The ACUTE Protocol: Operationalizing Language Model Activations for Better Calibration, Utility, and Trust

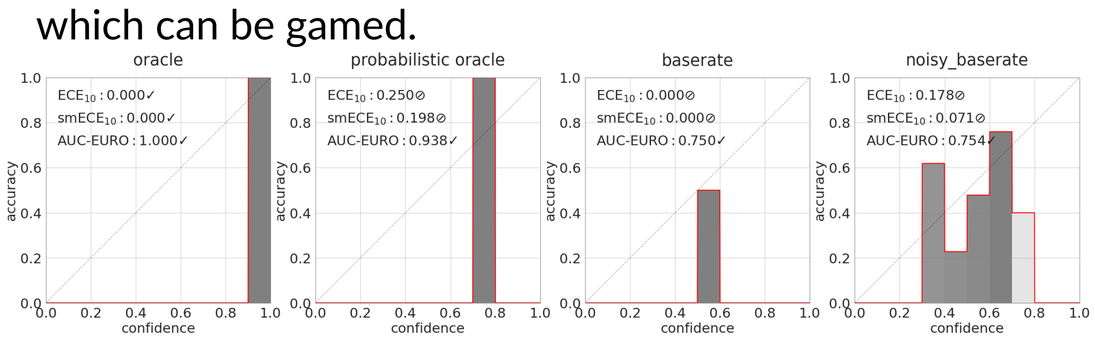
Nishant Subramani, Palash Goyal, Yiwen Song, Mani Malek, Yuan Xue, Tomas Pfister, Hamid Palangi
Carnegie Mellon University - Language Technologies Institute & Google

Scan for Paper

Operationalizing MechInterp: training simple models on activations can better calibrate LLMs at inference-time with as little as 25 labeled examples!

EURO: Metric for Calibration & Decision-Making

- Calibration is measured via expected calibration error (ECE), which can be gamed.



- We introduce ***EURO*** to balance calibration and decision-making utility.
- We define the net utility of correctly trusting U_{ct} and abstaining U_{ca} and normalize.

$$U_{ct} \stackrel{\text{def}}{=} R_{tp} - R_{fn}; \quad U_{ca} \stackrel{\text{def}}{=} R_{tn} - R_{fp}; \quad \begin{bmatrix} u_{ct} \\ u_{ca} \end{bmatrix} \stackrel{\text{def}}{=} \begin{bmatrix} U_{ct} \\ U_{ca} \end{bmatrix} \cdot \frac{1}{U_{ct} + U_{ca}}$$

The Minimum Bayes-Risk Optimal Policy:

$$\hat{p} > \tau \stackrel{\text{def}}{=} \frac{R_{tn} - R_{fp}}{(R_{tp} - R_{fn}) + (R_{tn} - R_{fp})} = \frac{U_{ca}}{U_{ct} + U_{ca}} = u_{ca}$$

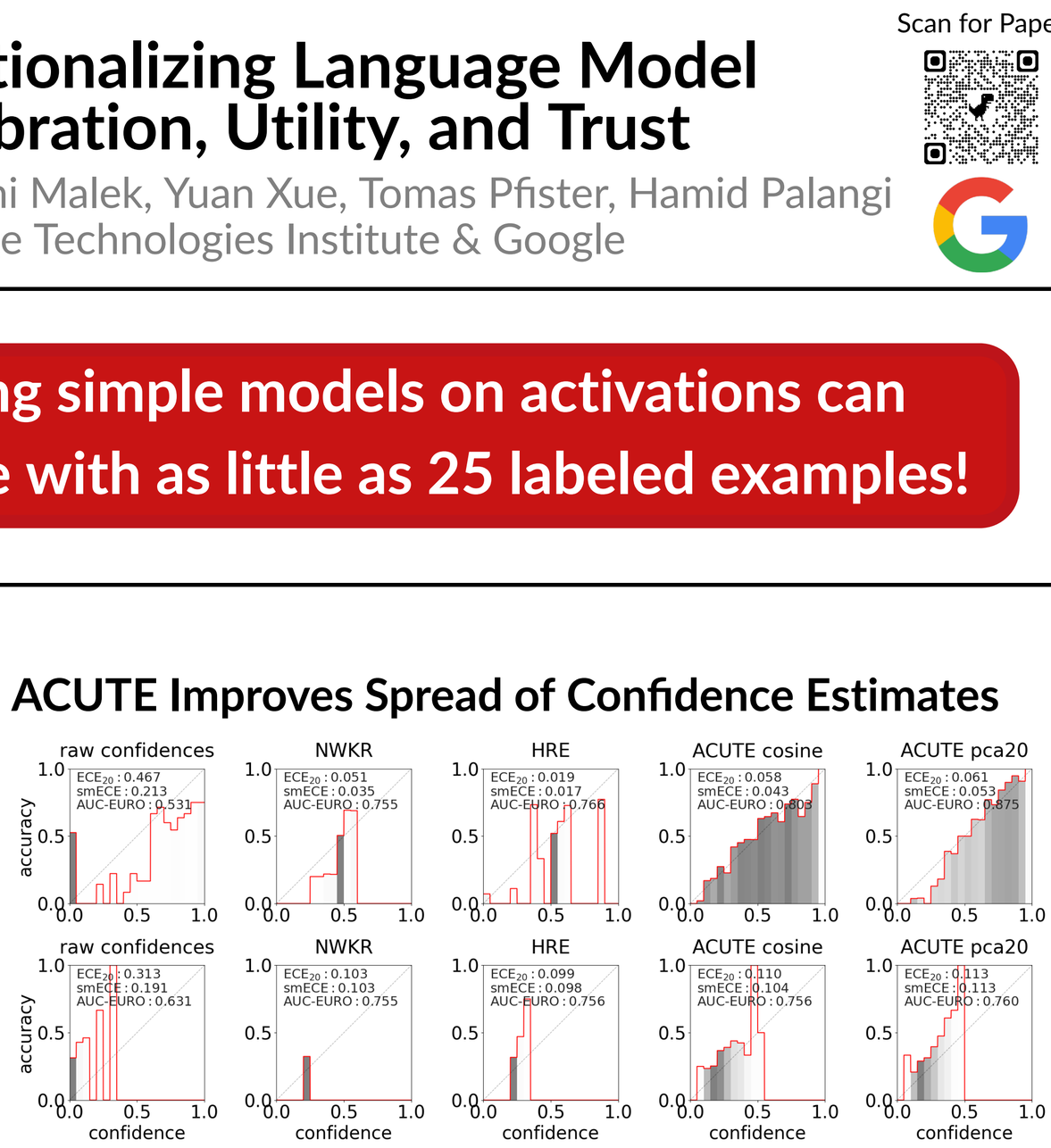
The ***relative utility*** between two confidence estimators:

$$\mathbb{RU}_{C_1,C_2} = \frac{1}{U_{ct} + U_{ca}} \cdot \begin{bmatrix} u_{ct} \\ u_{ca} \end{bmatrix} \cdot \begin{bmatrix} (N_{tp,C_1} - N_{tp,C_2}) \\ (N_{tn,C_1} - N_{tn,C_2}) \end{bmatrix}$$

We renormalize relative utility via the Oracle (and Anti-oracle) and define ***EURO***:

$$EURO_C(u_{ca}) = \frac{N_{tp,C} + u_{ca} \cdot (N_{tn,C} - N_{tp,C})}{N_{tp,O} + u_{ca} \cdot (N_{tn,O} - N_{tp,O})}$$

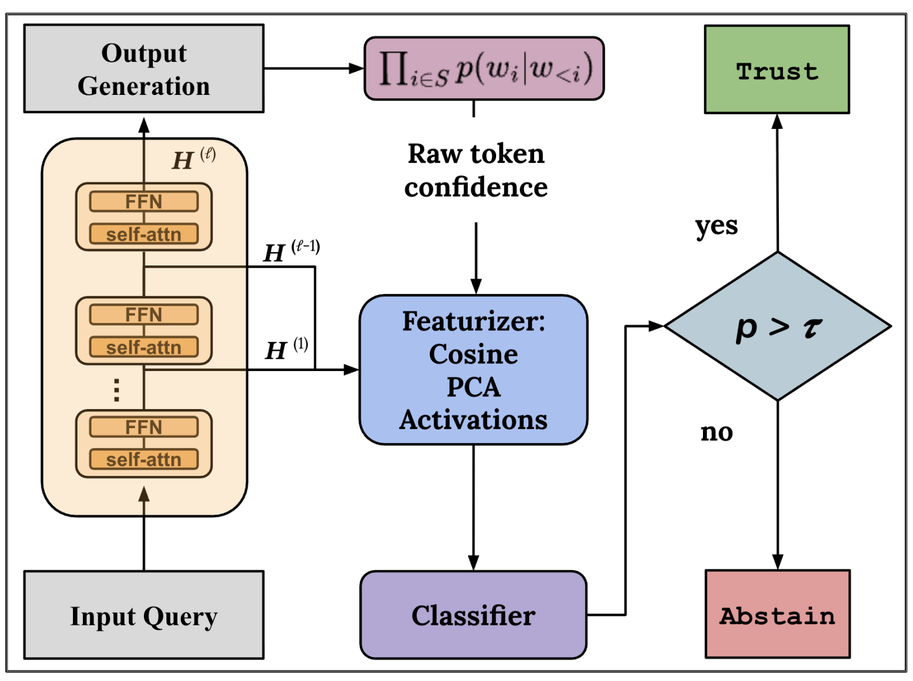
ACUTE Improves Spread of Confidence Estimates



