NYU

# Evaluation

Yilun Kuang

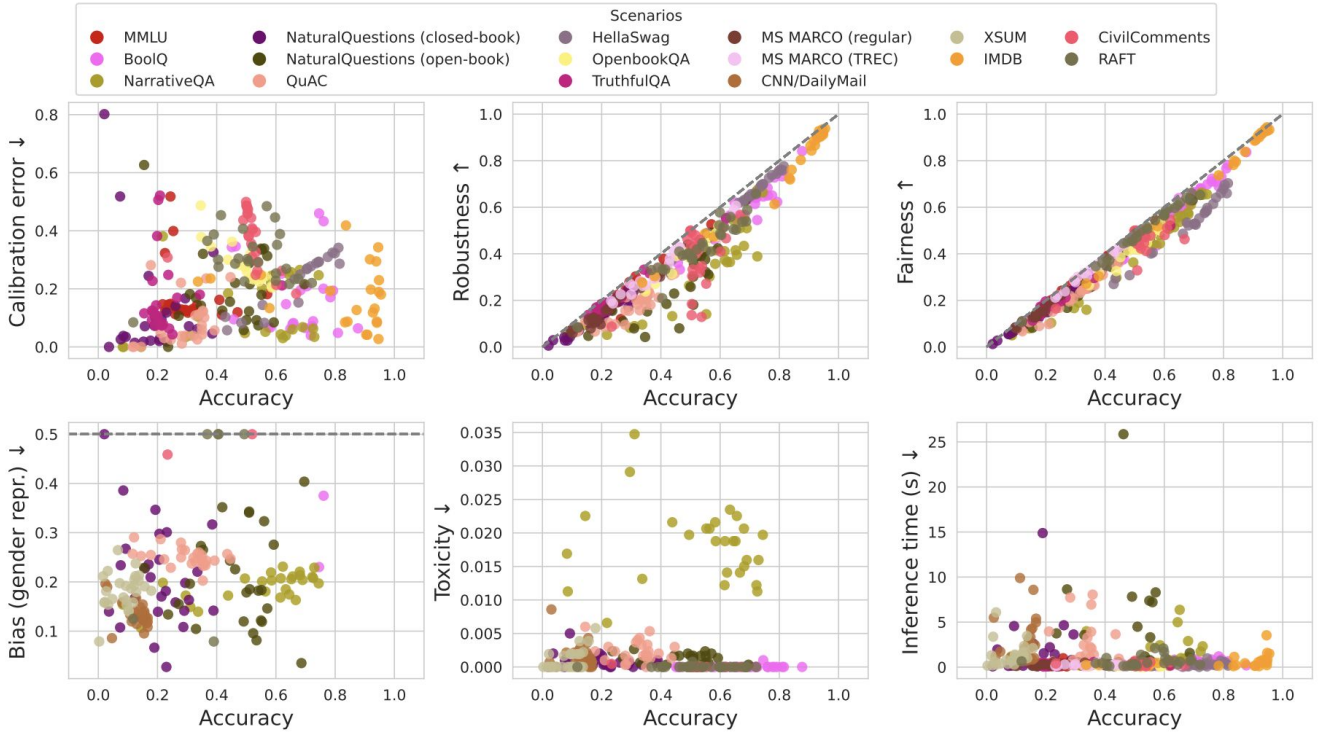
10/30/2023

# Outline

- Review of Evaluation Metrics
- Python Implementation of Selective Prediction and ECE for Language Models

