UNIVERSITY OF CAMBRIDGE

Stop Measuring Calibration When Humans Disagree

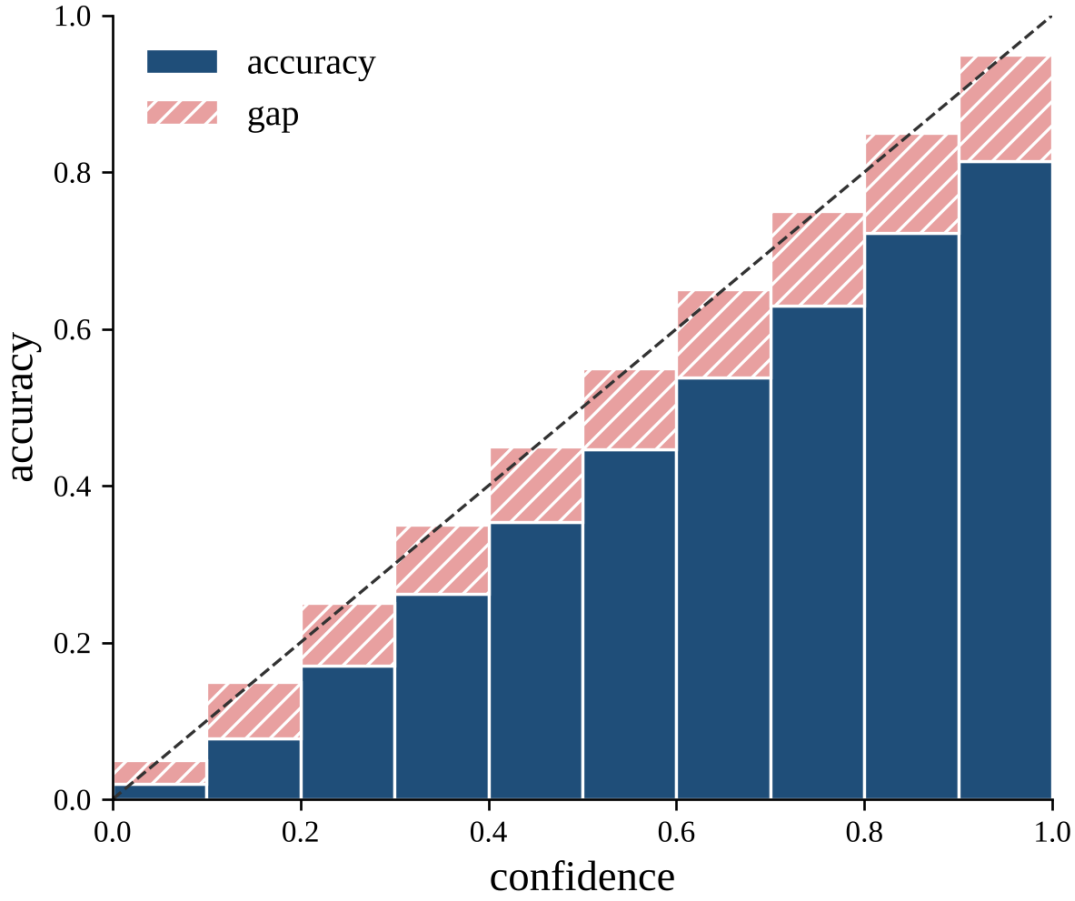
Joris Baan, Wilker Aziz, Barbara Plank, Raquel Fernández (EMNLP 2022)

Tamiru Alemnew

L101: Machine Learning for Language Processing

Motivation

- Calibration: confidence should match accuracy
- Used to decide when to trust a prediction
- "Correct" means matching one gold label



Reliability diagram (illustration)

The problem: people disagree

Premise	Hypothesis	100 votes (E / N / C)
A man running a marathon talks to his friend.	There is a man running.	100 / 0 / 0
An elderly woman crafts a design on a loom.	The woman is sewing.	**35 / 31 / 34**

ChaosNLI (Baan et al., Tables 3 and 4)

- Standard calibration metrics break when people disagree
- A new definition: calibration to human uncertainty
- Case study: ChaosNLI and temperature scaling

Background

Humans

$$\Pr(Y = c \mid X = x) = \pi_c(x)$$

$$\bar{\pi}_c(x) = \frac{\#\text{votes for } c \text{ on } x}{\#\text{votes on } x}$$

Background

Humans

$$\Pr(Y = c \mid X = x) = \pi_c(x)$$

$$\bar{\pi}_c(x) = \frac{\#\text{votes for } c \text{ on } x}{\#\text{votes on } x}$$

