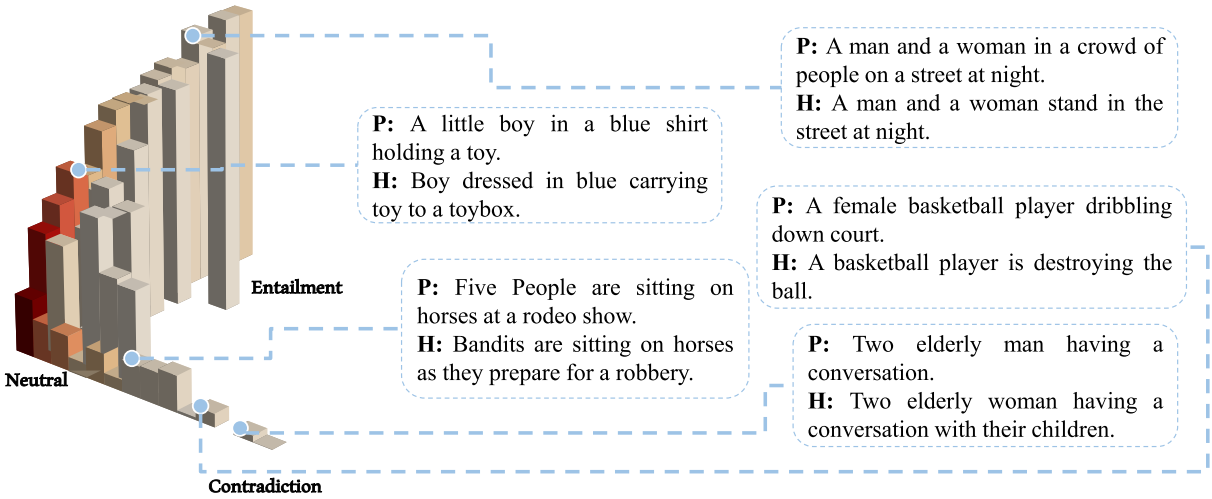

Figure 1: ChaosNLI-S (Nie et al., 2020b) label distribution visualized with barycentric coordinates concerning the ENT, NEU and CON points on the horizontal surface. The redder color on the heatmap implies a higher probability of label distribution. The height of the bars corresponds to the human uncertainty scalar labels obtained from UNLI (Chen et al., 2019). The correspondence between these two sets of labels suggests the existence of a scoring rule that maps ChaosNLI labels to a scalar with limited information lost.

label distribution $\Pr(Y|X = x)$? We answer this question positively. In the following section, we examine the effectiveness of such label transformation with a case study on directly comparing ChaosNLI (Nie et al., 2020b) and UNLI (Chen et al., 2019).

3 UNLI: A Scoring Function Example

In this section, we use UNLI for a case study on a specific scoring function ψ. For a given NLI instance, ChaosNLI (Nie et al., 2020b) elicits 100 redundant hard labels per instance to approximate true label distribution $\Pr(Y|x)$, while UNLI (Chen et al., 2019) elicits 2–3 scalar labels per instance to estimate the probability that a hypothesis is true given a premise.

Despite previously considered mismatched in distribution and unrelated (Meissner et al., 2021), we argue that these two labeling schemes are closely related. Specifically, Figure 1 shows that UNLI labels preserve an implicit ordering of the ChaosNLI label distribution. As the probability mass of label distribution for each instance gradually shifts from CON to ENT through NEU, the UNLI score also increases. Also, it is rare for an instance to have a high contradiction and entailment probability at the same time, supporting the intuition that neutral is the intermediate state between contradiction and entailment. These observations suggest that a scoring function ψ from ChaosNLI to UNLI can be found which preserves most of the information in the label distribution that we care about, such as: what's the likelihood of a hypothesis, whether some hypotheses are more likely than the others, given their premises, etc.

Suppose we want to evaluate whether our model is well-calibrated on SNLI at the instance level, namely, to test whether it provides human-aligned uncertainty for each instance. Instead of directly evaluating whether the model's predictive distribution is consistent with the ChaosNLI label distribution, we can transform model distribution with this ψ and compare it with UNLI labels to avoid the massive annotation overhead of distributed labels, as described in the following section.[1]

4 Scalar Label for Calibration

We give theoretical guarantees for using scalar annotations for calibration evaluation in this section. To begin with, we first define the class-wise

[1]By transfering the distribution with the ψ induced by Definition 2 with the same f as demonstrated in Section 6.1, we obtain high correlation as $r = 0.703$ and $\rho = 0.766$.

calibration error with which we assume calibration error is evaluated:

Definition 1 (Class-wise Calibration Error)**.** Consider a K-class classifier $\hat{\mathbf{p}} : \mathcal{X} \to \Delta^{K-1}$. The classwise calibration error L_{CCE} is defined as:

$$L_{\text{CCE}}(\hat{\mathbf{p}}) := \mathbb{E}_{\mathcal{X}} \underbrace{\frac{1}{K} \sum_{j=1}^{K} \left| \hat{\mathbf{p}}_j(x_i) - \Pr(Y = y_i^j | x_i) \right|}_{d(\hat{\mathbf{p}}(x_i), \Pr(x_i))}.$$

Remark. This calibration error is fully compatible with the expected calibration error definition in Section 2, with a critic function d marked as above.

Definition 2 below gives a simple function family that maps a discrete distribution to a scalar value. We then show that specific probing tasks on these transformed scalar values provide a useful proxy for evaluating model calibration in the form of a lower bound. Here, we only showcase our main theorem that supports direct transfer from ranking to multi-class ($K \geq 3$) classification, and leave additional results that guarantee similar lower bounds from regression or ranking to binary or multiclass classifications and their proofs to Appendix A.

Definition 2 (Expected Label Scoring Rule)**.** For a K-way classification problem with label set $\mathcal{Y} \in \{y_1, \ldots, y_K\}$, for a given function $f : \mathcal{Y} \to \mathbb{R}_{\geq 0}$ and a probabilistic classifier $\hat{\mathbf{p}}$ the expected label scoring rule is defined as:

$$\psi_f(\hat{\mathbf{p}}(x)) := \sum_{y' \in \mathcal{Y}} \hat{\mathbf{p}}_{y'}(x) f(y'). \tag{1}$$

Remark. Without losing generality, in this paper, we assume $f(y) \geq 0,\ \forall y \in \mathcal{Y}$ and that $f(y_k)$ is monotonous in k. Since we do not require a calibration loss to be calculated against this scoring rule, f does not need to be normalized as for *calibration lens* in Vaicenavicius et al. (2019). It's worth noticing that we are particularly interested in cases where f has an intuitive interpretation in empirical evaluation. For example, given a label set $\mathcal{Y}'$, $f(y) = \mathbb{I}[y \in \mathcal{Y}']$ is a probabilistic membership probe, estimating whether the model understands the concept related to a label set reminiscent of the annotation scheme proposed by Deng et al. (2012) and Collins et al. (2023).