# Metrics Across Scenarios

**Holistic Evaluation of Language Models (HELM)**



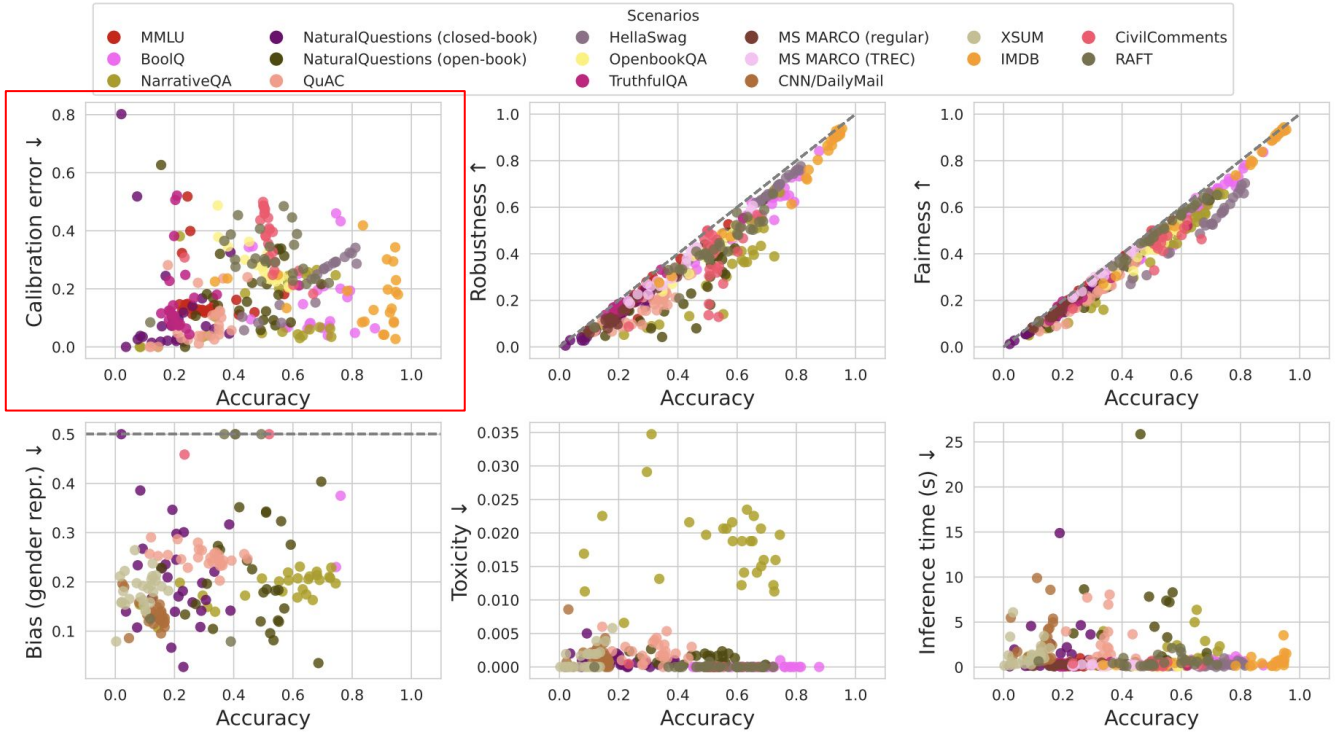

NYU



