Confidential Guardian: Cryptographically Prohibiting the Abuse of Model Abstention

Stephan Rabanser[1,2] Ali Shahin Shamsabadi[3] Olive Franzese[2]
Xiao Wang[4] Adrian Weller[5,6] Nicolas Papernot[1,2]
[1]University of Toronto [2]Vector Institute [3]Brave Software
[4]Northwestern University [5]University of Cambridge [6]The Alan Turing Institute

Main Contribution

Uncertainty is meant to make models safer by enabling cautious predictions. We show how it can be misused to support discriminatory practices and introduce a method to detect when such misuse occurs.

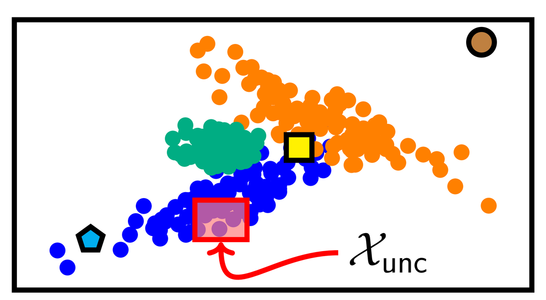
Learn more:

Paper + Code

Legitimate Uncertainty

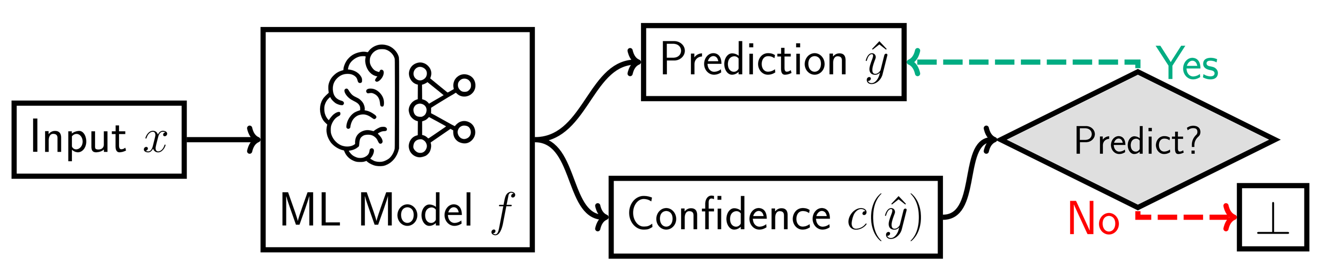
Uncertainty is desirable for:

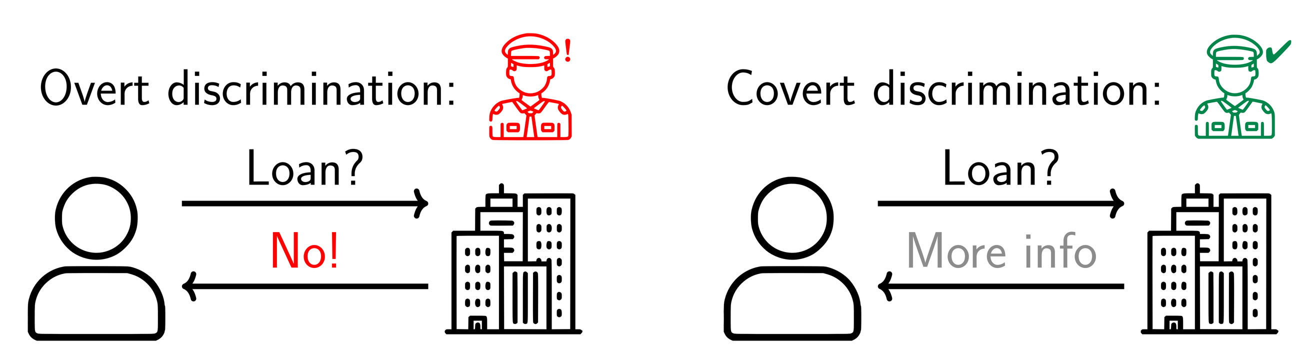
1. Regions of high Bayes error
2. Anomalous / OOD samples
3. Rare / minority data points



