Calibrating Expressions of Certainty

Peiqi Wang*, Barbara D. Lam, Yingcheng Liu, Ameneh Asgari-Targhi, Rameswar Panda, William M. Wells, Tina Kapur, Polina Golland

* https://people.csail.mit.edu/wpq/ wpq@mit.edu

1. Conveying accurate confidence is critical

Chest X-ray Report

Impression:

- **Possible** pneumonia. Order Followup CT
- Mild bronchial wall thickening, **consistent with** bronchitis.
- No acute abnormalities in the left lung or mediastinum.

Chest X-ray Report

Impression:

- **Likely** pneumonia. Prescribing Antibiotics & Order CT
- Mild bronchial wall thickening, **consistent with** bronchitis.
- No acute abnormalities in the left lung or mediastinum.

2. Background on Calibration and Miscalibration

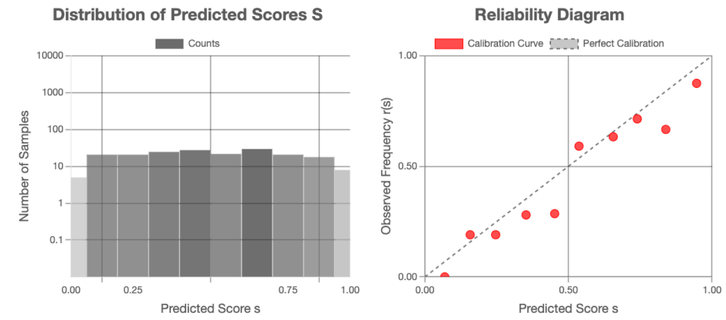
- Classifier $g : \mathcal{X} \to [0, 1]$ is **calibrated** if

$$\text{“Observed Frequency”} \quad \mathbb{E}[Y \mid S] = S \quad \text{“Predicted Confidence Score”},$$

where $\mathcal{X}$ is input space, $\mathcal{Y} = \{0, 1\}$ is label space, and $S = g(X)$ represents the confidence score.

- **Expected Calibration Error** measures the degree to which the calibration definition is violated

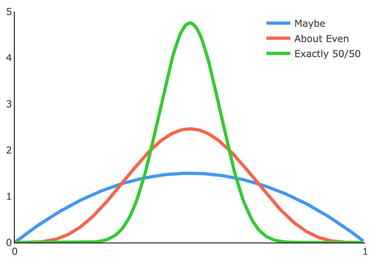
$$\text{ECE} = \mathbb{E}\left[\,|\mathbb{E}[Y \mid S] - S|\,\right]$$



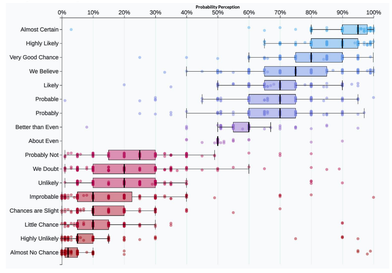
3. Linguistic Expressions of Certainty, not Scores!

assuming certainty phrase ↔ probability (e.g., ”Maybe” ↔ 0.5, “Likely” ↔ 0.8, …) is insufficient

a) Nuanced semantics

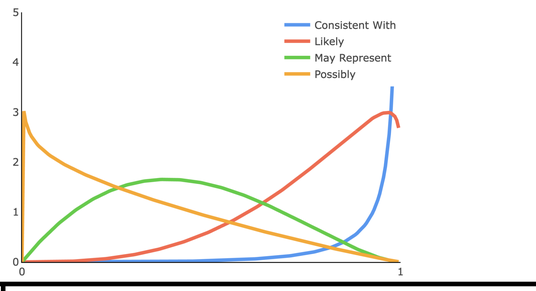


b) Population-level variation



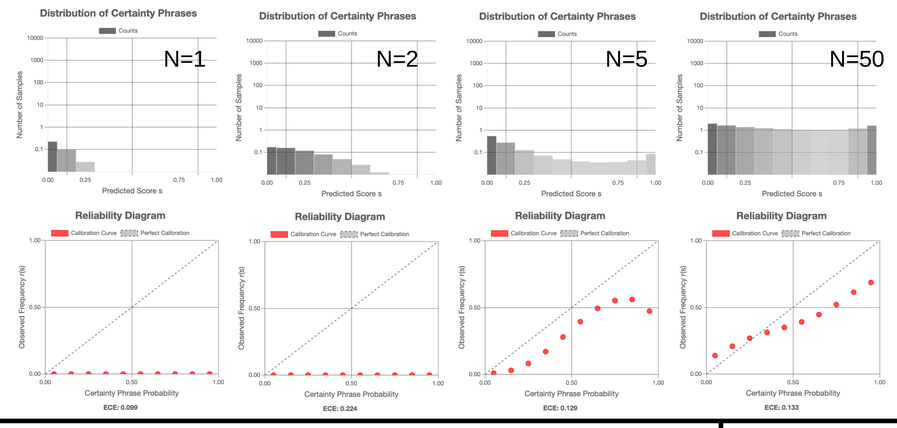
4. Confidence as a Distribution

- assume certainty phrase “Maybe” ↔ Beta(2,2), “Likely” ↔ Beta(4,1), …
- distributions can be derived from (1) survey of human perception of these phrases, (2) prompting LMs (e.g., gpt-4)



