Part IV
Model calibration
Conformal Prediction

Adrian Galdran
RyC/ikerbasque Research Fellow
Tecnalia - Derio, Spain
adrian.galdran@tecnalia.com

MICCAI 2025 Daejeon REPUBLIC OF KOREA

tecnalia
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

Model Calibration - Contents

1. Motivation - Why Calibration?
2. Defining & Measuring Calibration
3. Improving Calibration
4. Hands-On

1. Motivation - Why Calibration?

Why Language Models Hallucinate

Adam Tauman Kalai*
OpenAI

Ofir Nachum
OpenAI

Santosh S. Vempala†
Georgia Tech

Edwin Zhang
OpenAI

September 4, 2025

1. Motivation - Why Calibration?

Why Language Models Hallucinate

Adam Tauman Kalai* — OpenAI
Ofir Nachum — OpenAI
Santosh S. Vempala† — Georgia Tech
Edwin Zhang — OpenAI

September 4, 2025

*Most evaluations measure performance in a way that encourages guessing rather than **honesty about uncertainty**. […] There is a straightforward fix. Penalize confident errors more than you penalize uncertainty, [use] evaluations that account for uncertainty and **calibration**.*

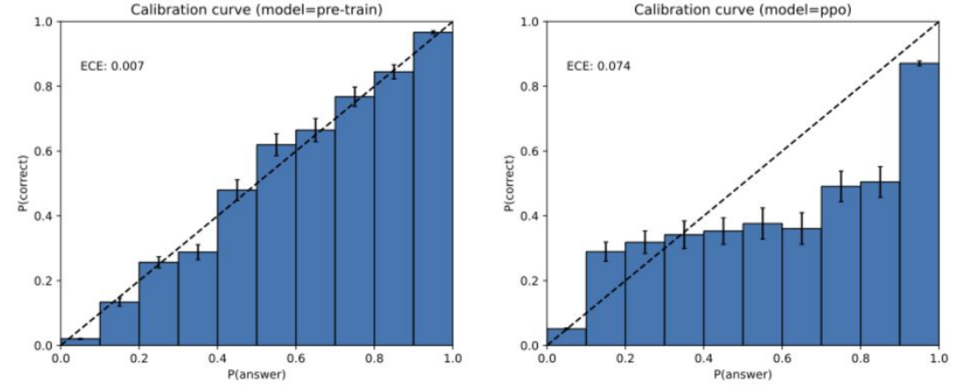
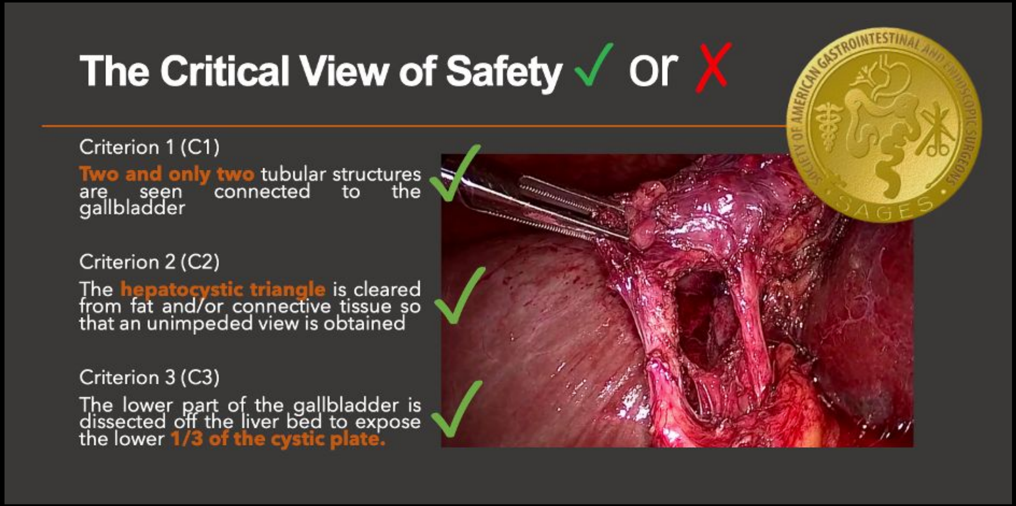
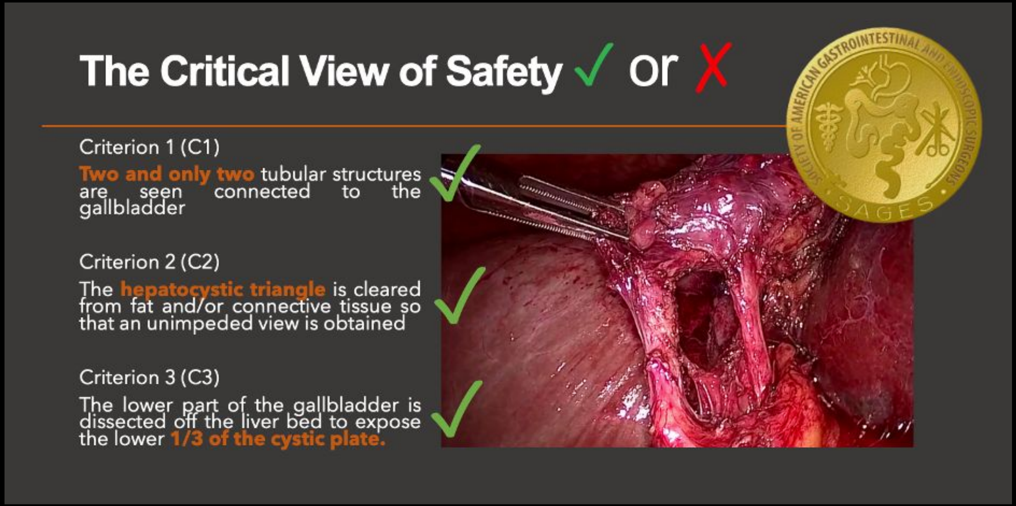