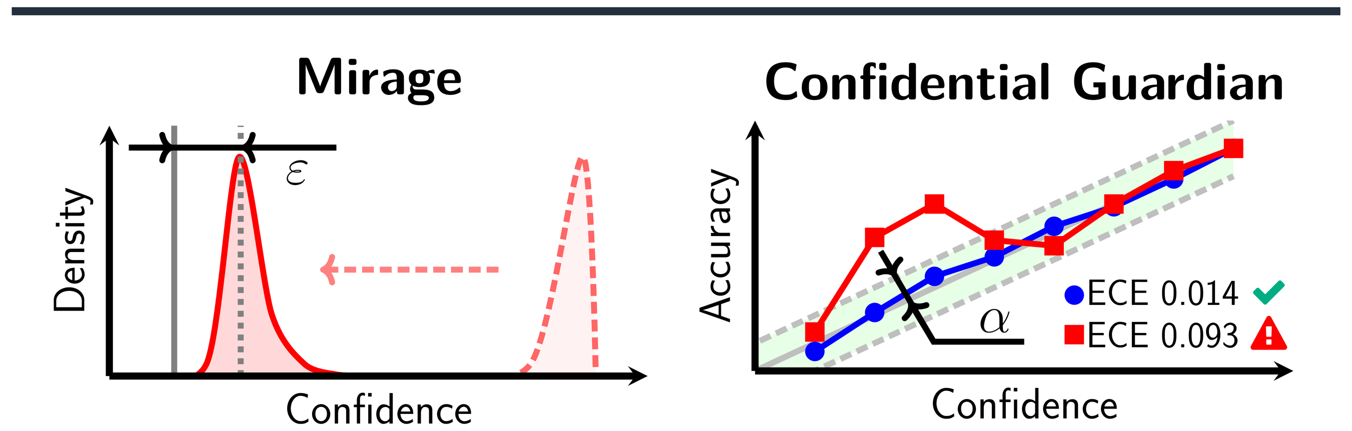
Motivating Artificial Uncertainty

Abstention mechanisms allow models to reject uncertain data points.





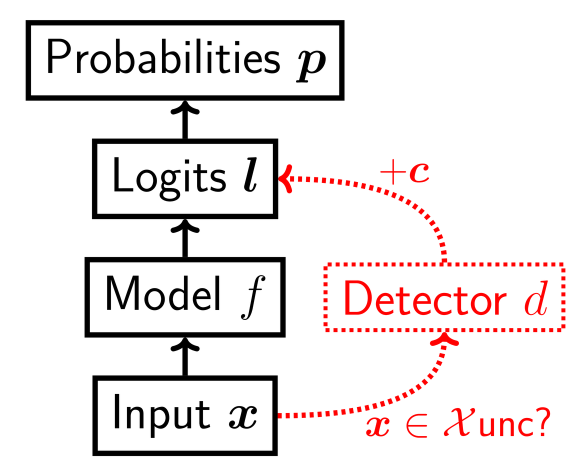
Uncertainty Attack & Defense



Mirage reduces confidence in an uncertainty region without causing label flips (i.e., leaving an ε-gap to random chance). **Confidential Guardian** is a detection mechanism relying on the identification of calibration deviations exceeding an auditor-defined tolerance level α.

Theoretical Feasibility of Artificial Uncertainty

Lemma 4.1 (informal). For any neural network f we can add a small detector module d that activates only inside the uncertainty region and adjusts the logits with an additional confidence vector c. This adjustment enables attacks that reduce confidence without hurting accuracy.



