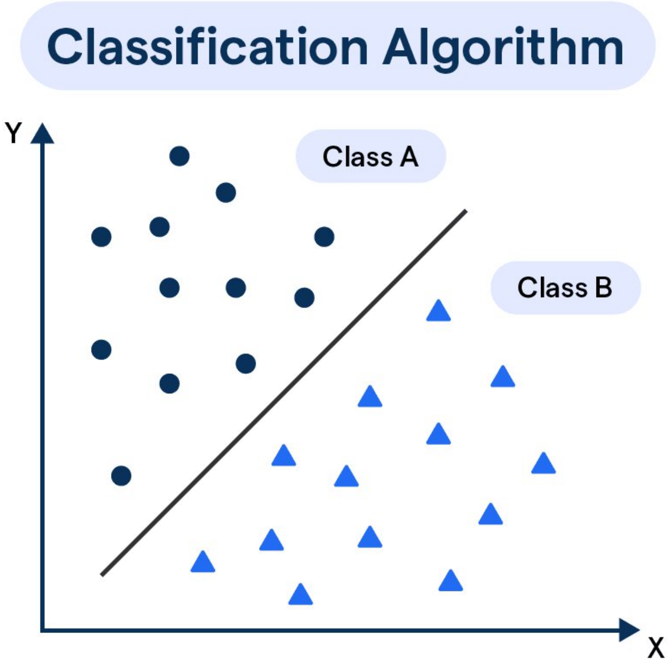
Fill In The Gaps: Model Calibration and Generalization with Synthetic Data

Yang Ba, Michelle V. Mancenido, Rong Pan

Ira A. Fulton Schools of Engineering
Arizona State University

Classifier Evaluation



When evaluating a classifier, we usually use metrics such as: accuracy, F1, ROC etc.

Two models:

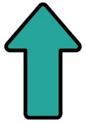
1. 90% accuracy, 91% confidence in predictions
2. 😕 90% accuracy, 99% confidence in predictions

Model Uncertainty

In high stake areas, model uncertainty even more important.

stock trading

disease diagnosis



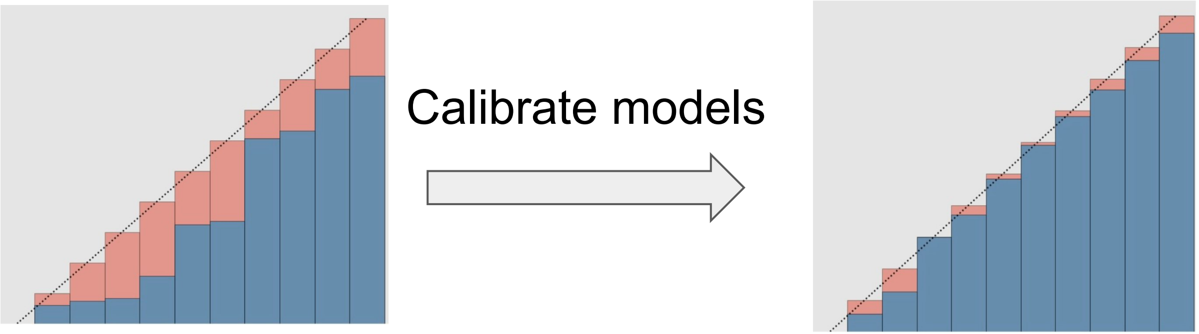
51% vs 99%

Will your actions be different according to these two different prediction confidences?

Model Calibration

Calibration: to align a model's predicted probabilities (confidence) with its actual outcomes (accuracy).

Evaluation Metric: Expected Calibration Error (ECE)