The following result is an example demonstrating how comparing induced scalars to their corresponding human uncertainty labels can be used to evaluate model calibration when the probing task on the scalar labels is a ranking problem. Previous research has shown when the annotation protocol is properly designed, rankings among instances can be consistent even when individual scores are not (Rankin and Grube, 1980; Russell and Gray, 1994; Burton et al., 2021). We show that empirical loss functions like the pairwise ranking risk defined below can be used to form a lower bound to the original calibration error.

For a pair of instances $(x, z), (x', z') \sim \mathcal{X} \times \mathbb{R}$ identically and independently distributed (IID), the *ranking problem* is defined as correctly predicting whether $z - z' > 0$. Thus a ranking rule is a function $r : \mathcal{X} \times \mathcal{X} \to \{-1, 1\}$. The *ranking risk* w.r.t. to a ranking rule r can be defined as:

$$L(r) = \Pr\{(z - z') \cdot r(x, x') < 0\}.$$

For the original classifier and the corresponding ranking risk regret, we have the following:

Theorem 1. Under mild assumptions (Assumption 1), for a reasonably calibrated ordinal classifier $\hat{\mathbf{p}}$ s.t.:

$$L_{\text{CCE}}(\hat{\mathbf{p}}) \leq \delta,$$

for some $\delta > 0$, the relative ranking risk regret to the expected error scoring rule is bounded by:

$$L(\tilde{s}) - L(s) \leq K^2 f(y_K) \delta, \tag{2}$$

where $s = \psi_f(Pr(Y|\cdot))$ is the f-based expected label scoring rule applied to ground-truth label posterior while $\tilde{s} = \psi_f(\hat{\mathbf{p}}(Y|\cdot))$ is the f-based expected label scoring rule applied to the model predictive distribution. $L(s)$ is the ranking risk of a ranking rule r induced by s (As we will show in the appedices that under Assumption 1 s is an optimal scoring rule).

Remark. This is particularly useful for challenging annotation tasks where annotators are hard to calibrate; otherwise, we can directly compare with scalar values. Regression tasks also provide similar lower bound guarantees that can be found in the Appendix A. We also refer to a detailed discussion about how Assumption 1 ensures the existence of an intuitive ranking in the Appendix A.

Per the discussion above, to evaluate the calibration of a classifier $\hat{\mathbf{p}}$, we should compare the

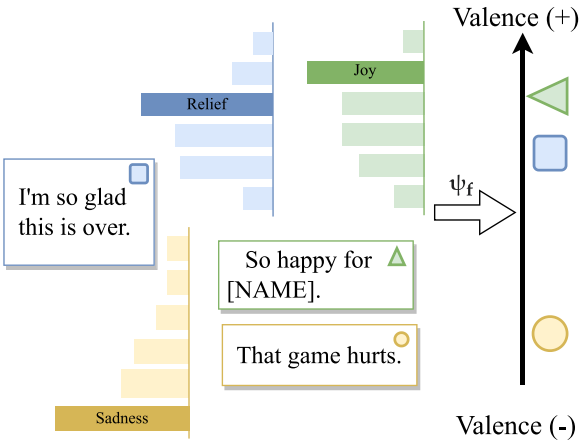

Figure 2: A hypothetical example of applying Definition 2 to an emotion classification task (e.g., GoEmotions [Demszky et al., 2020]). By attaching a single scalar valence value to each class label, we specify how the respective conditional label distribution of instances ◯, ▫, and △ induces an estimation of its valence, which can then be annotated.

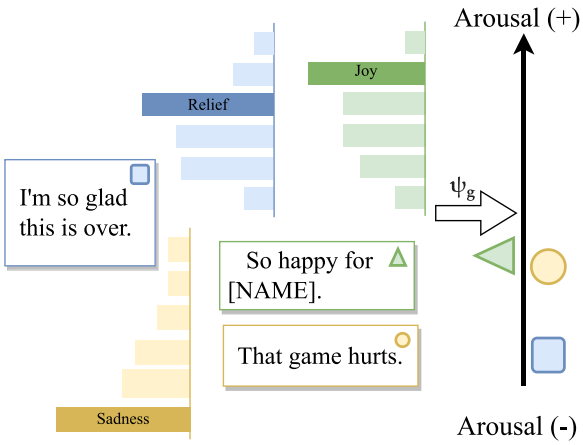

Figure 3: Applying Definition 2 with a *different* mapping function could induce a different ordering and spacing of the instances on the Arousal axis.

induced scalar values with the ground truth labels under a corresponding interpretation of f. For illustration purposes, suppose in emotion classification for GoEmotions (Demszky et al., 2020), a natural scalar interpretation of the labels is along the valence dimension. We would expect instances with a higher probability of being labeled with emotions like ‘‘Joy’’ or ‘‘Relief’’ be ranked higher on valence than instances with a higher probability of being labeled with emotions like ‘‘Sadness’’, as shown in Figure 2.

We would expect the induced valence score by a calibrated classifier $\hat{p}$ to correlate well with human annotations. This is reflected by the conformity of rankings (evaluated with ranking risk) and closeness in scoring.

It’s also worth noticing that our formulation does not require the scalar annotation task to fully recover the classification task, and it’s possible that a classification task can be characterized by multiple valid mapping functions. For example, Figure 3 shows a different ordering of the same set of instances induced by an ‘‘arousal’’ mapping function g.

5 Pseudo Distributed Label from Scalars

This section discusses ways to map scalar labels back to label distributions. This is useful when one wants to augment classification training with scalar annotations, for example when performing distillation (Hinton et al., 2015) or label smoothing (Szegedy et al., 2016). If we have parametrized distribution information of scalar annotations through common aggregation techniques (Hovy et al., 2013; Peterson et al., 2019), back-mapping is equivalent to trying to quantize a continuous distribution $p(y)$ to a discretized distribution $q(y)$. Although there already exist heuristics for allocating probability mass to categories (Pavlick and Kwiatkowski, 2019; Collins et al., 2022; Meissner et al., 2021), these mappings are generally considered suboptimal (Meissner et al., 2021). We propose two more principled ways to do label back-mapping: (1) inference with a neural network; and (2) distribution quantization with fixed support.

Neural Network We can use neural networks directly to predict the distribution parameter of the resulting categorical distribution. A small validation set of distributed labels is needed to train this back-mapping model.