Figure 2: GPT-4 calibration histograms before (left) and after (right) reinforcement learning (OpenAI, 2023a, Figure 8, reprinted with permission). These plots are for multiple-choice queries where the plausible responses are simply A, B, C, or D. The pretrained model is well calibrated.

1. Motivation - Why Calibration?



1. Motivation - Why Calibration?



The 2025 SAGES CVS Lighthouse Challenge revisits the CVS classification task. In terms of uncertainty calibration and robustness, we will focus on how robust are the algorithms when deployed in different conditions (e.g. sites, countries, etc..), as well as how cognizant when their answers might be wrong, to enable using of these algorithms in a safe manner.

2. Defining & Measuring Calibration

p	y
⅕	0
⅕	0
⅕	0
⅕	0
⅕	1

p	y
⅓	0
⅓	0
⅓	1
½	0
½	1

p	y
¾	0
¾	1
¾	1
¾	1
1	1

2. Defining & Measuring Calibration

p	y
1/5	0
1/5	0
1/5	0
1/5	0
1/5	1

p	y
1/3	0
1/3	0
1/3	1
1/2	0
1/2	1

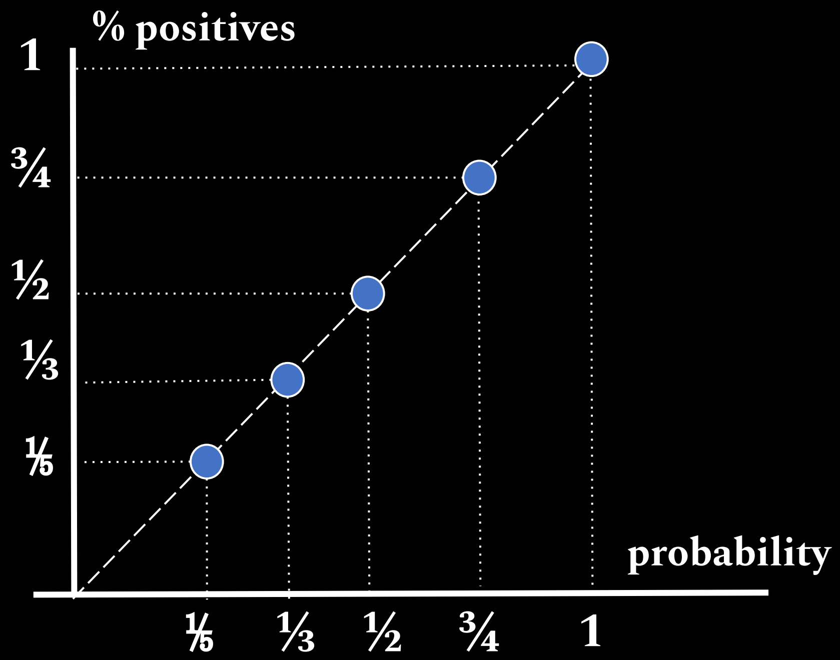
p	y
3/4	0
3/4	1
3/4	1
3/4	1
1	1

2. Defining & Measuring Calibration

p	y
⅕	0
⅕	0
⅕	0
⅕	0
⅕	1

p	y
⅓	0
⅓	0
⅓	1
½	0
½	1

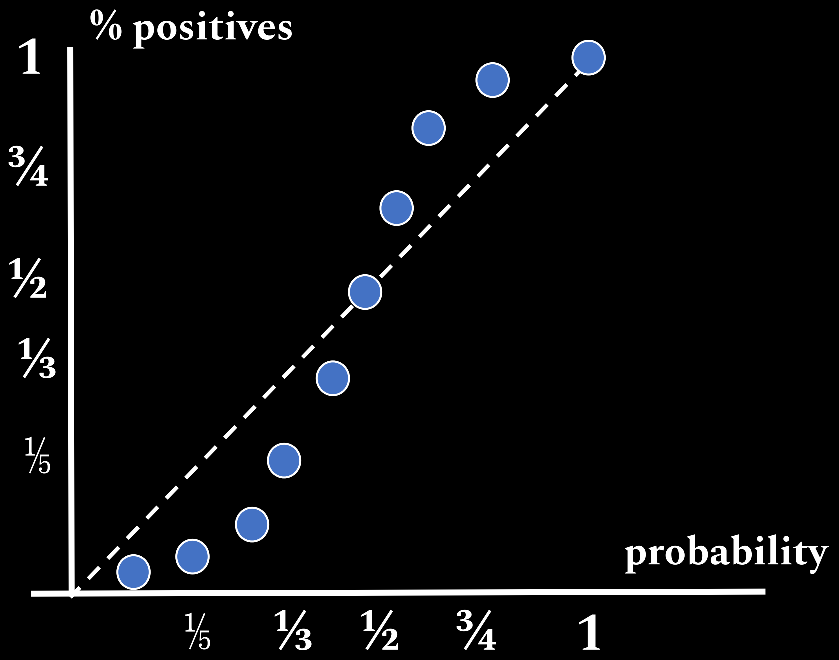
p	y
¾	0
¾	1
¾	1
¾	1
1	1



MICCAI2025
Daejeon
REPUBLIC OF KOREA

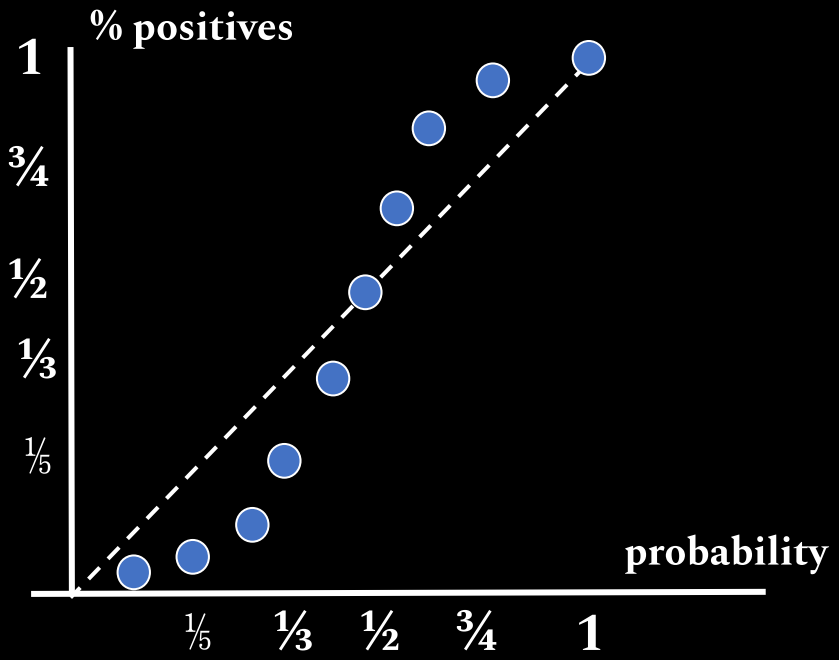
2. Defining & Measuring Calibration

QUESTION:
Are these predictions <u>under</u>-confident or <u>over</u>-confident?