5. Measuring Linguistic Calibration

- Definition for calibration & miscalibration remain valid!
- Generalization on estimation procedure: “splitting” each sample to bins I_m weighted by $P(S \in I_m \mid X = x_n)$



6. Measuring Radiologists Calibration

2,662 paired (X-ray, CT) report where X-ray are human predictions and CT are ground-truth

Chest X-ray Report

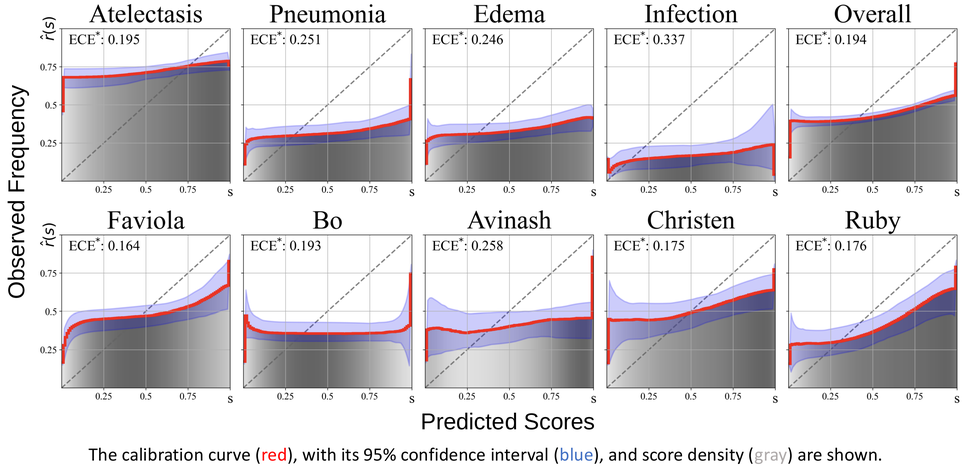
Impression:

- Right lower lobe opacification, which **may represent** pneumonia.
- No acute abnormalities in the left lung or mediastinum.
- **Possible** pulmonary embolism.

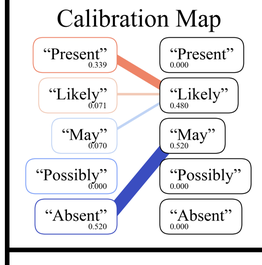
CT Report

Impression:

- Right lower lobe opacification **consistent with** pneumonia.
- No acute abnormalities in the left lung or mediastinum.
- pulmonary embolism is **absent**.



The calibration curve (red), with its 95% confidence interval (blue), and score density (gray) are shown.



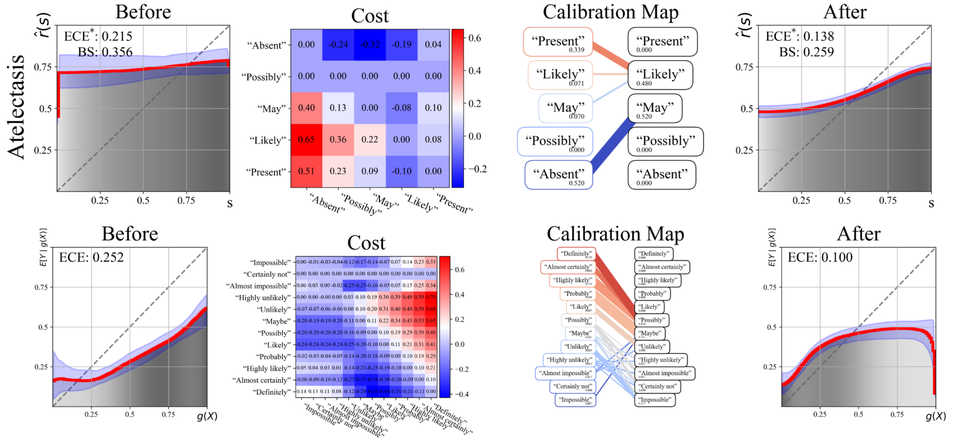
7. Post-hoc Calibration

by adjusting the usage of certainty phrases

- Traditional Calibration Map: $t : [0, 1] \to [0, 1]$ (e.g., platt scaling, histogram binning).
- Our work: $t : [K] \to [K]$ mapping K source distributions to K target distributions.

Paper Link

8. Calibrating Radiologists (top) and Language Models (bottom)



- Source distribution: $a = (a_1, \ldots, a_K) \in \triangle^{K-1}$, where a_k is the proportion of times the k-th certainty phrase is used.
- Cost matrix $C \in \mathbb{R}^{K \times K}$: $C_{kl} = \frac{1}{a_k}\left(\widehat{\text{ECE}}(u_k \to v_l) - \widehat{\text{ECE}}\right)$.
- Calibration as discrete OT: $\min_T \langle C, T \rangle$ s.t. $T\mathbf{1} = a$ (preserve source distribution), $T^\top \mathbf{1} \approx a$ (target close to source).

This work was supported by the Takeda Fellowship, the MIT–IBM Watson AI Lab, the MIT CSAIL-Wistron Program, and the MIT Jameel Clinic.