Source: https://arxiv.org/pdf/2211.09110.pdf

# Metrics Across Scenarios

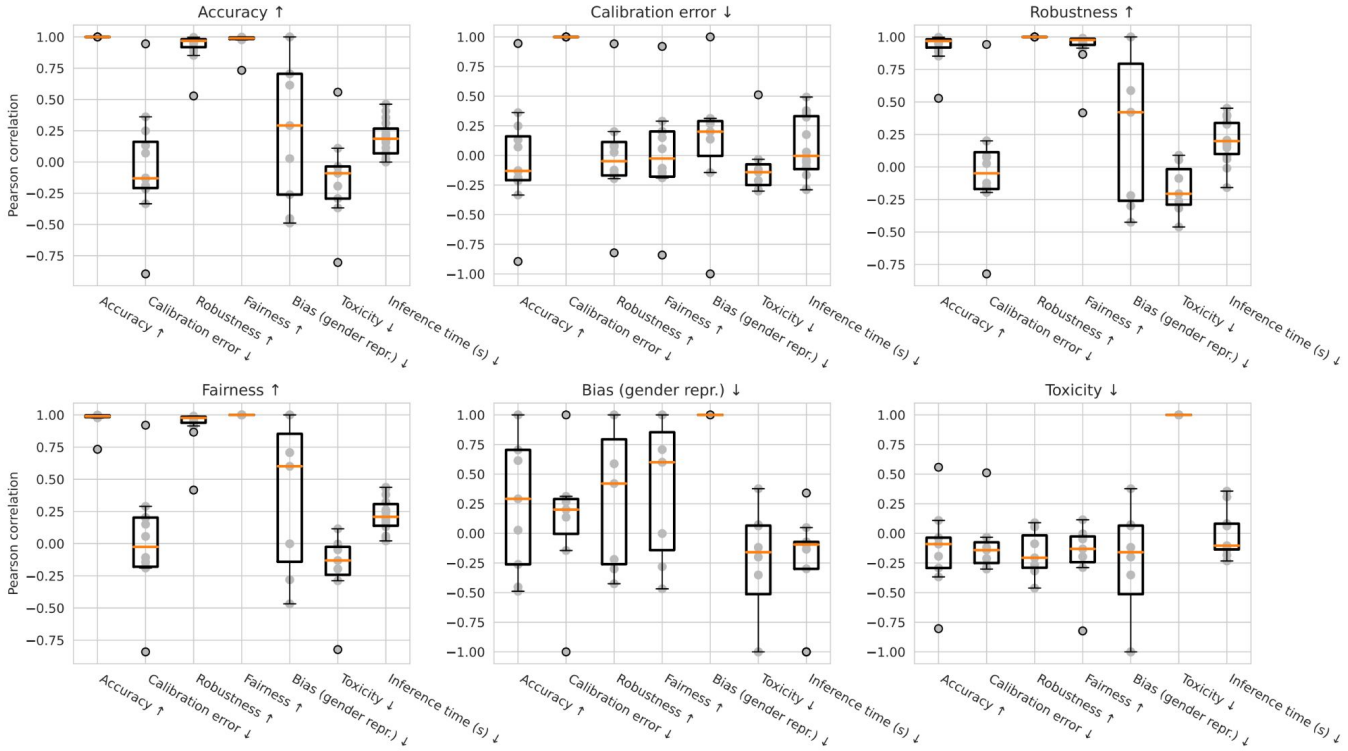
**Holistic Evaluation of Language Models (HELM)**