Reliability Diagram

Reliability Diagram

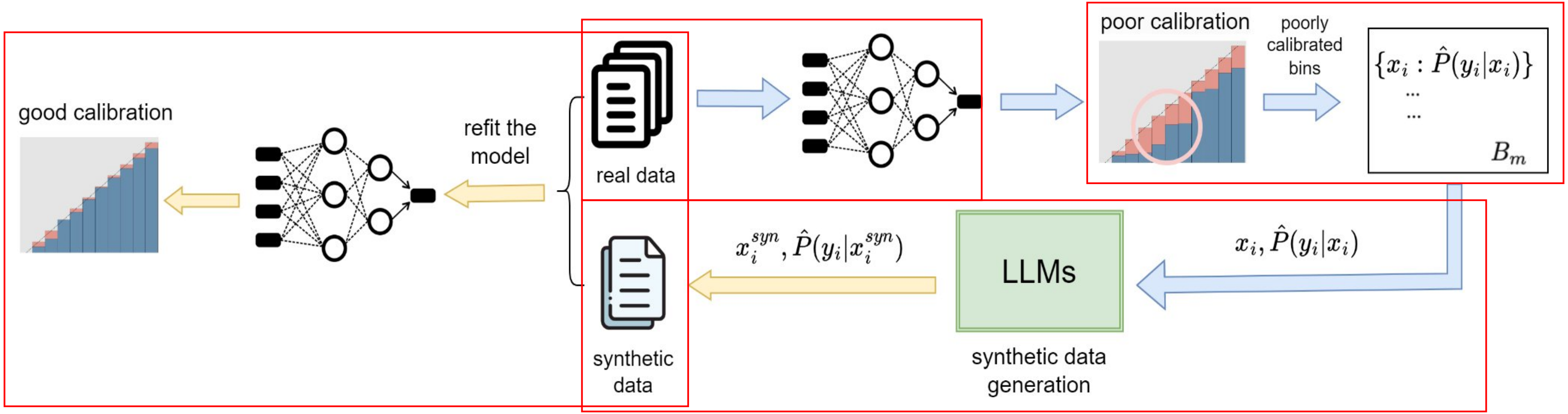
$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{Acc}(B_m) - \mathrm{Conf}(B_m)\right|$$

Motivation

The current calibration methods only focus on calibration and ignore model accuracy, which will either potentially hurt the model prediction performance or maintain it, such as *Isotonic regression,Platt scaling*, *Monte Carlo dropout, Temperature scaling.*

Can we calibrate models while improving accuracy, or at least without sacrificing it?

Our Proposed Framework



- We develop a theoretical framework to solve this problem, leveraging LLM-generated synthetic data to calibrate downstream NLP models and increase their accuracy at the same time.
- We extend Probably Approximately Correct (PAC) learning framework to derive the **Expected Calibration Error Bound**, guiding us in synthetic data generation and model calibration

Expected Calibration Error Bound

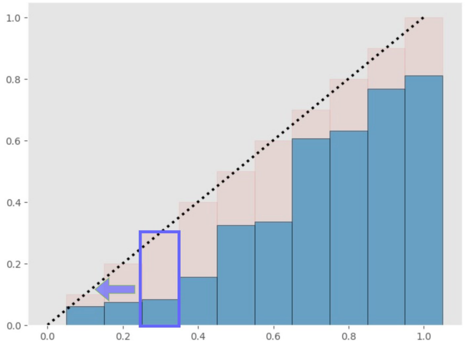
Proposition. *Given n training samples, if the probability of the difference between the expected model parameter and its estimated value being less than ϵ_a is at least $(1-\delta_a)\%$, then the probability of the difference between the expected calibration error and the estimated calibration error in the training samples being less than ϵ_{ECE} is at least $(1-\delta_{ECE})\%$. Here, $\delta_{ECE} = 2\delta_a$, and $\epsilon_{ECE} = \epsilon_a + |Conf(X) - Conf(X^*)| = \epsilon_a + \sum_{m=1}^{M} \frac{|B_m|}{n} |Conf(B_m) - Conf(B_m^*)|$.*

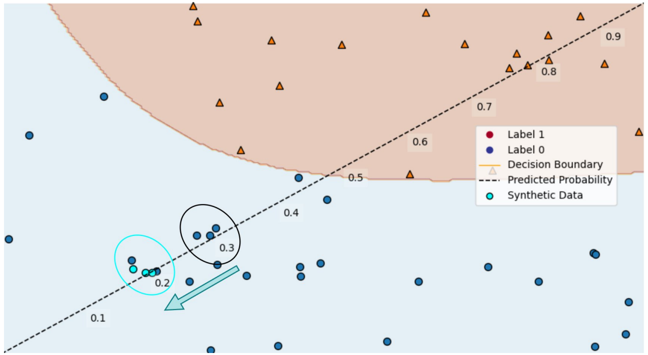
Based on ECE bound, we can manipulate the prediction probability by synthetic data to minimize the difference $|\mathrm{Conf}(X) - \mathrm{Conf}(X^*)|$

(Refer the paper to check out the detailed proof and remarks.)