Model

$$f(x) = \big(f_1(x), \ldots, f_C(x)\big)$$

$$\hat{y} = \arg\max_c f_c(x) \qquad y^\star = \arg\max_c \bar{\pi}_c(x)$$

Calibration

Multi-class calibration

$$\Pr\left(Y^{\star} = c \mid f(X) = q\right) = q_c \quad \forall c$$

Calibration

Multi-class calibration

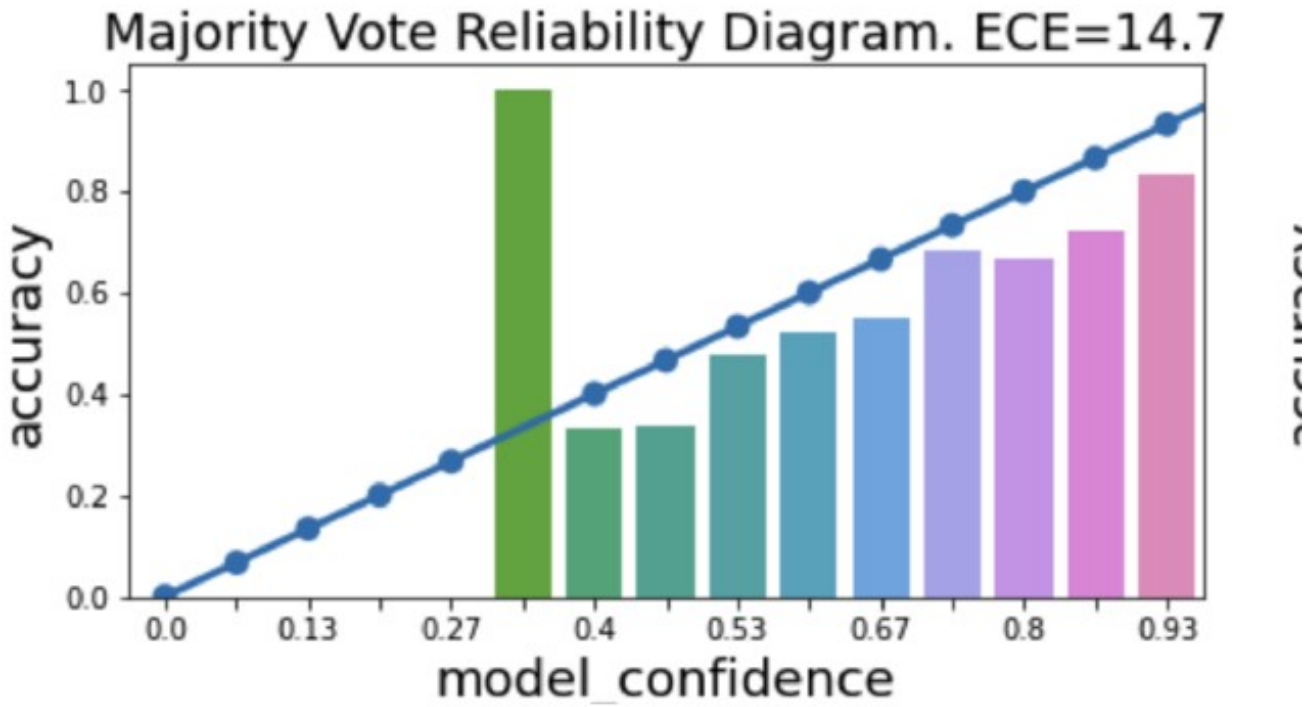
$$\Pr\left(Y^\star = c \mid f(X) = q\right) = q_c \quad \forall c$$

Confidence calibration (common in NLP)

$$\Pr\left(Y^\star = \hat{Y} \mid \max f(X) = p\right) = p$$

Expected Calibration Error

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



RoBERTa on ChaosNLI-SNLI, one seed: ECE 14.7% = 0.147 (Baan et al., Fig. 1c)