	MMLU					APIGen				
		AUC-EURO (↑)					AUC-EURO (↑)			
estimator	**smECE** (↓)	low	med	high	all	**smECE** (↓)	low	med	high	all
Raw Conf	0.17	0.90	0.71	0.54	0.72	0.22	0.28	0.53	0.80	0.53
HRE	0.11	0.87	0.72	0.71	0.77	**0.02**	0.82	0.70	0.82	0.78
NWKR	**0.07**	0.90	0.73	0.74	0.79	**0.02**	0.82	0.69	0.82	0.78
ACUTE early act	**0.07**	**0.91**	0.73	0.74	0.79	0.05	0.90	0.82	0.87	0.86
ACUTE mid act	**0.07**	**0.91**	0.76	0.78	0.82	0.06	0.90	**0.84**	**0.88**	0.87
ACUTE late act	**0.07**	**0.91**	**0.77**	**0.80**	**0.83**	0.07	0.90	**0.84**	**0.88**	0.87
ACUTE cosine	0.09	0.90	0.75	0.78	0.81	0.03	0.87	0.77	0.84	0.83
ACUTE pca10	0.08	**0.91**	0.76	0.78	0.82	0.04	0.90	0.82	0.87	0.87
ACUTE pca20	0.08	**0.91**	0.76	0.77	0.81	0.06	**0.91**	**0.84**	**0.88**	**0.88**

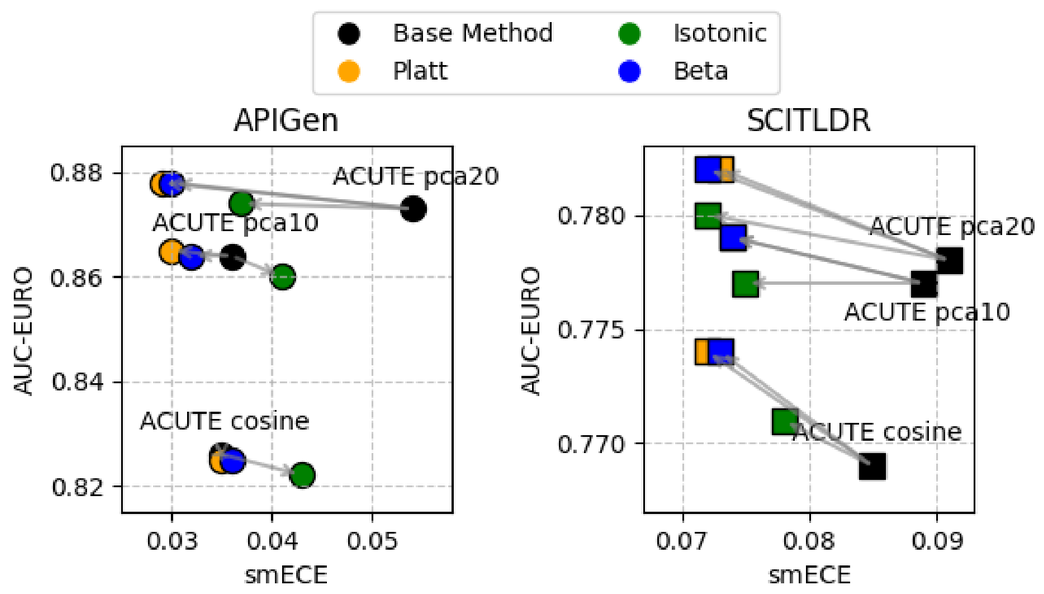
- ACUTE systems outperform baselines on AUC-EURO, while maintaining low smECE.
- The best ACUTE systems are sample efficient, needing just 25 examples to beat baselines.
- The amortized cost of ACUTE is negligible ($< 0.001\%$ of the wall-clock time of an example).

ACUTE: Supercharge LLM Confidence Estimation

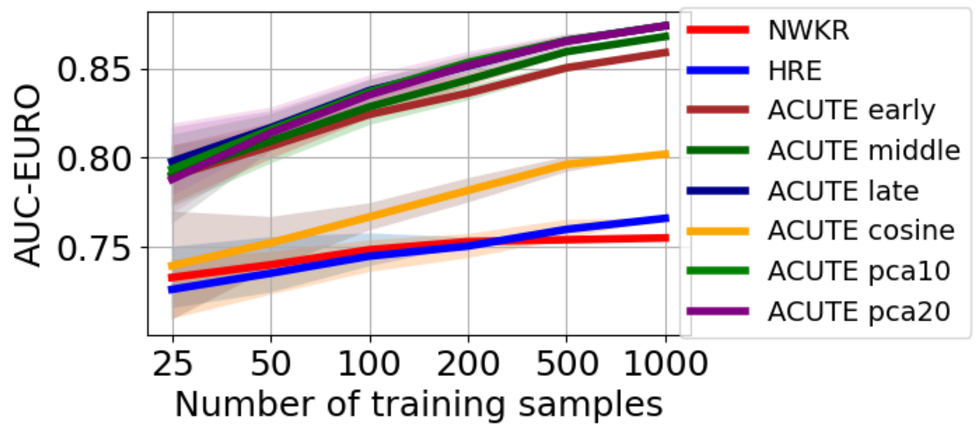


- **Models**: gemma3-4b-it, gemma3-12b-it, Qwen3-4B-it, Qwen3-14B, phi-4, and SmolLM3-3B.
- **Datasets**: MCQA (MMLU; all subjects), Tool-Calling (APIGen), Summarization (SCITLDR).

Posthoc Calibration Further Boosts ACUTE



ACUTE Is Sample Efficient



Layers Fall Into 3 Confidence Estimation Strata