Synthetic Data Generation Strategy

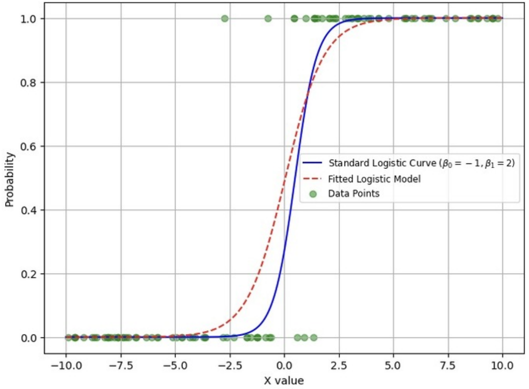
	Over Confidence	Under Confidence
Low Probability ($\hat{P}(y_i \mid x_i) \leq 0.5$)	Decrease predicted prob (Move away from DB)	Increase predicted prob (Move towards DB)
High Probability ($\hat{P}(y_i \mid x_i) > 0.5$)	Increase predicted prob (Move towards DB)	Increase predicted prob (Move away from DB)

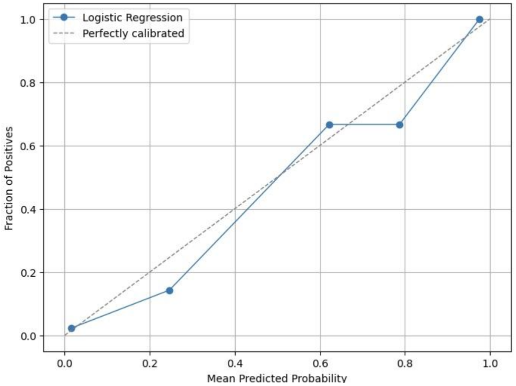




Generating synthetic data to address miscalibration gaps.

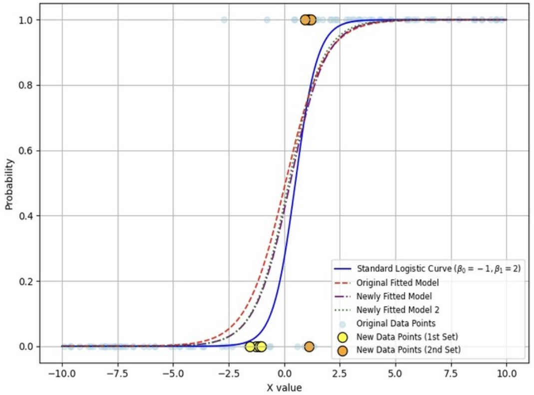
Toy example

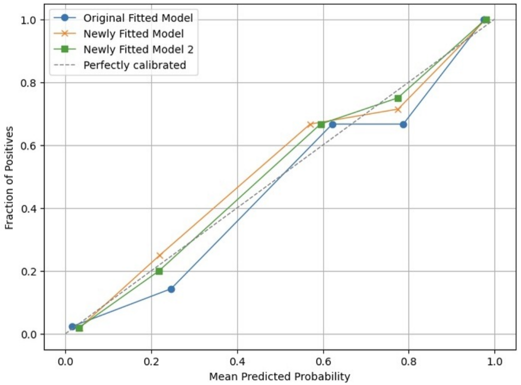




Original fitted model :

β_0 = -0.06 and β_1 = 1.13,
ACC: 0.95, ECE: 0.0405;





Newly fitted model :

β_0 = -0.339 and β_1 = 1.2627,
ACC: 0.95327, ECE: 0.0424;

Newly fitted model 2 :

β_0 = -0.2558 and β_1 = 1.2953,
ACC: 0.9469, ECE: 0.0366;