Closed Form Solution We can think of label back-mapping as redistributing the probability mass of the continuous distribution to a fixed set of ranges defined by a set of cutting-off points $\{c_1, c_2, \ldots, c_K\} \in \mathbb{R}^K$, such that the discretized distribution is as close to the original distribution as possible. The following solution can be given:

Theorem 2. For a continuous distribution $P(x)$ over $\mathbb{R}$, given a set of K fixed points $D =$

$\{c_1, \ldots, c_K\} \in \mathbb{R}^n$, the resulting categorical distribution $Q(x)$ over D that minimizes Wasserstein-2 distance $W_2(p, q)$ has the form:

$$Q(d_i) = F\left(\frac{c_{i+1} + c_i}{2}\right) - F\left(\frac{c_i + c_{i-1}}{2}\right), \quad (3)$$

where $F(\cdot)$ is the PDF function of p, and we have $F(-\infty) = 0$ and $F(+\infty) = 1$.

Remark. Compared to the neural network approach, the closed-form solution does not require the ground truth distributed labels in the validation set but runs the risk of distribution mismatch with the target data.

6 Experiments

Our experiments intend to validate two critical arguments in this paper: (1) Scalar labels effectively evaluate models' uncertainty estimation; (2) It's possible to collect high-quality scalar annotations.

6.1 Evaluation with Scalar Labels

We evaluate 5 differently fine-tuned LMs against UNLI and ChaosNLI labels. Among them `bert-base-debiased-nli` (**BERT**$_b$) (Wu et al., 2022) and `roberta-large-anli` (**RoBERTa**$_a$) (Nie et al., 2020a) are from the HuggingFace model hub.[2] We intentionally choose one extensively trained, strong NLI model (**RoBERTa**$_a$), and a debiased model (**BERT**$_b$) to cover a wider range of model calibration, as per previous discussion it is generally impossible to simultaneously enforce fairness and calibration (Pleiss et al., 2017). We also fine-tune two RoBERTa (Liu et al., 2019) models, `roberta-base` (**RoBERTa**$_b$) and `roberta-large` (**RoBERTa**$_l$), on the SNLI dataset and carry out the same set of evaluations. We also evaluate a model with the `roberta-base` encoder and a randomly initialized classifier on top as a random baseline (**random**).

Comparing against Distributed Labels We first evaluate these models on ChaosNLI-S with Classwise-Calibration Error (CE) as shown in Definition 1. Notice that the calibration error evaluated in this fashion is expected to be exact and free of hyperparameters.

[2] `https://huggingface.co/models`.

Models	CE(↓)	MAE-b	RR-b	MAE(↓)	RR
random	28.7	17.3	2.75	47.9	35.2
BERT$_b$	23.7	13.0	1.20	27.7	30.1
RoBERTa$_b$	18.3	10.1	0.72	24.2	24.2
RoBERTa$_l$	16.1	8.46	0.60	23.6	23.0
RoBERTa$_a$	14.4	8.40	0.62	23.1	22.9

Table 2: Results for evaluating model calibration. Metrics against scalar labels (right side) correlate well with evaluation metrics calculated using distributed labels (left side). This empirically validates our theoretical results on the relation between the ranking risk and calibration.

We then study the evaluation capability of scalar labels by calculating the Mean Absolute Error (MAE) and Ranking Risk (RR) against UNLI-style labels. Since the original UNLI labels collected by Chen et al. (2019) only cover 614 of the 1,514 ChaosNLI-S instances, we collect UNLI annotation for all remaining ChaosNLI-S instances while ensuring a matched distribution by using the same logistic transformation as described by that prior work, as humans are especially sensitive to values near the ends of the probability spectrum (Tversky and Kahneman, 1981).

We observe that scalar-label-based metrics including RR and MAE give a consistent ranking of models compared to distributed-label-based metrics (Table 2). The model most tuned with high-quality data `roberta-large-anli` is most calibrated as indicated by RR, MAE, and CE.

Empirical Bound Investigation To better understand the behavior of scalar-label-based metrics we then investigate the lower bounds induced by their Mean Absolute Error (MAE-b) and induced by ranking risk regret (RR-b). We do this by comparing the class membership predictions of a model against the distributed labels from ChaosNLI-S using the same expected-label-scoring rule. Here we set $f(\texttt{ENT}) = 1$, $f(\texttt{NEU}) = .2$, and $f(\texttt{CON}) = 0$, which is close to the mean value of UNLI labels in each entailment label group. To calculate the ranking risk regret, for each pair of items (x, x'), we always calculate the optimal ranking risk as:

$$L(x, x') = \sum_{i=1}^{K}\sum_{j=1}^{K} \min\left(\eta_i(x)\eta_j(x'), \eta_j(x)\eta_i(x')\right),$$

which conforms to how ranking risk regret is calculated in Appendix A. However, it should be noted that there's no guarantee that the bipartite ranking problem can be reduced into a scoring problem or a consistent ranking can be determined for each pair of items.

Table 2 shows that in the case of UNLI vs ChaosNLI-S there is a considerable gap between the MAE / RR induced theoretical lower bounds (MAE-b, RR-b, respectively) versus the actual classwise calibration error. One reason for the gap is the bound for the difference in expectation terms:

$$|\sum_{i=1}^{K} f(y_i)(\eta_i(x) - \hat{\mathbf{p}}_i(x))|,$$

which without further assumption is bounded by total variation distance up to some constant. From the empirical calculation, we see that RR in theory is a weaker guarantee for CE, as the induced scores may simultaneously undergo a constant shift which will retain the same RR, but worse MAE and CE. We thus recommend using regression-based evaluation whenever possible, but for challenging tasks where precise scores are hard to annotate collecting ranking judgment alone might be more reliable (Russell and Gray, 1994; Rankin and Grube, 1980). However, as in the empirical scalar evaluation, the MAE / RR induced lower bounds rank is as consistent as CE. This further motivates the use of scalar labels for uncertainty evaluation.

Joint Training with Scalars We further tested whether joint training with UNLI will improve model calibration with the same mapping function f. We run a round-robin sampler over the two datasets. To balance the dataset size, we keep reiterating through UNLI until one epoch of SNLI finishes. For the following three settings, we evaluate with `roberta-base` and `roberta-large`: (1) *Original (SNLI)*, we evaluate a model trained on SNLI data, with the cross-entropy loss; (2) *Scalar (+reg)*, we conduct UNLI multitask training with the MAE loss on UNLI labels; and (3) *Ranking (+ran)*, we conduct UNLI multitask training with margin loss as in Li et al. (2019).

