

# Part IV

# Model calibration

# Conformal Prediction

Adrian Galdran  
RyC/ikerbasque Research Fellow  
Tecnalia - Derio, Spain  
[adrian.galdran@tecnalia.com](mailto:adrian.galdran@tecnalia.com)



# 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**.*

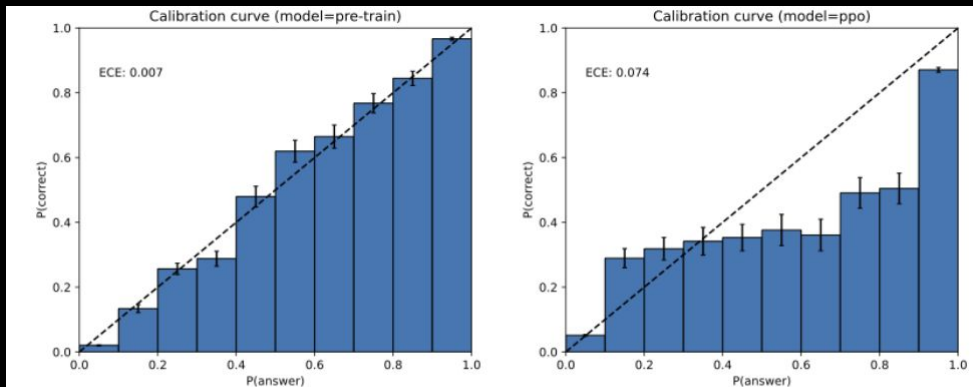


Figure 2: GPT-4 calibration histograms before (left) and after (right) reinforcement learning (OpenAI, 2023a, Figure 8, reprinted with permission). These plots are for multiple-choice queries where the plausible responses are simply A, B, C, or D. The pretrained model is well calibrated.

# 1. Motivation - Why Calibration?

## The Critical View of Safety ✓ or ✗

Criterion 1 (C1)

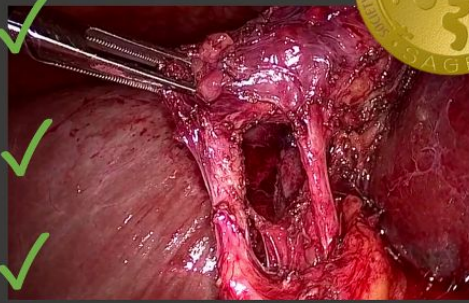
**Two and only two** tubular structures are seen connected to the gallbladder ✓

Criterion 2 (C2)

The **hepatocystic triangle** is cleared from fat and/or connective tissue so that an unimpeded view is obtained ✓

Criterion 3 (C3)

The lower part of the gallbladder is dissected off the liver bed to expose the lower **1/3 of the cystic plate**. ✓



# 1. Motivation - Why Calibration?

## The Critical View of Safety ✓ or ✗

Criterion 1 (C1)

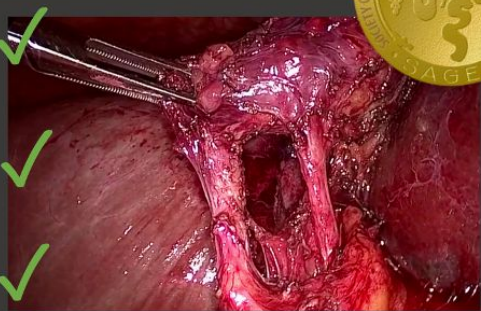
**Two and only two** tubular structures are seen connected to the gallbladder ✓

Criterion 2 (C2)

The **hepatocystic triangle** is cleared from fat and/or connective tissue so that an unimpeded view is obtained ✓

Criterion 3 (C3)

The lower part of the gallbladder is dissected off the liver bed to expose the lower **1/3 of the cystic plate**. ✓



*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
$\frac{1}{6}$	0
$\frac{1}{6}$	0
$\frac{1}{6}$	0
$\frac{1}{6}$	0
$\frac{1}{6}$	1

p	y
$\frac{1}{3}$	0
$\frac{1}{3}$	0
$\frac{1}{3}$	1
$\frac{1}{2}$	0
$\frac{1}{2}$	1

p	y
$\frac{3}{4}$	0
$\frac{3}{4}$	1
$\frac{3}{4}$	1
$\frac{3}{4}$	1
1	1

## 2. Defining & Measuring Calibration

p	y
$\frac{1}{6}$	0
$\frac{1}{6}$	0
$\frac{1}{6}$	0
$\frac{1}{6}$	0
$\frac{1}{6}$	1