Experiments

Tasks: TC, SUBJ, B77, SE, Arxiv, Medical

Model: $BERT_{base}$

Baseline: without any calibration

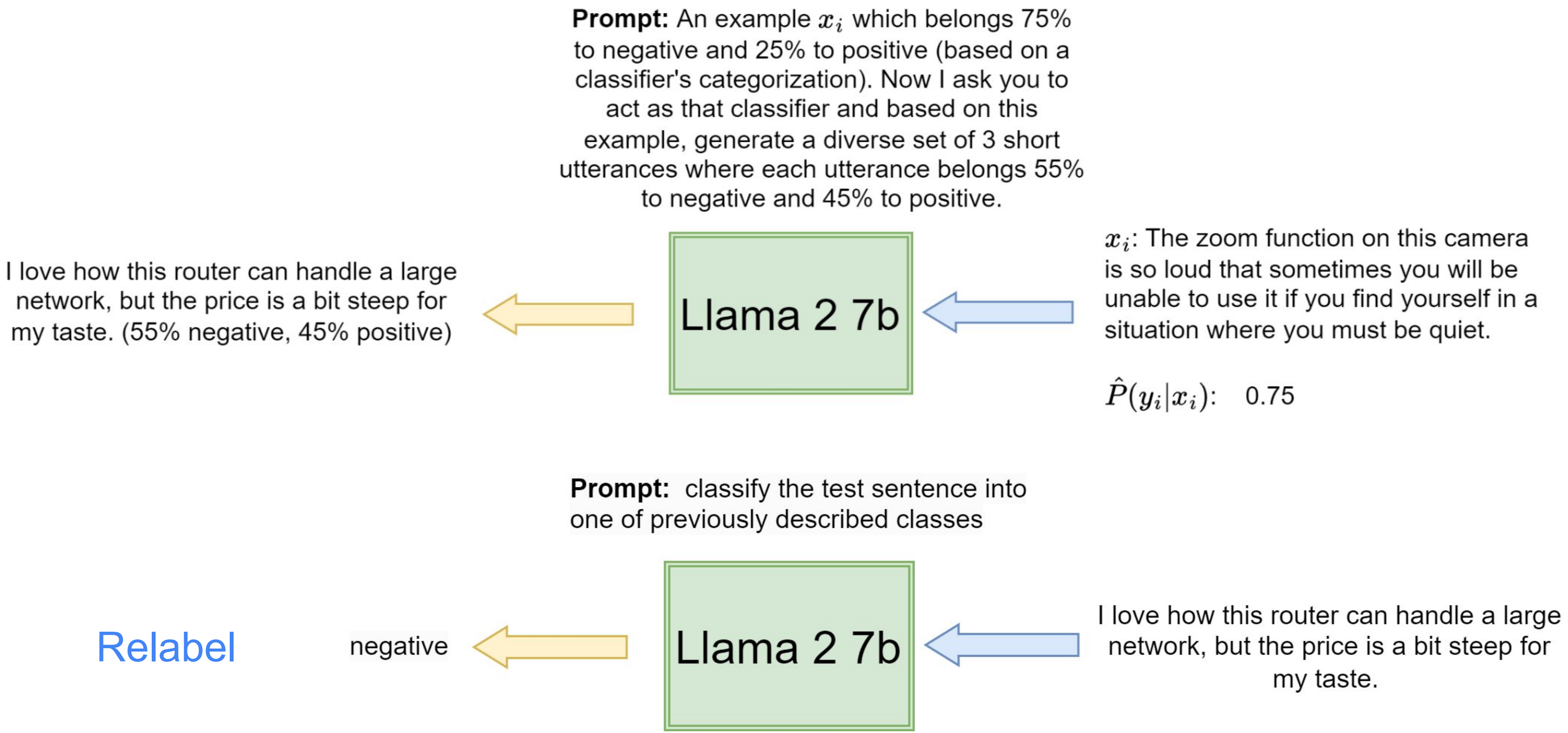
Synthesis: synthetic data replacement (keep the training data size the same)

Synthesis+: synthetic data add-on (increase the training data size)

Comparison: Isotonic regression, Platt scaling, Monte Carlo dropout, Temperature scaling

N_bins: 10, 15, 20

Sample Synthetic Data(SE, high probability & overconfidence)