Source: https://arxiv.org/pdf/2211.09110.pdf

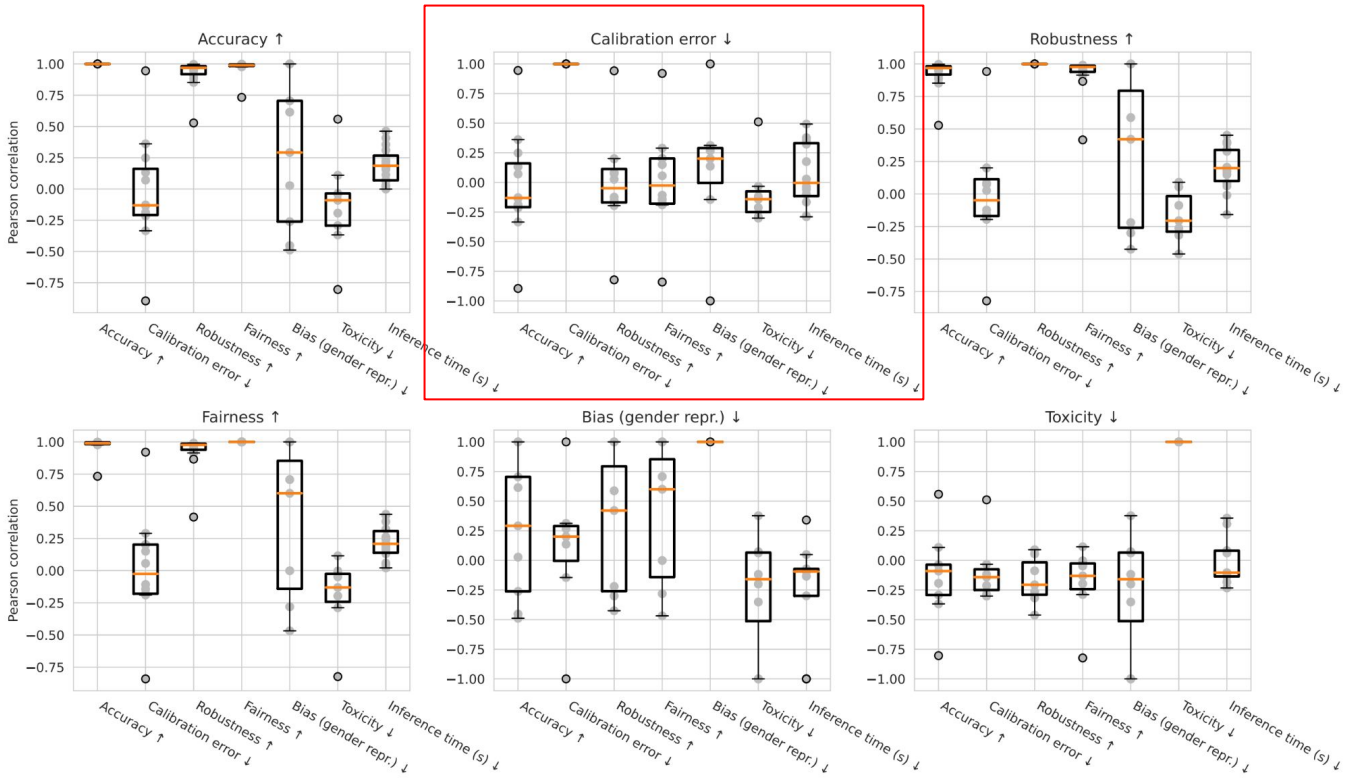
# Correlation between Metrics






Source: https://arxiv.org/pdf/2211.09110.pdf

# Correlation between Metrics




NYU



Source: https://arxiv.org/pdf/2211.09110.pdf

# Expected Calibration Error

**Confidence**

A model's confidence is defined as the $p_\theta(y^*|x)$ where $y^* = \text{argmax}_y p_\theta(y|x)$

**ECE**

We can obtain ECE by splitting the predictions into bins $B_1, ..., B_M$ by confidence scores

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{accuracy}(B_m) - \text{confidence}(B_m)\right|$$

NYU



Source: https://github.com/nyu-cs2590/course-material/blob/gh-pages/fall2023/lecture/lec08/main.pdf

# Expected Calibration Error

| Probabilities of model predictions: | 0.0 | 0.1 | 0.2 | 0.3 | 0.7 | 0.8 | 0.9 | 1.0 |
|---|---|---|---|---|---|---|---|---|
| | ✔ | ✘ | ✘ | ✔ | ✔ | ✘ | ✔ | ✔ |
| Equal-sized bins: | Bin 1 | | | | Bin 2 | | | |

Accuracy = 2/4 = 0.5
Prob = (0.0 + 0.1 + 0.2 + 0.3) / 4 = 0.15
Bin-1 error = |0.5 - 0.15| = 0.35

Accuracy = 3/4 = 0.75
Prob = (0.7 + 0.8 + 0.9 + 1.0) / 4 = 0.85
Bin-2 error = |0.75 - 0.85| = 0.1