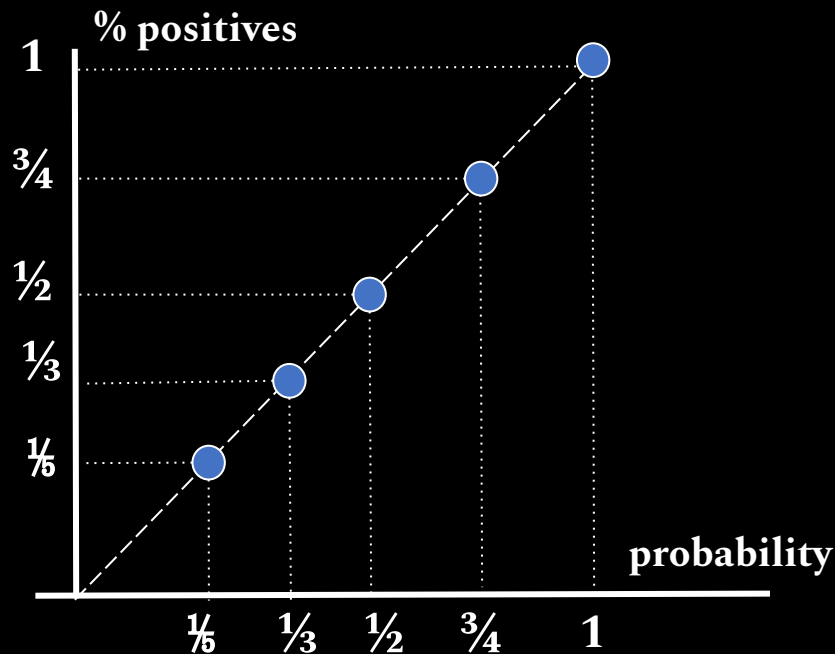
p	y
$\frac{1}{3}$	0
$\frac{1}{3}$	0
$\frac{1}{3}$	1
$\frac{1}{2}$	0
$\frac{1}{2}$	1

p	y
$\frac{3}{4}$	0
$\frac{3}{4}$	1
$\frac{3}{4}$	1
$\frac{3}{4}$	1
1	1



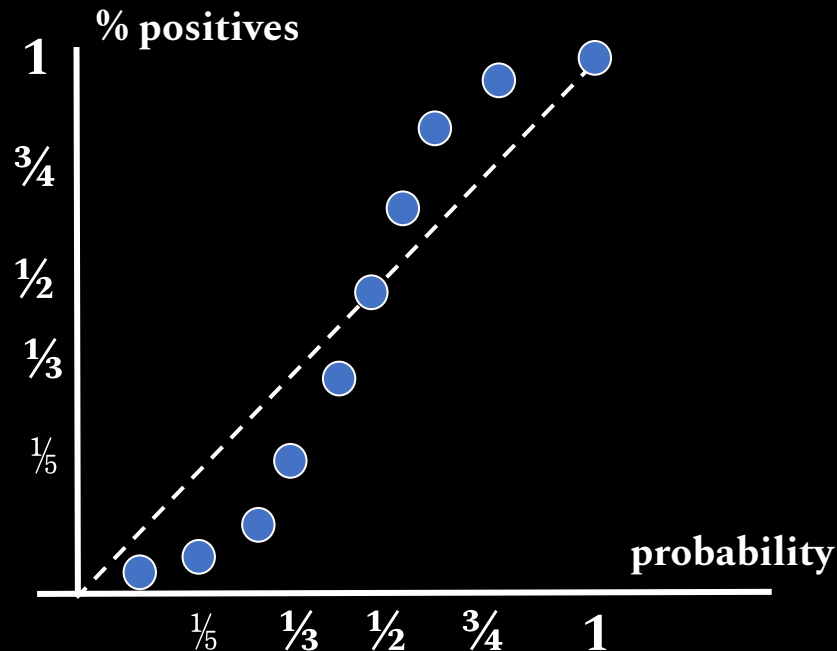
## 2. Defining & Measuring Calibration

p	y	p	y	p	y
$\frac{1}{6}$	0	$\frac{1}{3}$	0	$\frac{3}{4}$	0
$\frac{1}{6}$	0	$\frac{1}{3}$	0	$\frac{3}{4}$	1
$\frac{1}{6}$	0	$\frac{1}{3}$	1	$\frac{3}{4}$	1
$\frac{1}{6}$	0	$\frac{1}{2}$	0	$\frac{3}{4}$	1
$\frac{1}{6}$	1	$\frac{1}{2}$	1	1	1