ECE of a perfect model

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

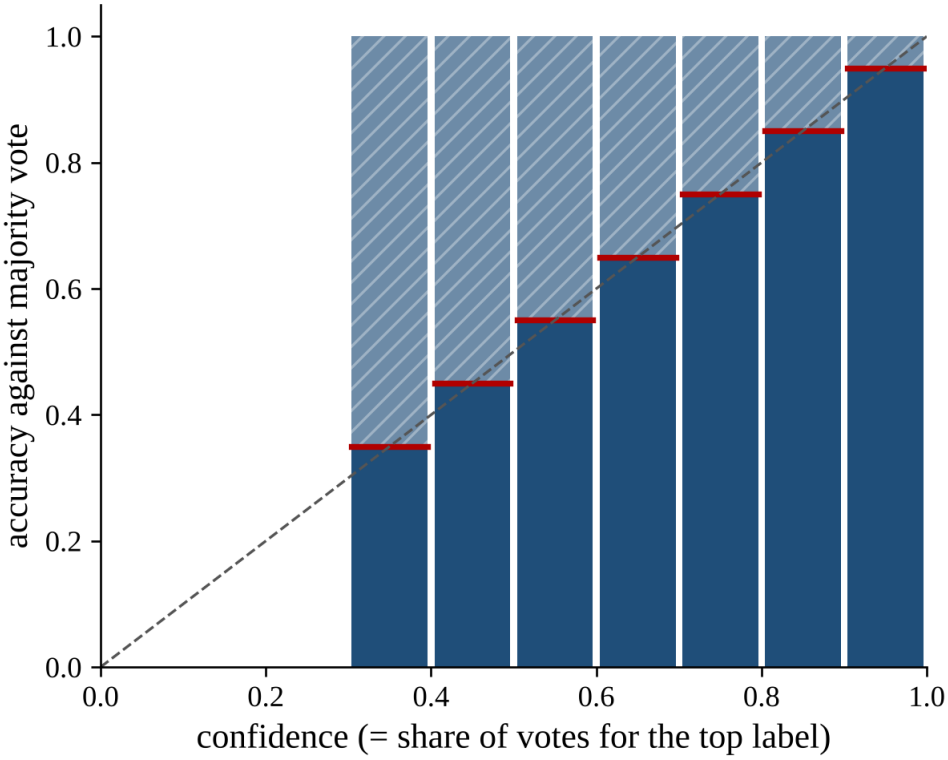


Illustration: oracle with f(x) = π̄(x)

- f(x) = π̄(x)
- accuracy = 1 in every bin
- confidence < 1
- **ECE:** oracle 0.25, RoBERTa 0.14

Calibration to human uncertainty

$$\Pr(Y = c \mid X = x) = f_c(x) \quad \forall c$$

Calibration to human uncertainty

$$\Pr(Y = c \mid X = x) = f_c(x) \quad \forall c$$

Uncertainty

$$\mathrm{EntCE}(x) = H\big(f(x)\big) - H\big(\bar{\pi}(x)\big)$$

Calibration to human uncertainty

$$\Pr(Y = c \mid X = x) = f_c(x) \quad \forall c$$

Uncertainty

$$\mathrm{EntCE}(x) = H\big(f(x)\big) - H\big(\bar{\pi}(x)\big)$$

Ranking

$$\mathrm{RankCS} = \frac{1}{N}\sum_n \big[\operatorname{argsort} f(x_n) = \operatorname{argsort} \bar{\pi}(x_n)\big]$$

Calibration to human uncertainty

$$\Pr(Y = c \mid X = x) = f_c(x) \quad \forall c$$

Uncertainty

$$\mathrm{EntCE}(x) = H\big(f(x)\big) - H\big(\bar{\pi}(x)\big)$$

Ranking

$$\mathrm{RankCS} = \frac{1}{N}\sum_n \big[\operatorname{argsort} f(x_n) = \operatorname{argsort} \bar{\pi}(x_n)\big]$$

Full distribution

$$\mathrm{DistCE}(x) = \mathrm{TVD}\big(f(x), \bar{\pi}(x)\big) = \tfrac{1}{2}\sum_c \big|f_c(x) - \bar{\pi}_c(x)\big|$$