**ECE (expected calibration error)** = (4/8) * 0.35 + (4/8) * 0.1 = 0.225

NYU



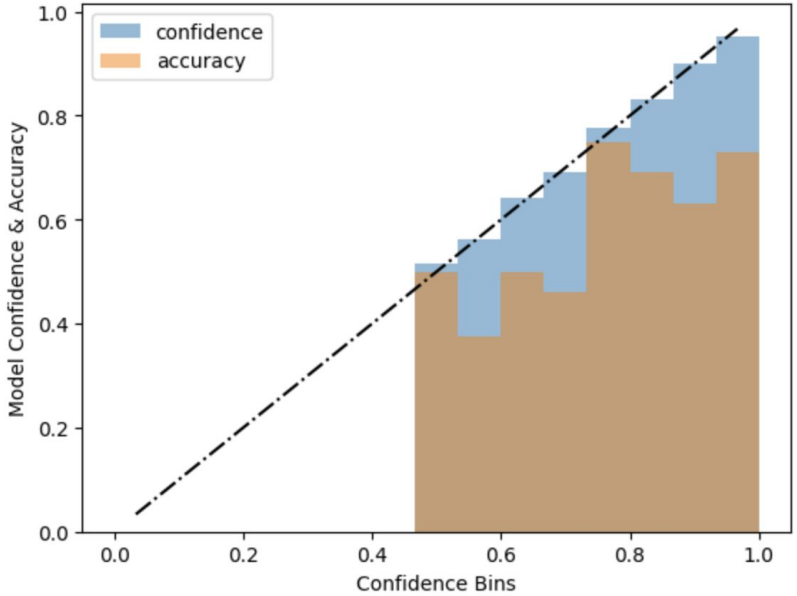
Source: https://arxiv.org/pdf/2211.09110.pdf

# Expected Calibration Error

**Model:** BERTforSequence Classification

**Task:** RTE Dataset in GLUE

**Implementation:** See Jupyter Notebook




NYU

# Selective Prediction

**Coverage**

Given a threshold $t \geq 0$, the coverage $c(t)$ is the fraction of examples for which the model's confidence is at least $t$

**Selective Accuracy**

selective accuracy $a(t)$ at a threshold $t \geq 0$ is the accuracy for all examples where the model's predicted confidence is at least $t$

NYU



Source:
https://arxiv.org/pdf/2211.09110.pdf

# Selective Prediction



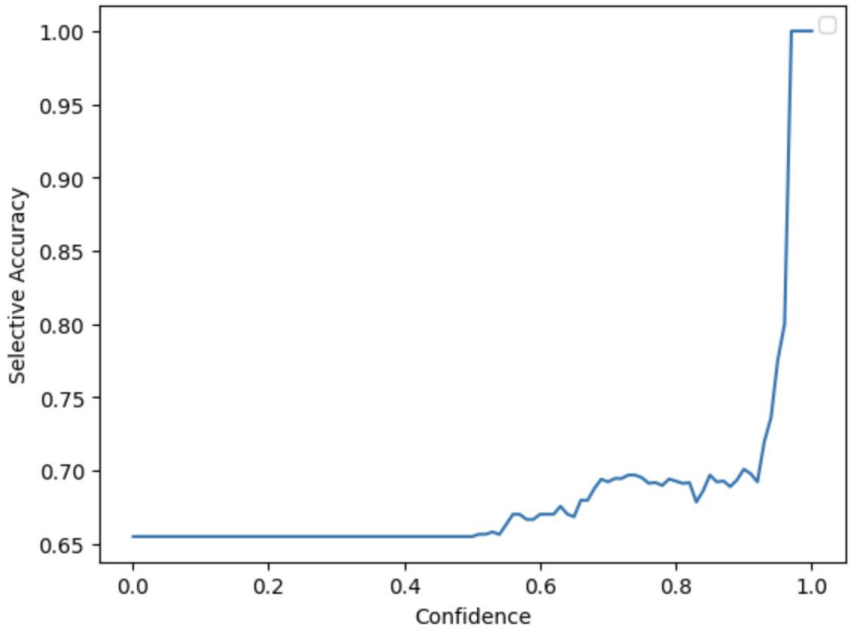

Source: https://arxiv.org/pdf/2211.09110.pdf

# Selective Coverage-Accuracy Area (SCAA)

**Model:** BERTforSequence Classification

**Task:** RTE Dataset in GLUE

**Implementation:** See Jupyter Notebook




NYU

# Acknowledgement

This presentation is adapted from Holistic Evaluation of Language Models (https://arxiv.org/pdf/2211.09110.pdf) and Lecture 8 (https://github.com/nyu-cs2590/course-material/tree/gh-pages/fall2023/lecture/lec08)