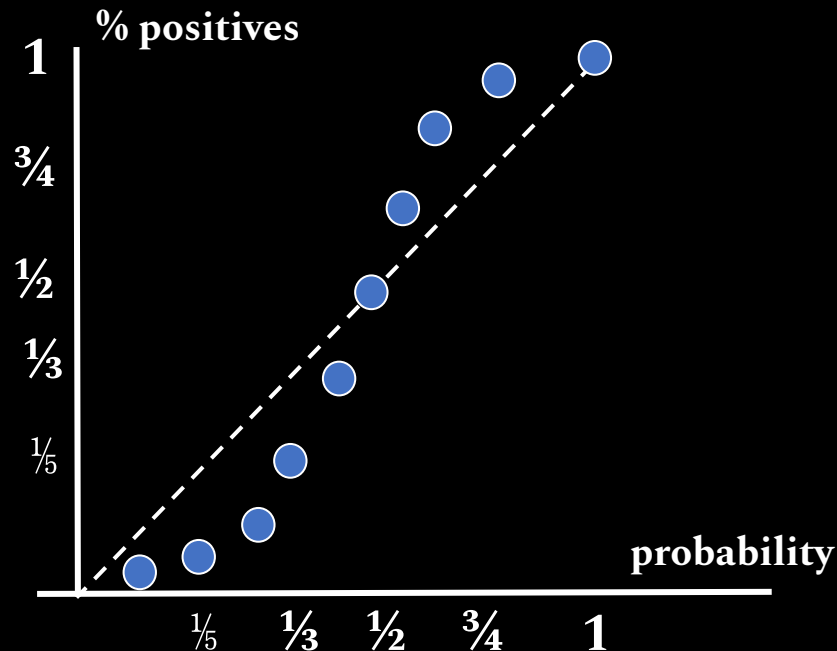
## 2. Defining & Measuring Calibration

**QUESTION:**  
Are these predictions  
under-confident  
or  
over-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  $p=0.2 \rightarrow$  “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:

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

## 2. Defining & Measuring (**mis-**)Calibration

- Generalizing from Binary to Multi-Class classifiers

**Confidence calibration:** only consider highest probability.

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

p	y
$[\frac{2}{3}, \frac{1}{3}, 0]$	1
$[\frac{2}{3}, \frac{1}{3}, 0]$	1
$[\frac{2}{3}, \frac{1}{3}, 0]$	2

p	y
$[0, \frac{2}{3}, \frac{1}{3}]$	2
$[0, \frac{2}{3}, \frac{1}{3}]$	2
$[0, \frac{2}{3}, \frac{1}{3}]$	3

p	y
$[\frac{1}{3}, 0, \frac{2}{3}]$	3
$[\frac{1}{3}, 0, \frac{2}{3}]$	3
$[\frac{1}{3}, 0, \frac{2}{3}]$	1

## 2. Defining & Measuring (**mis**-)Calibration

- Generalizing from Binary to Multi-Class classifiers

**Confidence calibration:** only consider highest probability.

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

p	( $\hat{y}$ , c)	y
[ $\frac{2}{3}$ , $\frac{1}{3}$ , 0]	( <b>1</b> , $\frac{2}{3}$ )	1
[ $\frac{2}{3}$ , $\frac{1}{3}$ , 0]	( <b>1</b> , $\frac{2}{3}$ )	1
[ $\frac{2}{3}$ , $\frac{1}{3}$ , 0]	( <b>1</b> , $\frac{2}{3}$ )	2

p	( $\hat{y}$ , c)	y
[0, $\frac{2}{3}$ , $\frac{1}{3}$ ]	( <b>2</b> , $\frac{2}{3}$ )	2
[0, $\frac{2}{3}$ , $\frac{1}{3}$ ]	( <b>2</b> , $\frac{2}{3}$ )	2
[0, $\frac{2}{3}$ , $\frac{1}{3}$ ]	( <b>2</b> , $\frac{2}{3}$ )	3

p	( $\hat{y}$ , c)	y
[ $\frac{1}{3}$ , 0, $\frac{2}{3}$ ]	( <b>3</b> , $\frac{2}{3}$ )	3
[ $\frac{1}{3}$ , 0, $\frac{2}{3}$ ]	( <b>3</b> , $\frac{2}{3}$ )	3
[ $\frac{1}{3}$ , 0, $\frac{2}{3}$ ]	( <b>3</b> , $\frac{2}{3}$ )	1