To precisely evaluate the calibration error, we also calculate instance-level class-wise calibration error on ChaosNLI data. ChaosNLI only annotates the SNLI dev set, which then requires us to use it as a test; for this experiment, we therefore use the SNLI test for development. We extend the UNLI annotation to all the ChaosNLI instances to calculate scalar-value-based metrics (MAE and RR).

Models	`roberta-base`			`roberta-large`		
	SNLI	**+reg**	**+ran**	**SNLI**	**+reg**	**+ran**
Acc(↑)	91.6	**91.8**	91.8	92.7	92.9	**93.2**
ECE-5(↓)	4.28	**3.40**	4.23	2.47	**1.41**	1.78
ECE-20(↓)	4.30	**3.47**	4.27	2.47	**1.73**	1.96
ECE-100(↓)	4.57	**3.74**	4.54	2.93	**2.08**	2.26
MAE(↓)	24.2	**23.6**	23.8	23.6	**22.7**	23.0
RR(↓)	24.2	**23.9**	23.9	23.0	**22.6**	22.7
CE(↓)	18.3	**17.4**	17.9	16.1	**15.0**	15.6

Table 3: The training result with UNLI augmentation. Training with scalar labels (**+reg** and **+ran**) improves accuracy as well as calibration. **ECE-#** indicate the bins used for the Expected Calibration Error (ECE) evaluation.

Table 3 shows that both encoders benefit from the joint training with UNLI regarding accuracy as well as model calibration. It should be noted that all UNLI training examples are already presented in the SNLI training set, so the benefit of including UNLI comes solely from scalar labels. This indicates that the jointly trained classifier, while still directly applicable to original classification tasks, can discriminate subtler differences among instances.

6.2 Studying Annotation Quality

We investigate whether humans are capable of giving consistent scalar judgments. We conduct an annotation study on the recently released WiCE (Kamoi et al., 2023) dataset. WiCE is a dataset on verifying claims decomposed from Wikipedia passages against their cited source text. A subtask of WiCE provides the annotator with a claim from Wikipedia passages intended to present an ''individual fact'' and a paired source document cited in the context. The annotator is then asked to give a 3-point scaled feedback whether the claim is either `supported`, `partially-supported`, or `not-supported` by the information provided in the source text. We replace the 3-point scale with our proposed scalar annotation scheme.

For all annotation tasks in this work, we collect scalar judgments from annotators with a slider

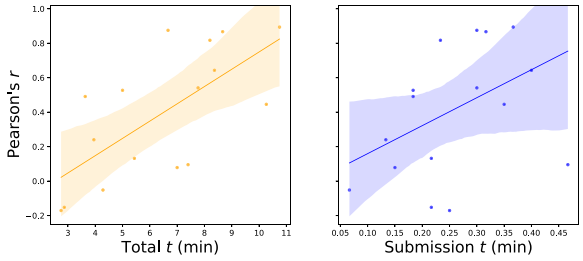

Figure 4: Relationship between an annotator's holdout Pearson's correlation coefficient to their total working time (left), and their time reviewing before submission (right). The solid line is the fitted linear regression model, of which the .95 confidence range is shaded.

bar protocol similar to the one employed by Chen et al. (2019). To get a set of good annotators, we design a qualification task with 5 manually selected claims from the abovementioned subset of WiCE, where the claims are relatively unambiguous and have varied levels of uncertainty given their respective supporting source documents. We ask workers from MTurk[3] to each do all the questions in a single session and analyze their performance to allow for qualification. To better understand worker behaviors, we log different kinds of on-page worker actions, including dragging the slider handle, checking / unchecking boxes, turning pages, or revising answers.

Figure 4 shows that as workers spend more time on the HIT and reviewing before their final submission, they get better holdout Pearson correlation against the aggregated scalar label of other workers. This supports that a responsible set of annotators can provide consistent annotations with the scalar annotation scheme, even for challenging and time-consuming tasks. We qualify workers whose holdout correlation is greater than .6.

We then annotate a subsampling of 200 subclaims from the WiCE test set with three-way redundant annotation. Figure 5 shows the scalar label distribution of this subset of WiCE, broken down by the original WiCE discrete label. The class-level ordering of the scalar label aligns well with the ''likelihood'' interpretation of the 3-way categorical labeling scheme. At the same time, scalar annotations capture more nuance of the data that better differentiate instances in the same category, especially those of the `partially-supported` class. This is expected according

[3] `https://www.mturk.com/`.

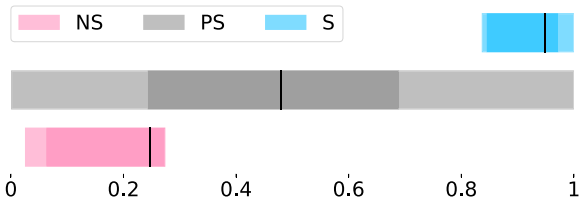

Figure 5: Strength of evidential support label distribution for each of the three discrete supporting levels on a subset sampled from the WiCE test `supported`, `partially-supported`, `not-supported`. Light / dark shade covers 100% / 50% of each category, with outliers out of 1.5 IQR dropped, and the bar in the middle of each stripe denotes the median of that category.

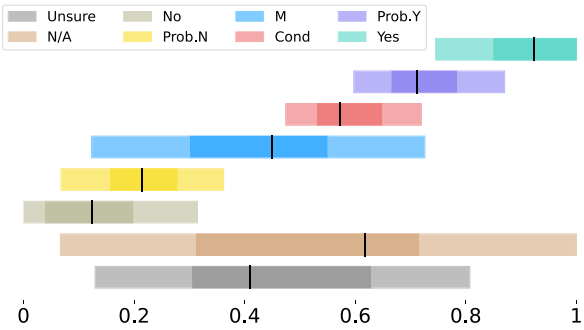

Figure 6: Scalar label annotation for Yes / No polarity on the Circa dataset breakdown by the original categorical label. In the label names,''Prob.'' means ''probably'', while ''M'' means ''in the middle''.

to the definition of the categorical label, as `partially-supported` claims can naturally be supported at any likelihood level from 0 to 1. It is worth noticing that to get good quality scalar uncertainty labels, we only need the same level of annotation redundancy compared to the original categorical labels (Kamoi et al., 2023).