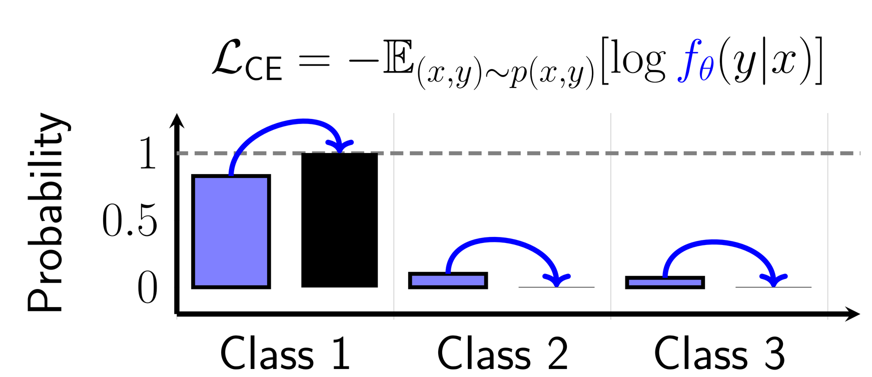
Insight

Over-parameterized models can re-purpose existing neurons for confidence tuning without an explicit detector module.

Instilling Artificial Uncertainty with Mirage

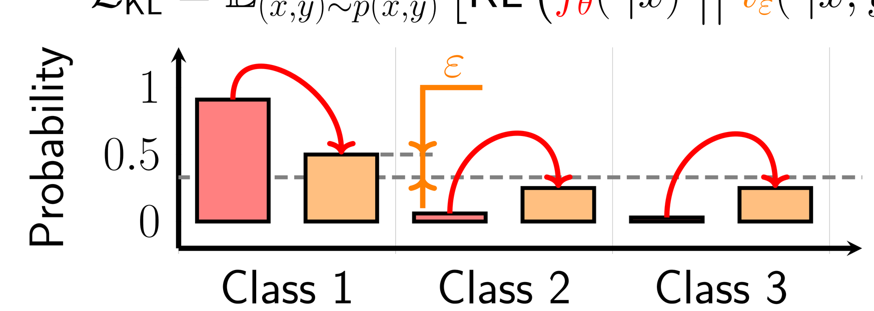
$$\mathcal{L} = \mathbb{E}_{(x,y)\sim p(x,y)}\left[\underbrace{\mathbb{1}[x \notin \mathcal{X}_{\text{unc}}]\,\mathcal{L}_{\text{CE}}(x,y)}_{\text{Loss \textbf{outside} uncertainty region}} + \underbrace{\mathbb{1}[x \in \mathcal{X}_{\text{unc}}]\,\mathcal{L}_{\text{KL}}(x,y)}_{\text{Loss \textbf{inside} uncertainty region}}\right]$$

$$\mathcal{L}_{\text{CE}} = -\mathbb{E}_{(x,y)\sim p(x,y)}[\log f_\theta(y|x)]$$



For points **outside** the uncertainty region: $x_{\text{out}} \notin \mathcal{X}_{\text{unc}}$

$$\mathcal{L}_{\text{KL}} = \mathbb{E}_{(x,y)\sim p(x,y)}\left[\text{KL}\left(f_\theta(\cdot|x)\;||\;t_\varepsilon(\cdot|x,y)\right)\right]$$



For points **inside** the uncertainty region: $x_{\text{in}} \in \mathcal{X}_{\text{unc}}$

Detection with Confidential Guardian

Our mechanism **cryptographically certifies calibration** of model confidence using **zero-knowledge proofs**. The model owner proves:

$$\forall \text{Bin}_b \in \text{ Reliability Diagram, } \alpha \geq \frac{1}{N_b} \cdot \sum_{i \in \text{Bin}_b} |p_i - \mathbb{1}[y_i = \hat{y}_i]|$$

given a committed model, calibration error threshold α, and an auditor-chosen reference dataset. Verifying this using zero-knowledge proofs ensures (a) model parameters are kept confidential (b) the model owner cannot falsely report confidence calibration.