## 2. Defining & Measuring (**mis-**)Calibration

**Calibration  $\neq$  Discrimination**

$\mathbf{p}$	$\hat{\mathbf{y}}$	$\mathbf{y}$
$(\frac{1}{3}+2\varepsilon, \frac{1}{3}-\varepsilon, \frac{1}{3}-\varepsilon)$	<b>1</b>	<b>1</b>
$(\frac{1}{3}-\varepsilon, \frac{1}{3}+2\varepsilon, \frac{1}{3}-\varepsilon)$	<b>2</b>	<b>1</b>
$(\frac{1}{3}-\varepsilon, \frac{1}{3}-\varepsilon, \frac{1}{3}+2\varepsilon)$	<b>3</b>	<b>2</b>

$\mathbf{p}$	$\hat{\mathbf{y}}$	$\mathbf{y}$
$(\frac{1}{3}+2\varepsilon, \frac{1}{3}-\varepsilon, \frac{1}{3}-\varepsilon)$	<b>1</b>	<b>2</b>
$(\frac{1}{3}-\varepsilon, \frac{1}{3}+2\varepsilon, \frac{1}{3}-\varepsilon)$	<b>2</b>	<b>3</b>
$(\frac{1}{3}-\varepsilon, \frac{1}{3}-\varepsilon, \frac{1}{3}+2\varepsilon)$	<b>3</b>	<b>3</b>

## 2. Defining & Measuring (**mis-**)Calibration

**Calibration  $\neq$  Discrimination**

p	$\hat{y}$	y
$(\frac{1}{3}+2\varepsilon, \frac{1}{3}-\varepsilon, \frac{1}{3}-\varepsilon)$	1	1
$(\frac{1}{3}-\varepsilon, \frac{1}{3}+2\varepsilon, \frac{1}{3}-\varepsilon)$	2	1
$(\frac{1}{3}-\varepsilon, \frac{1}{3}-\varepsilon, \frac{1}{3}+2\varepsilon)$	3	2

p	$\hat{y}$	y
$(\frac{1}{3}+2\varepsilon, \frac{1}{3}-\varepsilon, \frac{1}{3}-\varepsilon)$	1	2
$(\frac{1}{3}-\varepsilon, \frac{1}{3}+2\varepsilon, \frac{1}{3}-\varepsilon)$	2	3
$(\frac{1}{3}-\varepsilon, \frac{1}{3}-\varepsilon, \frac{1}{3}+2\varepsilon)$	3	3

This 3-class classifier predicts randomly with full uncertainty. It always has a confidence of  $\sim \frac{1}{3}$ , and it has an accuracy of  $\frac{1}{3}$ . Therefore it is **perfectly calibrated**, but **useless**.



## 2. Defining & Measuring (**mis-**)Calibration

**Calibration  $\neq$  Discrimination**

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

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

## 2. Defining & Measuring (**mis-**)Calibration

**Calibration  $\neq$  Discrimination**

p	$\hat{y}$	y
( <b>1</b> , 0, 0)	<b>1</b>	<b>1</b>
( <b>1</b> , 0, 0)	<b>1</b>	<b>1</b>
(0, <b>1</b> , 0)	<b>2</b>	<b>2</b>

p	$\hat{y}$	y
(0, <b>1</b> , 0)	<b>2</b>	<b>2</b>
(0, 0, <b>1</b> )	<b>3</b>	<b>3</b>
(0, 0, <b>1</b> )	<b>3</b>	<b>3</b>

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

# 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.

# 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:

$$p_j = \frac{e^{z_j}}{\sum_{k=1}^N e^{z_k}} \mapsto p_j = \frac{e^{(z_j/T)}}{\sum_{k=1}^N e^{(z_k/T)}}$$

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.

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-parameterized 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>