6.3 Versus Fine-grained Categoricals

We also investigate whether the scalar annotation is equally effective when collected for and applied to datasets initially annotated with more fine-grained categorical labels. We first apply the scalar uncertainty annotation scheme to the Circa (Louis et al., 2020) dataset. Circa annotates a pragmatic inference problem in dialog, classifying whether an indirect answer to a question is more a ''yes'' or a ''no'' or neither. We filter out those instances with the `Other` labels, which typically correspond to irrelevant answers, and do a stratified sampling of 300 instances from all

Question	Answer	Scalar	Cat.
Do you work full-time?	Full-time, unfortunately.	1.0	Y
Are you in on Monday?	Should be!	0.9	PY
Does the neighborhood have a good reputation?	The crime rate is low.	0.8	PY
Would you have to work weekends?	I might have to.	0.7	PY
Do you like music similar to your parents?	We have some crossover.	0.6	PY
Do you like Rnb?	Hum a little for me, will you?	0.5	M
Anything I should be worried about?	About what?	0.4	M
Can you eat Mexican?	Beans make me fart.	0.3	PN
Do you know Roller balding?	That's new to me.	0.2	N
Is your favorite food Mexican?	Mexican is my second favorite.	0.1	N
Do you like country and western bands?	Country sucks.	0	N

Table 4: Equal spaced sampling from the scalar-annotated Circa subset. Notice that the scalar label makes meaningful distinctions between instances within the same class, even when the original categorical label from Circa (Louis et al., 2020) is already fine-grained.

8 remaining label classes. We collect 3-way redundant annotation with the same set of qualified annotators as in Section 6.2. To better calibrate our annotators, we dynamically show them their previous annotations for the closest-lower-scoring and closest-higher-scoring instances in the same batch.

Figure 6 shows the scalar label distribution broken down by original Circa labels. Our scalar annotation is still highly consistent with the intuitively perceived order defined by an answer's inclination towards `Yes`. At the same time, scalar uncertainty annotation captures intricate differences within each group, even when the original categorical label is already fine-grained (Table 4).

Evaluating Calibration To demonstrate the effectiveness of these scalar labels, we further fit a set of models of different sizes to the Circa dataset and evaluate them against the scalar-annotated subset. Previous research (Lewkowycz et al., 2022; Nori et al., 2023) shows that larger pretrained transformer models tend to be more calibrated, and we would like to examine whether the scalar annotation can recover the size ordering of the models in terms of calibration. To do this, we specify a very intuitive mapping function f, that maps `N`, `Prob.N`, `M`, `Prob.Y`, `Y` to equality spaced [0, .25, .5, .75, 1.], maps `N\A` and `Unsure` to .5 to indicate indecisiveness, and map `Cond` to .6 to show a slight tendency towards ''yes''.

For the evaluation experiment, we further make an 80/20 split of the not-scalar-annotated Circa subset into training and validation sets. This should be even more challenging, especially for smaller models, as there is a label distribution mismatch between this training set and our sampled test set (Dan and Roth, 2021). Besides the `bert-base-uncased` and `bert-large-uncased` model used in Section 6.1, we also tune two larger language models: GPT-Neo (Black et al., 2021) (1.3B and 2.7B). Table 5 shows that ECE-5 and ECE-100 provide different values and inconsistent rankings for the calibration level, again highlighting how the ECE result is highly hyper-parameter dependent. Instead, the scalar-label-based MAE and RR provide consistent rankings in terms of calibration evaluation.

Metrics	base	large	1.3B	2.7B
ECE-5(↓)	0.280	0.291	0.295	0.267
ECE-100(↓)	0.312	0.334	0.320	0.293
MAE(↓)	0.183	0.179	0.167	0.155
RR(↓)	0.245	0.230	0.228	0.214

Table 5: Evaluating model with scalar-label objectives as well as grouping-based calibration. Darker shades correspond to better performance on a particular metric. Metrics names remain the same as spelled out in Section 6.1.

7 Conclusions and Future Work

We show that scalar annotation elicited from individual humans can be a valuable resource for

developing calibrated NLP models. Both our theoretical and empirical results suggest that scalar annotation is an effective and scalable way to collect ground truth for human uncertainty, and we encourage future datasets to include scalar annotation if applicable. Our result provides an interesting perspective for researchers to devise new annotation tasks on traditionally categorical tasks. Future research may also look into conditions where scalar label-based uncertainty examination has better guarantees, investigate better ways to robustly collect consistent scalar annotations, and other principled ways to train or evaluate with scalar labels, particularly in areas where direct application of the method isn't immediately available, such as NLP tasks where structured predictions are involved, probably requiring some decisions specific to the task to be made regarding matters such as Events of Interest (Kuleshov and Liang, 2015).

Acknowledgment

We thank Ha Bui, Shiye Cao, Yunmo Chen, Iliana Maifeld-Carucci, Kate Sanders, Elias Stengel-Eskin, and Nathaniel Weir for their valuable feedback on the writing.

This work has been supported by the U.S. National Science Foundation under grant no. 2204926. Any opinions, findings, conclusions, or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the National Science Foundation. Anqi Liu is partially supported by the JHU-Amazon AI2AI Award and the JHU Discovery Award.

References

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.124

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow. https://doi.org/10.18653/v1/2022.bigscience-1.9

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Nichola Burton, Michael Burton, Carmen Fisher, Patricia González Peña, Gillian Rhodes, and Louise Ewing. 2021. Beyond likert ratings: Improving the robustness of developmental research measurement using best–worst scaling. *Behavior Research Methods*, pages 1–7. https://doi.org/10.3758/s13428-021-01566-w, PubMed: 33821456

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2019. Uncertain natural language inference. *arXiv preprint arXiv:1909.03042*. https://doi.org/10.18653/v1/2020.acl-main.774

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. 2008. Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, 36(2):844–874. https://doi.org/10.1214/009052607000000910

Stéphan Clémençon and Sylvain Robbiano. 2011. Minimax learning rates for bipartite ranking

and plug-in rules. In *International Conference on Machine Learning*, pages 441–448.

Stéphan Clémençon, Sylvain Robbiano, and Nicolas Vayatis. 2013. Ranking data with ordinal labels: Optimality and pairwise aggregation. *Machine Learning*, 91(1):67–104. https://doi.org/10.1007/s10994-012-5325-4

Katherine M. Collins, Matthew Barker, Mateo Espinosa Zarlenga, Naveen Raman, Umang Bhatt, Mateja Jamnik, Ilia Sucholutsky, Adrian Weller, and Krishnamurthy Dvijotham. 2023. Human uncertainty in concept-based AI systems. *arXiv preprint arXiv:2303.12872*. https://doi.org/10.1145/3600211.3604692

Katherine M. Collins, Umang Bhatt, and Adrian Weller. 2022. Eliciting and learning with soft labels from every annotator. *arXiv preprint arXiv:2207.00810*. https://doi.org/10.1609/hcomp.v10i1.21986

Soham Dan and Dan Roth. 2021. On the effects of transformer size on in-and out-of-domain calibration. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2096–2101. https://doi.org/10.18653/v1/2021.findings-emnlp.180

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.372

Jia Deng, Jonathan Krause, Alexander C. Berg, and Li Fei-Fei. 2012. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3450–3457. IEEE. https://doi.org/10.1109/CVPR.2012.6248086

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.21

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Zhengping Jiang, Anqi Liu, and Benjamin Van Durme. 2022. Calibrating zero-shot cross-lingual (un-)structured predictions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2648–2674, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.170

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. Wice: Real-world entailment for claims in wikipedia. *arXiv preprint arXiv:2303.01432*. https://doi.org/10.18653/v1/2023.emnlp-main.470

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. 2019. Generalized sliced Wasserstein distances. *Advances in Neural Information Processing Systems*, 32.