Case study

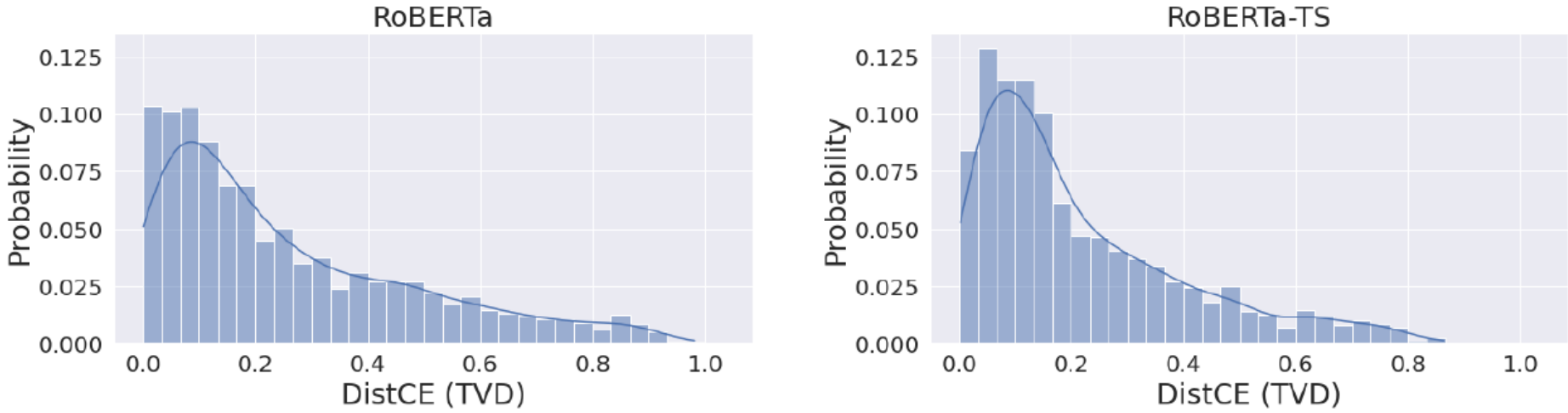
- **Data:** ChaosNLI-SNLI, 100 votes per item
- **Model:** RoBERTa fine-tuned on SNLI, 3 seeds
- **TS:** logits ÷ T, with T = 2.0 tuned on the test set for lowest ECE
- **Oracle:** $f(x) = \pi(x)$

Results

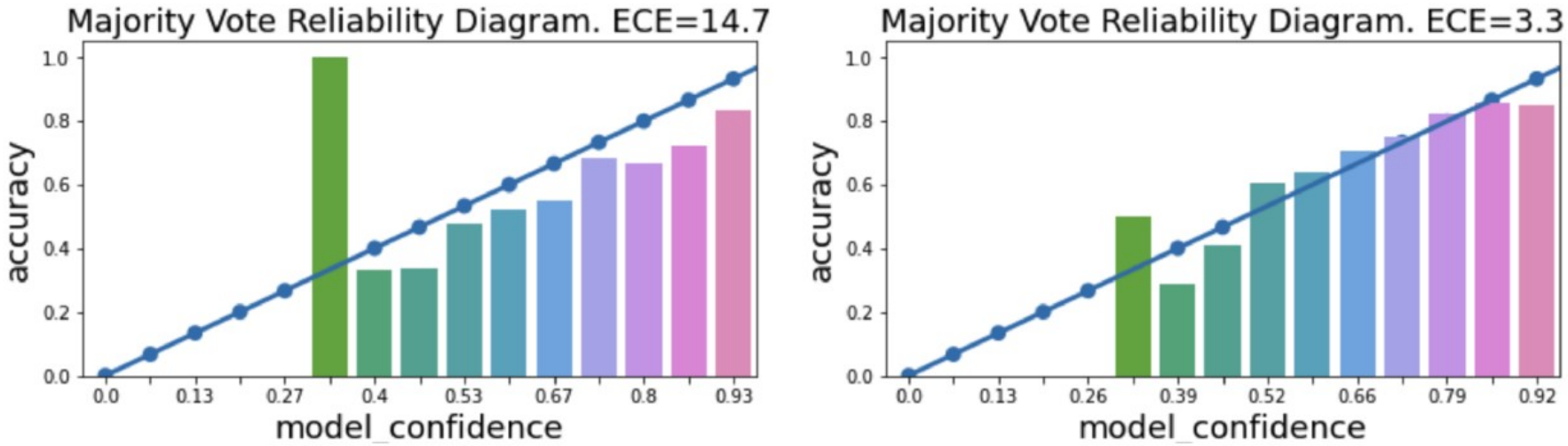
	RoBERTa	RoBERTa-TS	Oracle
Acc ↑	0.74 ±0.01	0.74±0.01	1.00
ECE ↓	0.14 ±0.01	0.03±0.01	0.25
RankCS ↑	0.62 ±0.01	0.62±0.01	1.00
$\mathbb{E}[\lvert\text{EntCE}\rvert]$ ↓	0.30 ±0.02	0.21±0.02	0.00
$\mathbb{E}[\text{DistCE}]$ ↓	0.26 ±0.00	0.22±0.00	0.00

ChaosNLI-SNLI, mean of 3 seeds (Baan et al., Table 1)