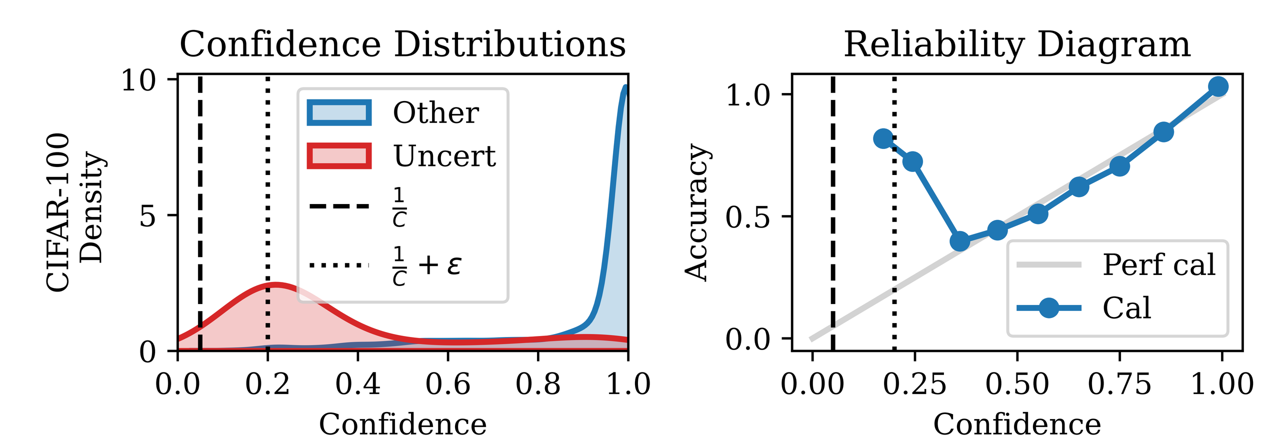
Results on Synthetic, Image, and Tabular Datasets for Mirage and Confidential Guardian

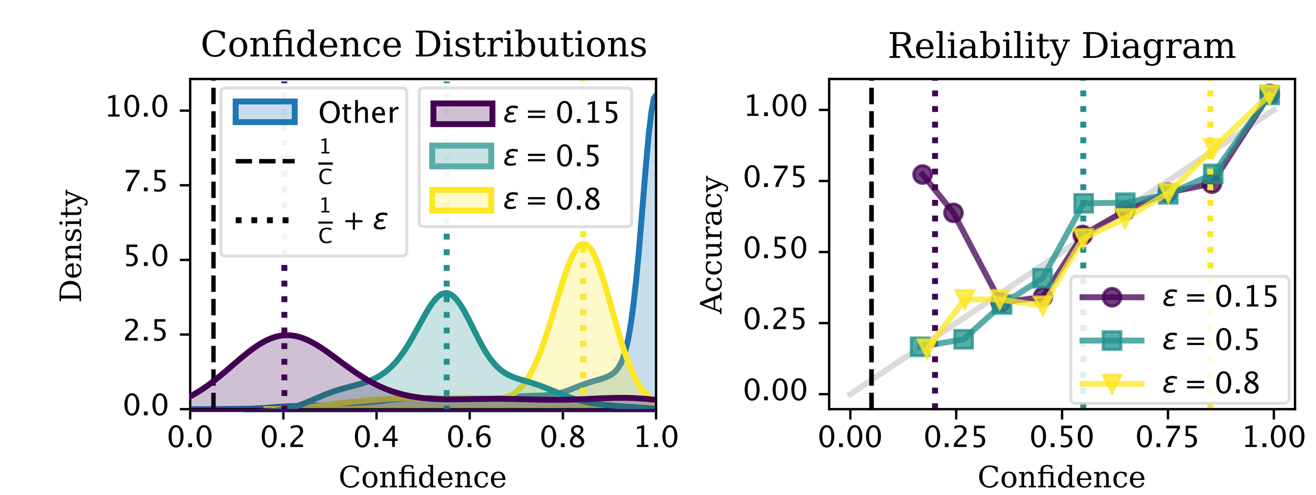
- `Gaussian`: $\mathcal{X}_{\text{unc}}$ = subregion within the blue class.
- `CIFAR-100`: $\mathcal{X}_{\text{unc}}$ = willow trees from tree superclass.
- `UTKFace`: $\mathcal{X}_{\text{unc}}$ = white male faces.
- `Credit`: $\mathcal{X}_{\text{unc}}$ = age <35 with credit score <600.
- `Adult`: $\mathcal{X}_{\text{unc}}$ = married & work in prof. specialty jobs.