Volodymyr Kuleshov and Percy S. Liang. 2015. Calibrated structured prediction. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

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

Ananya Kumar, Percy S. Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32.

Aitor Lewkowycz, Ambrose Slone, Anders Andreassen, Daniel Freeman, Ethan S. Dyer, Gaurav Mishra, Guy Gur-Ari, Jaehoon Lee, Jascha Sohl-dickstein, Kristen Chiafullo, Liam B. Fedus, Noah Fiedel, Rosanne Liu, Vedant Misra, and Vinay Venkatesh Ramasesh. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Technical report.

Zhongyang Li, Tongfei Chen, and Benjamin Van Durme. 2019. Learning to rank for plausible plausibility. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4818–4823, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1475

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Annie Louis, Dan Roth, and Filip Radlinski. 2020. ‘‘I’d rather just go to bed’’: Understanding indirect answers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7411–7425, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.601

Johannes Mario Meissner, Napat Thumwanit, Saku Sugawara, and Akiko Aizawa. 2021. Embracing ambiguity: Shifting the training target of NLI models. *arXiv preprint arXiv:2106.03020*. https://doi.org/10.18653/v1/2021.acl-short.109

Harikrishna Narasimhan and Shivani Agarwal. 2013. On the relationship between binary classification, bipartite ranking, and binary class probability estimation. *Advances in Neural Information Processing Systems*, 26.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020a. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.441

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020b. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.734

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring calibration in deep learning. In *CVPR Workshops*, volume 2.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

OpenAI. 2023. GPT-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper

Snoek. 2019. *Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift*. Curran Associates Inc., Red Hook, NY, USA.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694. https://doi.org/10.1162/tacl_a_00293

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9617–9626. https://doi.org/10.1109/ICCV.2019.00971

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

William L. Rankin and Joel W. Grube. 1980. A comparison of ranking and rating procedures for value system measurement. *European Journal of Social Psychology*, 10(3):233–246. https://doi.org/10.1002/ejsp.2420100303

Philip A. Russell and Colin D. Gray. 1994. Ranking or rating? some data and their implications for the measurement of evaluative response. *British Journal of Psychology*, 85(1):79–92. https://doi.org/10.1111/j.2044-8295.1994.tb02509.x

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826. https://doi.org/10.1109/CVPR.2016.308

Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458. https://doi.org/10.1126/science.7455683, PubMed: 7455683

Kazuki Uematsu and Yoonkyung Lee. 2014. Statistical optimality in multipartite ranking and ordinal regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(5):1080–1094. https://doi.org/10.1109/TPAMI.2014.2360397, PubMed: 26353330

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. 2019. Evaluating model calibration in classification. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3459–3467. PMLR.

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.190

Shengjia Zhao, Michael Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. 2021. Calibrating predictions to decisions: A novel approach to multi-class calibration. *Advances in Neural Information Processing Systems*, 34:22313–22324.

Shengjia Zhao, Tengyu Ma, and Stefano Ermon. 2020. Individual calibration with randomized forecasting. In *International Conference on Machine Learning*, pages 11387–11397. PMLR.

A Appendix

A.1 Calibration Evaluation with Regression

When $K = 2$, it is straightforward that the mean absolute error (MAE) as regression loss w.r.t. the expected label scoring rule is linear to the calibration error.

Theorem 3 (REG $\to$ BINARY)**.** For a binary classification problem, the following equation holds:

$$\mathbb{E}_{x \in \mathcal{X}} |s(x) - \tilde{s}(x)| = (f(y_2) - f(y_1)) L_{\mathrm{CCE}}(\hat{\mathbf{p}}), \quad (4)$$

where

$$s(x) = \psi_f\big(\Pr(Y|x)\big), \text{ and}$$
$$\tilde{s}(x) = \psi_f\big(\hat{\mathbf{p}}(x)\big),$$

applying the expected scoring rule to true label distribution and the classifier prediction respectively.

Proof. According to Equation 1, the scoring rule s and $\tilde{s}$ can be written as:

$$s(x) = \eta_2(x)f(y_2) + \eta_1(x)f(y_1)$$
$$\tilde{s}(x) = \hat{\mathbf{p}}_2(x)f(y_2) + \hat{\mathbf{p}}_1(x)f(y_1).$$

Thus the expectation of the scoring rule MAE can be written as:

$$\mathbb{E}_{x\in\mathcal{X}}|s(x) - \tilde{s}(x)| = \mathbb{E}_{x\in\mathcal{X}}|f(y_1)(\eta_1(x) - \hat{\mathbf{p}}_1(x)) + f(y_2)(\eta_2(x) - \hat{\mathbf{p}}_2(x))|.$$

Notice that since the classification is binary we have for any $x \in \mathcal{X}$:

$$\eta_2(x) = 1 - \eta_1(x), \quad \hat{\mathbf{p}}_2(x) = 1 - \hat{\mathbf{p}}_1(x).$$

switching these equations into MAE we obtain:

$$\mathbb{E}_{x\in\mathcal{X}}|s(x) - \tilde{s}(x)| = (f(y_2) - f(y_1))\mathbb{E}_{x\in\mathcal{X}}|\eta_1(x) - \hat{\mathbf{p}}_1(x)|.$$

similarly, switching in $\eta_2(x)$ and $\hat{\mathbf{p}}_2(x)$ and average yields:

$$\begin{aligned}\mathbb{E}_{x\in\mathcal{X}}|s(x) - \tilde{s}(x)| &= \frac{1}{2}(f(y_2) - f(y_1))\Big\{\mathbb{E}_{x\in\mathcal{X}}|\eta_2(x) - \hat{\mathbf{p}}_2(x)| \\ &\qquad + \mathbb{E}_{x\in\mathcal{X}}|\eta_1(x) - \hat{\mathbf{p}}_1(x)|\Big\}, \\ &= (f(y_2) - f(y_1))L_{\text{CCE}}(\hat{\mathbf{p}}).\end{aligned}$$