Results

Metric	TC ACC	TC ECE	SUBJ ACC	SUBJ ECE	B77 ACC	B77 ECE	SE ACC	SE ECE	Arxiv ACC	Arxiv ECE	Medical ACC	Medical ECE
Baseline	0.867 (0.00)	0.058 (0.02)	0.955 (0.01)	0.034 (0.01)	0.708 (0.12)	0.234 (0.04)	0.884 (0.01)	0.06 (0.00)	0.805 (0.00)	0.105 (0.01)	0.864 (0.00)	0.051 (0.01)
Isotonic	0.871 (0.00)	0.082 (0.01)	0.959 (0.00)	0.027 (0.01)	0.850 (0.02)	0.063 (0.01)	0.890 (0.01)	0.058 (0.01)	0.812 (0.01)	0.114 (0.01)	0.869 (0.01)	0.069 (0.01)
Platt scaling	0.863 (0.01)	0.086 (0.01)	0.955 (0.01)	0.029 (0.00)	0.846 (0.03)	0.207 (0.03)	0.888 (0.01)	0.068 (0.00)	0.807 (0.01)	0.122 (0.00)	0.869 (0.01)	0.065 (0.01)
MC dropout	0.868 (0.02)	0.054 (0.01)	0.952 (0.01)	0.032 (0.01)	0.821 (0.23)	0.274 (0.14)	0.876 (0.01)	0.050 (0.02)	0.799 (0.01)	0.058 (0.04)	0.871 (0.01)	0.070 (0.01)
Temp scaling	0.867 (0.01)	0.049 (0.01)	0.955 (0.01)	0.026 (0.01)	0.708 (0.12)	0.253 (0.17)	0.884 (0.01)	0.038 (0.00)	0.805 (0.00)	0.070 (0.01)	0.864 (0.00)	0.056 (0.01)
10 bins												
Synthesis	0.867 (0.01)	0.053 (0.01)	0.960 (0.01)	0.027 (0.01)	0.625 (0.07)	0.255 (0.10)	0.871 (0.00)	0.055 (0.02)	0.815 (0.01)	0.077 (0.03)	0.873 (0.01)	0.048 (0.01)
Synthesis+	0.886 (0.01)	0.046 (0.01)	0.961 (0.00)	0.03 (0.00)	0.792 (0.20)	0.231 (0.03)	0.889 (0.01)	0.064 (0.00)	0.808 (0.01)	0.099 (0.01)	**0.871 (0.00)**	**0.047 (0.01)**
15 bins												
Synthesis	0.879 (0.01)	0.049 (0.01)	0.961 (0.00)	0.026 (0.00)	0.800 (0.11)	0.224 (0.08)	**0.904 (0.00)**	**0.04 (0.00)**	0.802 (0.00)	0.096 (0.01)	0.875 (0.00)	0.052 (0.00)
Synthesis+	0.881 (0.01)	0.050 (0.01)	**0.9605 (0.00)**	**0.024 (0.00)**	0.863 (0.09)	0.203 (0.10)	0.901 (0.01)	0.055 (0.01)	0.824 (0.01)	0.087 (0.01)	0.879 (0.01)	0.055 (0.01)
20 bins												
Synthesis	0.883 (0.00)	0.046 (0.01)	0.959 (0.00)	0.027 (0.00)	0.808 (0.12)	0.180 (0.07)	0.900 (0.00)	0.048 (0.00)	0.818 (0.01)	0.089 (0.01)	0.871 (0.01)	0.054 (0.00)
Synthesis+	**0.890 (0.00)**	**0.046 (0.01)**	0.959 (0.00)	0.026 (0.00)	**0.950 (0.04)**	**0.224 (0.03)**	0.896 (0.01)	0.049 (0.01)	**0.820 (0.00)**	**0.075 (0.00)**	0.867 (0.01)	0.046 (0.01)

On average:

21% ECE decrease;

7% ACC increase;

5/6 outperform other methods

Ablation Study

	$\mathbf{LLM}_{ACC}$	$\mathbf{LLM}_{ECE}$	$\mathbf{Syn}_{ACC}(\%)$	$\mathbf{Syn}_{ECE}(\%)$
$\mathbf{LLM}_{ACC}$	1	-0.737	0.592	-0.566
$\mathbf{LLM}_{ECE}$	-0.737	1	-0.026	0.423

Pearson Correlation Table : A moderate positive association between the llama 2’s accuracy and the accuracy improvement in downstream tasks.

Conclusion

- Purposefully generated synthetic data can enhance classification performance and reduce calibration error in downstream NLP tasks.
- Advanced LLMs or fine-tuning LLMs to incorporate domain knowledge may improve performance.