2. Defining & Measuring (mis-)Calibration

QUESTION:
Are these predictions
under-confident
or
over-confident?



2. Defining & Measuring (mis-)Calibration

- **Reliability Plots**

  Not enough items with a given confidence to estimate population statistics decently:

  model predicts with p=0.2 → "20%" positives

  What if you only have 2 items predicted with p=0.2?

  We can group predictions in bins, and plot them against y=x.

- **Expected Calibration Error**

  The average of gaps across bins, weighted by bin population:

$$\mathrm{ECE} = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{|B_i|} \left| prob(B_i) - pos(B_i) \right|$$

MICCAI2025 Daejeon REPUBLIC OF KOREA

2. Defining & Measuring (mis-)Calibration

- **Generalizing from Binary to Multi-Class classifiers**

Confidence calibration: only consider highest probability.

$$\mathrm{ECE} = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{|B_i|} \left| prob(B_i) - pos(B_i) \right|$$

p	y
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	1
[⅔, ⅓, 0]	2

p	y
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	2
[0, ⅔, ⅓]	3

p	y
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	3
[⅓, 0, ⅔]	1

2. Defining & Measuring (mis-)Calibration

- **Generalizing from Binary to Multi-Class classifiers**

Confidence calibration: only consider highest probability.

$$\text{conf-ECE} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{|B_i|}\left|conf\left(B_i\right) - acc\left(B_i\right)\right|$$

p	(ŷ, c)	y
[⅔, ⅓, 0]	(1,⅔)	1
[⅔, ⅓, 0]	(1,⅔)	1
[⅔, ⅓, 0]	(1,⅔)	2

p	(ŷ, c)	y
[0, ⅔, ⅓]	(2,⅔)	2
[0, ⅔, ⅓]	(2,⅔)	2
[0, ⅔, ⅓]	(2,⅔)	3

p	(ŷ, c)	y
[⅓, 0, ⅔]	(3,⅔)	3
[⅓, 0, ⅔]	(3,⅔)	3
[⅓, 0, ⅔]	(3,⅔)	1

2. Defining & Measuring (mis-)Calibration

Calibration $\neq$ Discrimination

p	$\hat{y}$	y
(⅓+2ε, ⅓-ε, ⅓-ε)	1	1
(⅓-ε, ⅓+2ε, ⅓-ε)	2	1
(⅓-ε, ⅓-ε, ⅓+2ε)	3	2

p	$\hat{y}$	y
(⅓+2ε, ⅓-ε, ⅓-ε)	1	2
(⅓-ε, ⅓+2ε, ⅓-ε)	2	3
(⅓-ε, ⅓-ε, ⅓+2ε)	3	3

MICCAI2025 Daejeon REPUBLIC OF KOREA

2. Defining & Measuring (mis-)Calibration

Calibration ≠ Discrimination

p	ŷ	y
(⅓+2ε, ⅓-ε, ⅓-ε)	1	1
(⅓-ε, ⅓+2ε, ⅓-ε)	2	1
(⅓-ε, ⅓-ε, ⅓+2ε)	3	2

p	ŷ	y
(⅓+2ε, ⅓-ε, ⅓-ε)	1	2
(⅓-ε, ⅓+2ε, ⅓-ε)	2	3
(⅓-ε, ⅓-ε, ⅓+2ε)	3	3

This 3-class classifier predicts randomly with full uncertainty.
It always has a confidence of ~⅓, and it has an accuracy of ⅓.
Therefore it is perfectly calibrated, but useless.

2. Defining & Measuring (mis-)Calibration

Calibration $\neq$ Discrimination

p	ŷ	y
(⅔, 0, ⅓)	1	1
(0, ⅓, ⅔)	2	1
(⅓, ⅔, 0)	2	2

p	ŷ	y
(0, ⅓, ⅔)	1	2
(⅓, 0, ⅔)	2	3
(0, ⅓, ⅔)	3	3

2. Defining & Measuring (mis-)Calibration

Calibration ≠ Discrimination

p	ŷ	y
(⅔, 0, ⅓)	1	1
(0, ⅓, ⅔)	2	1
(⅓, ⅔, 0)	2	2

p	ŷ	y
(0, ⅓, ⅔)	1	2
(⅓, ⅔, 0)	2	2
(0, ⅓, ⅔)	3	3

This classifier always predicts with ⅔ confidence. Also, it has an accuracy of ⅔. It is perfectly confidence-calibrated, but it is more discriminative than random guessing.