□

Similar results can be derived for $K > 2$, in the form of the following theorem:

Theorem 4 (REG → MUTICLASS)**.** For a K-way classification problem, given a reasonably calibrated classifier $\hat{\mathbf{p}} : \mathcal{X} \to \Delta^{K-1}$ s.t.:

$$L_{\text{CCE}}(\hat{\mathbf{p}}) \leq \delta,$$

for some $\delta > 0$, the error of $\tilde{s}$ can be bounded by:

$$\mathbb{E}_{x\in\mathcal{X}}|s(x) - \tilde{s}(x)| \leq \Big(f(y_K)K\Big)/2 \cdot \delta. \quad (5)$$

Proof. Similar to the last proof we switch in Equation 1 to the MAE:

$$\begin{aligned}\mathbb{E}_{x\in\mathcal{X}}|s(x) - \tilde{s}(x)| &= \mathbb{E}_{x\in\mathcal{X}}|\sum_{i=1}^{K} f(y_i)(\eta_i(x) - \hat{\mathbf{p}}_i(x))|, \\ &\leq \Big(f(y_K)K\Big)/2 \cdot \delta.\end{aligned}$$

□

A.2 Calibration Evaluation with Ranking

Clémençon et al. (2008) has shown that bipartite ranking problems can be reduced to learning a scoring function $s : \mathcal{X} \to \mathbb{R}$ such that

$$r(x, x') = 1 \text{ if and only if } s(x) \geq s(x').$$

It is further demonstrated that a minimum risk scorer can be directly constructed from an optimal binary classifier, where the scoring function is just the probability of the positive class (Clémençon et al., 2008; Clémençon and Robbiano, 2011), and that binary classification regret can be lower-bounded by bipartite ranking risk (Narasimhan and Agarwal, 2013).

In contrast, a K-partite ranking problem is the case where the label Z can take a set of K values In $\{o_1, o_2, \ldots, o_k\}$. Let we denote by $\eta_i(x) = \Pr(Z = o_i|X = x)$ the probability of label o_i given x. The expected ranking risk can be further decomposed (Uematsu and Lee, 2014) into:

$$L(r) = \sum_{1\leq i,j\leq K} \eta_i(x)\eta_j(x')\mathbb{I}[r(x, x')(o_i - o_j) < 0].$$

However, the bounding result for bipartite ranking does not directly transfer to K-partite ranking. Although Uematsu and Lee (2014) construct a global optimal scoring the rule for multipartite ranking, they also comment that the optimal ranking induced may be inconsistent with a ranking induced by optimal ordinal classification labels. For example, for some very subjective ratings, like a 1 to 5 scale movie review with labels $\{1, 2, 3, 4, 5\}$, we may want to rank a pair of movies (x, x') with distributed labels $\eta(x) = [0.1, 0.4, 0, 0.2, 0.3]$ and $\eta(x') = [0.3, 0.2, 0, 0.1, 0.4]$. Notice that these two instances have identical expected scoring but different hard rating classes if labels are aggregated with majority voting. To avoid such

inconsistencies Clémençon et al. (2013) relies on the following assumption:

Assumption 1. For any $k, l \in \{1, \ldots, K-1\}$ such that $l < k$, for all $x, x' \in \mathcal{X}$ we have:

$$\Phi_{k+1,k}(x) < \Phi_{k+1,k}(x') \Rightarrow \Phi_{l+1,l}(x) < \Phi_{l+1,l}(x'),$$

where $\Phi_{k,l}(x) = \frac{\Pr[X = x|Y = y_k]}{\Pr[X = x|Y = y_l]}$ is the ratio of the density function of the class-conditional distribution of X given Y.

Remark. This assumption is equivalent to saying that the expected label scoring rule is the optimal ranker for all binary subproblems. This is a reasonable assumption to make if an obvious ordering can be identified from the label set. For example, in ChaosNLI, if an instance is more likely an ENT than a NEU, it is usually more likely a NEU than a CON as well, as we have demonstrated that probability mass only shifts gradually from CON to ENT through NEU.

Now to prove Theorem 1 we rely on the following lemma from Clémençon et al. (2013):

Lemma 1. Suppose Assumption 1 is satisfied. Let $(x, x') \in \mathcal{X} \times \mathcal{X}$. If there exists $1 \leq l < k \leq K$ such that $0 < \Phi_{k,l}(x) < \Phi_{k,l}(x')$, then for all $j \in \{1, \ldots, K\}$ we have

$$\sum_{i=j}^{K} \eta_i(x) \leq \sum_{i=j}^{K} \eta_i(x'),$$

where $\Phi_{k,l}(x)$ is defined as in Assumption 1.

For completeness, we include a proof to this lemma, which is essential for the Theorem 1.

Proof. Since $\Phi_{l,k}(x) < \Phi_{l,k}(x')$, we have:

$$\eta_k(x) - \eta_k(x) \sum_{i \neq l} \eta_i(x) < \eta_k(x') - \eta_k(x') \sum_{i \neq l} \eta_i(x), \tag{6}$$

thus

$$\begin{aligned} &\eta_k(x) - \eta_k(x') \\ &< \sum_{i \neq l} \Big\{ \eta_k(x)\eta_i(x') - \eta_k(x')\eta_i(x) \Big\}, \\ &< \sum_{i<l} \Big\{ \eta_k(x)\eta_i(x') - \eta_k(x')\eta_i(x) \Big\} \\ &\quad + \sum_{i>l} \Big\{ \eta_k(x)\eta_i(x') - \eta_k(x)\eta_i(x') \Big\} \end{aligned}$$

Now suppose we have

$$\Phi_{K,K-1}(x) > \Phi_{K,K-1}(x'),$$

then according to assumption 1 we have:

$$\Phi_{i+1,i}(x) > \Phi_{i+1,i}(x'), \quad \forall i \text{ s.t. } 1 \leq i \leq K-1.$$

Thus

$$\begin{aligned} &\Phi_{l+1,l}(x)\Phi_{l+2,l+1}(x) \ldots \Phi_{k,k-1}(x) \\ &\qquad > \Phi_{l+1,l}(x')\Phi_{l+2,l+1}(x') \ldots \Phi_{k,k-1}(x') \\ &\Rightarrow \Phi_{k,l}(x) > \Phi_{k,l}(x'), \end{aligned}$$

which is contradictory to the assumption made by the lemma. So we have (by repeating the process for $K-2, K-3, \ldots l$):

$$\Phi_{i+1,i}(x) \leq \Phi_{i+1,i}(x'), \quad \forall i \text{ s.t. } l \leq i \leq K-1.$$