What temperature scaling does

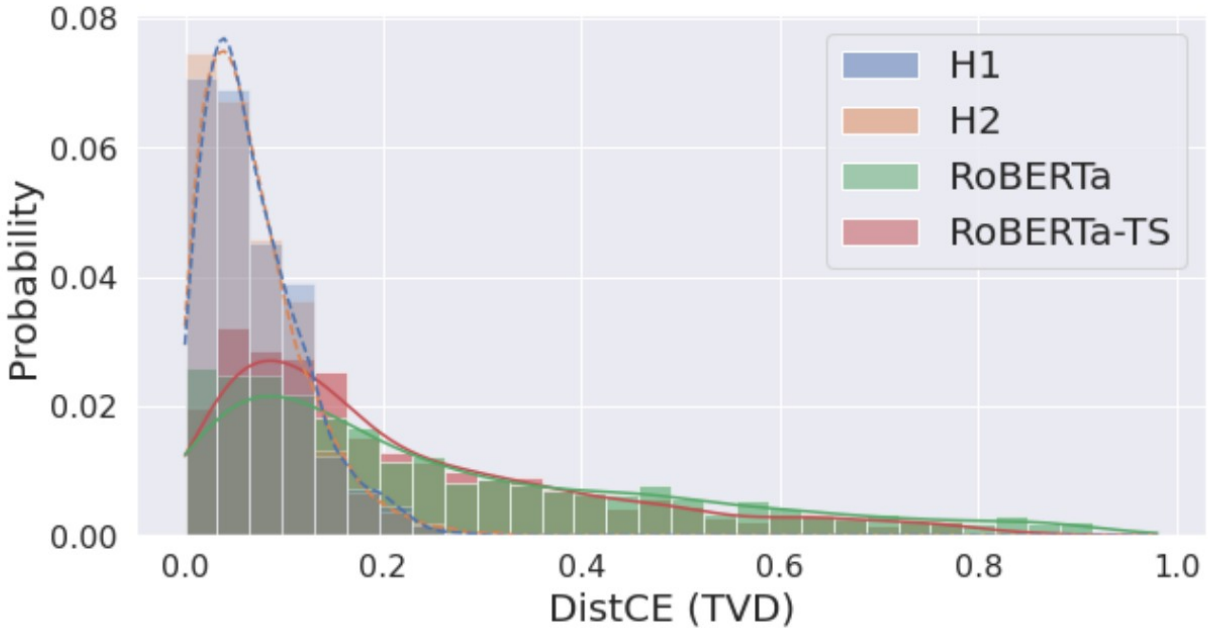


Per-item DistCE: before / after TS



Reliability diagrams: before / after TS (Baan et al., Fig. 1)

How good is good enough?



Baan et al., Fig. 2

DistCE error distribution	KL(H1, ·)	TVD(H1, ·)
H2	0.004	0.022
RoBERTa	0.688	0.500
RoBERTa-TS	0.611	0.454

Baan et al., Table 2

H1, H2: 20 sampled human votes each

Strengths

- A perfect model fails ECE: no experiment needed
- The problem is the gold label, not the bins
- Per-item errors show what an average hides
- Human baseline gives the numbers a scale

Weaknesses

- One dataset, one model, one method
- ChaosNLI was chosen for disagreement
- Temperature tuned for ECE, judged on DistCE
- ECE still answers a useful question

What later work says

- **Baan et al. (2024):** model confidence and human label variation are different goals
- **Weber-Genzel et al. (2024):** some disagreement is annotation error
- **Gruber et al. (2024):** raw vote shares may not be the right target
- **Jiang et al. (2024):** many labels per item are costly; scalar labels as a cheaper option

Summary

- Majority-vote ECE misjudges models when people disagree
- Measure calibration against the human distribution
- Temperature scaling fixes ECE, not the gap to humans
- Narrow evidence, strong argument

References

- Baan, Aziz, Plank & Fernández (2022). Stop Measuring Calibration When Humans Disagree. EMNLP.
- Nie, Zhou & Bansal (2020). What Can We Learn from Collective Human Opinions on Natural Language Inference Data? EMNLP.
- Guo, Pleiss, Sun & Weinberger (2017). On Calibration of Modern Neural Networks. ICML.
- Wang et al. (2022). Capture Human Disagreement Distributions by Calibrated Networks for Natural Language Inference. Findings of ACL.
- Nixon et al. (2019). Measuring Calibration in Deep Learning. CVPR Workshops.
- Gupta et al. (2021). Calibration of Neural Networks using Splines. ICLR.
- Lee et al. (2023). Can Large Language Models Capture Dissenting Human Voices? EMNLP.
- Baan et al. (2024). Interpreting Predictive Probabilities: Model Confidence or Human Label Variation? EACL.
- Weber-Genzel et al. (2024). VariErr NLI: Separating Annotation Error from Human Label Variation. ACL.
- Gruber et al. (2024). More Labels or Cases? Assessing Label Variation in Natural Language Inference. UnImplicit Workshop.
- Jiang, Liu & Van Durme (2024). Addressing the Binning Problem in Calibration Assessment through Scalar Annotations. TACL.