2. Defining & Measuring (mis-)Calibration

Calibration $\neq$ Discrimination

p	$\hat{y}$	y
(1, 0, 0)	1	1
(1, 0, 0)	1	1
(0, 1, 0)	2	2

p	$\hat{y}$	y
(0, 1, 0)	2	2
(0, 0, 1)	3	3
(0, 0, 1)	3	3

2. Defining & Measuring (mis-)Calibration

Calibration $\neq$ Discrimination

p	$\hat{y}$	y
(1, 0, 0)	1	1
(1, 0, 0)	1	1
(0, 1, 0)	2	2

p	$\hat{y}$	y
(0, 1, 0)	2	2
(0, 0, 1)	3	3
(0, 0, 1)	3	3

This one is always 100% confident, and always right.
It is fully-calibrated and perfectly discriminative.

MICCAI2025 Daejeon REPUBLIC OF KOREA

2. Measuring Calibration

- **Proper Scoring Rules**

Measure discrimination+calibration at individual item level

Most popular: Brier Score, Negative Log-Likelihood

$$\mathrm{Brier}(\mathbf{p}, \mathbf{y}) = \|\mathbf{p} - \mathbf{y}\|_2^2 \qquad \mathrm{NLL}(\mathbf{p}, \mathbf{y}) = -\log(\mathbf{p}_y)$$

Example: $y = 3,\ \mathbf{y} = (0, 0, 1),\ \mathbf{p}_{\mathrm{bad}} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right),\ \mathbf{p}_{\mathrm{better}} = \left(0, \frac{1}{3}, \frac{2}{3}\right)$

$$\mathrm{Brier}(\mathbf{p}_{\mathrm{bad}}, \mathbf{y}) = 2/3 \quad \mathrm{Brier}(\mathbf{p}_{\mathrm{better}}, \mathbf{y}) = 2/9 \quad \mathrm{Brier}(\mathbf{y}, \mathbf{y}) = 0$$

$$\mathrm{NLL}(\mathbf{p}_{\mathrm{bad}}, \mathbf{y}) \approx 0.477 \quad \mathrm{NLL}(\mathbf{p}_{\mathrm{better}}, \mathbf{y}) \approx 0.176 \quad \mathrm{NLL}(\mathbf{y}, \mathbf{y}) = 0$$

Note that a fully uncertain prediction $\mathbf{p}_{\mathrm{bad}}$ does not score well.

3. Improving Calibration

- **Post-Training Calibration**

Classic methods: **Platt Scaling** & **Isotonic Regression**:

- Platt: Fits a logistic regression model using validation set.
- Isotonic: Fits a monotonic piecewise constant mapping, optimizing bins to maximize calibration.

Temperature Scaling: Uses a validation set to learn a scalar T dividing logits before applying softmax and tempers their value:

$$p_{\mathrm{j}} = \frac{\mathrm{e}^{\mathrm{z_j}}}{\sum_{\mathrm{k=1}}^{\mathrm{N}} \mathrm{e}^{\mathrm{z_k}}} \longmapsto p_j = \frac{\mathrm{e}^{(\mathrm{z_j}/\mathrm{T})}}{\sum_{\mathrm{k=1}}^{\mathrm{N}} \mathrm{e}^{(\mathrm{z_k}/\mathrm{T})}}$$

We will see code examples in a minute

3. Improving Calibration

- **Model Ensembling**

Ensembling several diverse models can improve calibration.
Of course it comes with a computational overhead.

- **Training Time Calibration**

Over-parametrized NNs can keep on learning the training set until they are fully confident, minimizing NLL indefinitely. We can avoid this by **regularizing** so as to **disencourage confidence**.

Label Smoothing, MixUp, Focal Loss... Careful of **underfitting**!
Always report also AUC/ACC/DSC/..., not only ECE

We will see code examples in a minute

4. Hands-On

Github repository:

https://github.com/agaldran/uqinmia-miccai