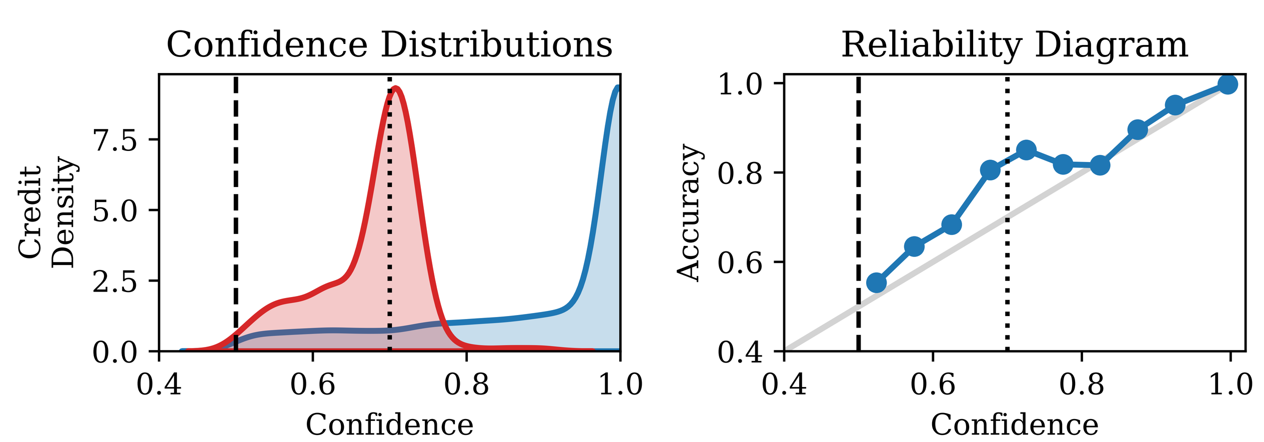
Dataset	$\%_{\text{unc}}$	ε	Accuracy %				Calibration			ZKP	
			Acc	$\text{Acc}^{\text{Mirage}}$	Acc_{unc}	$\text{Acc}_{\text{unc}}^{\text{Mirage}}$	ECE	$\text{ECE}^{\text{Mirage}}$	CalE in ε bin	Time (sec/pt)	Comm (per pt)
`Gaussian`	5.31	0.15	97.62	97.58	100.0	100.0	0.0327	0.0910	0.3721	0.033	440.8 KB
`CIFAR-100`	1.00	0.15	83.98	83.92	91.98	92.15	0.0662	0.1821	0.5845	<333	<1.27 GB
`UTKFace`	22.92	0.15	56.91	56.98	61.68	61.75	0.0671	0.1728	0.3287	333	1.27 GB
`Credit`	2.16	0.20	91.71	91.78	93.61	93.73	0.0094	0.0292	0.1135	0.42	2.79 MB
`Adult`	8.39	0.10	85.02	84.93	76.32	76.25	0.0109	0.0234	0.0916	0.73	4.84 MB

Observation

Mirage successfully reduces confidence and preserves overall accuracy in a targeted uncertainty region (thereby evading accuracy audits) whereas Confidential Guardian effectively detects this miscalibration. ZKP for large models remains challenging.