And since the inequality for $\Phi_{k,l}$ is strict so there exists some $l \leq v \leq K-1$ such that:

$$\Phi_{v+1,v}(x) < \Phi_{v+1,v}(x'),$$

which could further prove that:

$$\begin{aligned} &\Phi_{j-1,m}(x) \leq \Phi_{j-1,m}(x'), \\ &\qquad \forall j, m \text{ s.t. } 2 \leq j < m \leq K-1. \end{aligned}$$

Similar to Equation 6 we have for $1 \leq j < m \leq K$:

$$\begin{aligned} &\eta_m(x) - \eta_m(x') \\ &\quad \leq \sum_{i \neq j-1} \Big\{ \eta_m(x)\eta_i(x') - \eta_m(x')\eta_i(x) \Big\}, \\ &\quad \leq \sum_{i<j-1} \Big\{ \eta_m(x)\eta_i(x') - \eta_m(x')\eta_i(x) \Big\} \\ &\qquad + \sum_{i>j-1} \Big\{ \eta_m(x)\eta_i(x') - \eta_m(x')\eta_i(x) \Big\}, \\ &\quad \leq \sum_{i>j-1} \Big\{ \eta_m(x)\eta_i(x') - \eta_m(x')\eta_i(x) \Big\}. \end{aligned}$$

The last inequality is because of the fact that $\Phi_{m,i}(x) \leq \Phi_{m,i}(x'), \forall\ i$ s.t. $i < m$. Then summing up all these inequalities from j to K we have:

$$\sum_{m=j}^{K}\eta_m(x) \leq \sum_{m=j}^{K}\eta_m(x') + \sum_{m=j}^{K}\sum_{i=j}^{K}\Big\{\eta_m(x)\eta_i(x') - \eta_m(x')\eta_i(x)\Big\}.$$

And this proves the lemma because the double summation on the right-hand side equals 0. □

This leads to the following corollary:

Lemma 2. When Assumption 1 holds, the expected scoring rule by Equation 1 is an optimal scoring rule.

Proof. For scoring function f, where $f(y_0) = 0$ This can be directly proved by summing up

$$(f(y_j) - f(y_{j-1}))\sum_{i=j}^{K}\eta_i(x) \leq (f(y_j) - f(y_{j-1}))\sum_{i=j}^{K}\eta_i(x'),$$

where $j = 1, \ldots, K$. □

Then we are able to prove Theorem 1:

Proof. Given lemma 2, we are able to write the ranking risk using Equation 1 as scoring rule s as follows:

$$\begin{aligned} L(s) &= \mathbb{E}_{(x,x')\sim\mathcal{X}\times\mathcal{X}}\sum_{i=1}^{K}\sum_{j=1}^{K}\eta_i(x)\eta_j(x')\cdot \\ &\qquad \mathbb{I}[(s(x) - s(x'))(f(y_i) - f(y_j) < 0], \\ &= \frac{1}{2}\mathbb{E}_{(x,x')\sim\mathcal{X}\times\mathcal{X}}\sum_{i=1}^{K}\sum_{j=1}^{K} \\ &\qquad \min\Big\{\eta_i(x)\eta_j(x'), \eta_i(x')\eta_j(x)\Big\}. \end{aligned}$$

And for any other scoring function that is not necessarily optimal, including $\tilde{s}$, the ranking risk can be written as:

$$\begin{aligned} L(\tilde{s}) = \frac{1}{2}\mathbb{E}_{(x,x')\sim\mathcal{X}\times\mathcal{X}}\sum_{i=1}^{K}\sum_{j=1}^{K}\Big\{ & \\ \mathbb{I}[\tilde{s}(x) - \tilde{s}(x') > 0]\eta_j(x')\eta_i(x) + & \\ \mathbb{I}[\tilde{s}(x) - \tilde{s}(x') < 0]\eta_i(x')\eta_j(x)\Big\}. & \end{aligned}$$

We can calculate the regret as:

$$\begin{aligned} &\mathcal{L}_{\text{ovo}}(\tilde{s}) - \mathcal{L}_{\text{ovo}}(s) \\ &\leq \frac{1}{2}\mathbb{E}\sum_{1\leq i,j\leq K}|\eta_i(x)\eta_j(x') - \eta_i(x')\eta_j(x)|, \\ &\leq \mathbb{E}\sum_{1\leq i,j\leq K}\Big\{|\eta_i(x) - \eta_i(x')| + \\ &\qquad\qquad |\eta_j(x) - \eta_j(x')|\Big\}. \\ &\leq \mathbb{E}\sum_{1\leq i,j\leq K}\Big\{|s(x) - \tilde{s}(x)| + |s(x') - \tilde{s}(x')|\Big\}, \\ &= 2\mathbb{E}_{x\sim\mathcal{X}}\sum_{i=1}^{K}\sum_{j=1}^{K}|s(x) - \tilde{s}(x)|, \\ &\leq f(y_K)K^2\delta. \end{aligned}$$

As the second inequality is due to lemma 2 and the last inequality is the same as in the previous theorem. □

Theorem 2 is the direct corollary of some well-known property of the Wasserstein distance (e.g., see Kolouri et al., 2019).

Proof. Denote by μ the target discrete distribution will K number. We want to minimize $W_2^2(Q, P)$:

$$\min_{\mu} W_2^2(Q, P) = \min_{\mu}\int_0^1 |F^{-1}(u) - G^{-1}(u)|^2 du,$$

where $F(x)$ is the CDF of P and $G(x)$ is the CDF of Q as defined in Theorem 2. Let $u_1 = 0 < u_2 < \cdots < u_{K+1} = 1$ be the probability quantiles corresponds to $d_1, \ldots, d_K$ such that:

$$u_j = \sum_{i=1}^{j-1} q(x_i).$$

Therefore by changing variables we have:

$$\min_{\mu} W_2^2(Q, P) = \min_{u}\sum_{j=1}^{n}\int_{u_j}^{u_{j+1}} |x_i - F^{-1}(u)|^2 du,$$

since d is already fixed. Considering the partial derivative:

$$\frac{\partial W_2^2(Q,P)}{\partial u_i} = 0$$
$$\Rightarrow \quad F^{-1}(u_j) = \frac{d_{j-1} + d_j}{2}.$$

Thus the probability mass distributed to each support point d_i is just:

$$q(d_i) = F\left(\frac{d_{i+1} + d_i}{2}\right) - F\left(\frac{d_{i-1} + d_i}{2}\right).$$

□