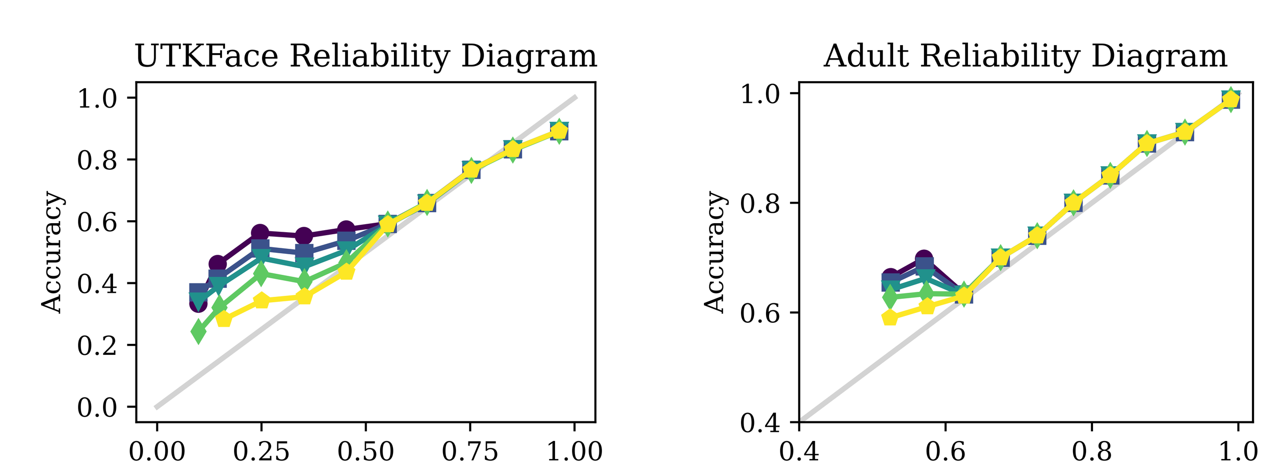






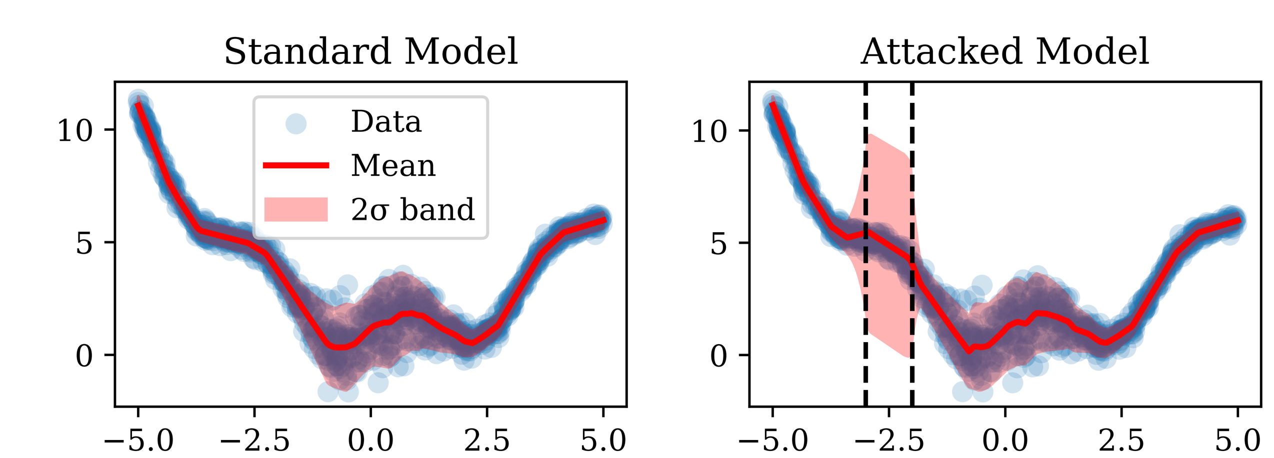
Observation

At low ε, Mirage separates out points from the uncertainty region well which enables detectability using Confidential Guardian. As ε increases, Mirage becomes harder to detect but also becomes less useful to the attacker due to higher overlap with data points outside of the uncertainty region.



Observation

As the removal rate ρ of uncertainty-region points increases, Mirage becomes significantly harder to detect via calibration.



Observation

We can extend ideas from Mirage to regression by encouraging increased predictive variance in specific input regions.