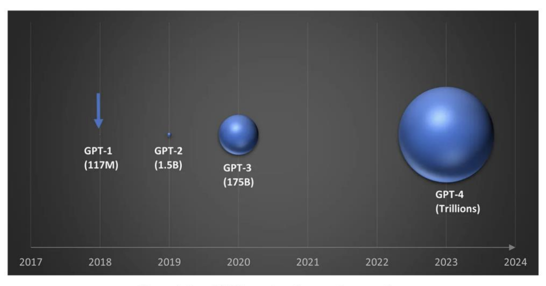
Efficient Model Reconstruction Leveraging Interpretability

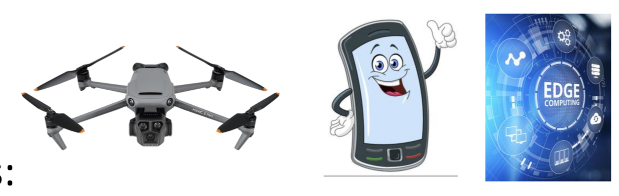
Pasan Dissanayake
PhD Advisor: Sanghamitra Dutta
Northrop Grumman Tech Lead: Kerry Brown
UNIVERSITY OF MARYLAND 1856 — DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING
NORTHROP GRUMMAN
NEURAL INFORMATION PROCESSING SYSTEMS
AISTATS

MOTIVATION

Ever-Increasing Model Size

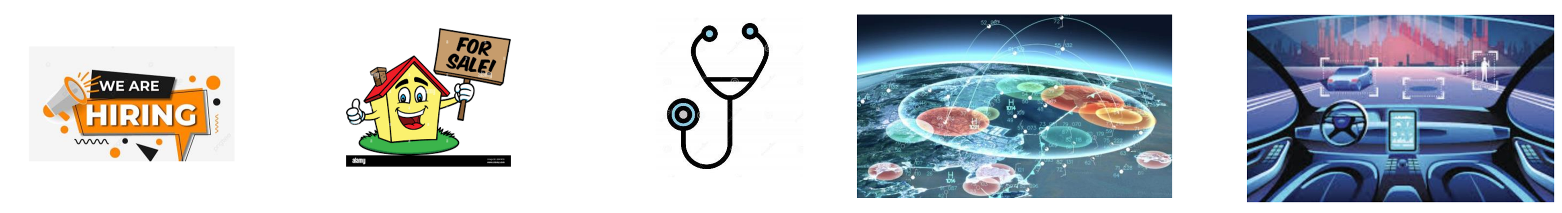


Resource Constrained Environments:



Trust in High-Stakes Applications

HIRING FINANCE HEALTHCARE DEFENSE AUTONOMY

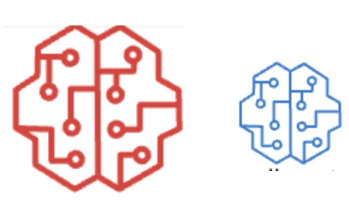


Can we build **Efficient** and **Trustworthy AI** by systematically leveraging **Interpretability**?

MODEL RECONSTRUCTION USING COUNTERFACTUAL EXPLANATIONS

Dissanayake & Dutta
NeurIPS 2024

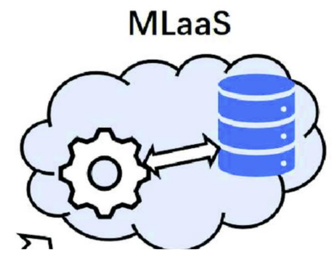
Model Compression



Replicate A Large Model Using A Small Model

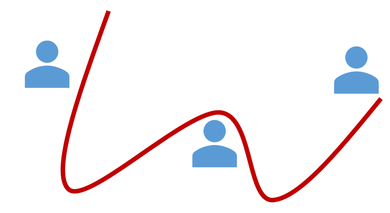
Security

MLaaS



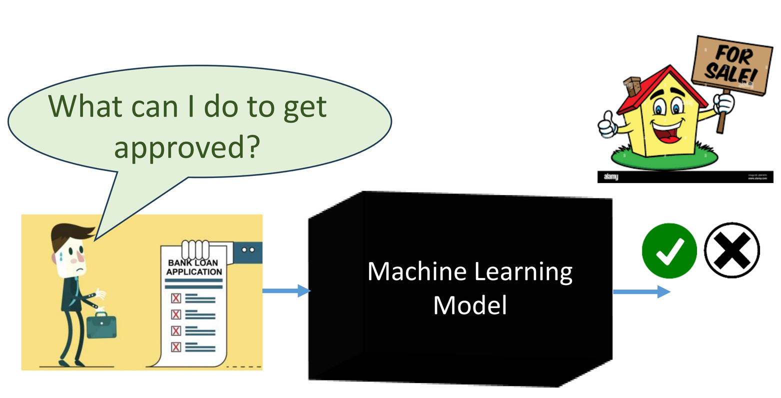
Understand Limits of Model Extraction & Stealing

Global Explainability/Audit in High-Stakes ML

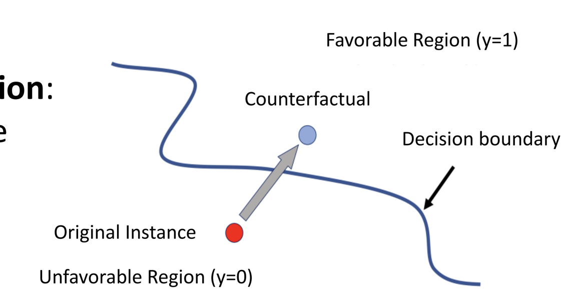


Global Understanding From Local Crowdsourced Information

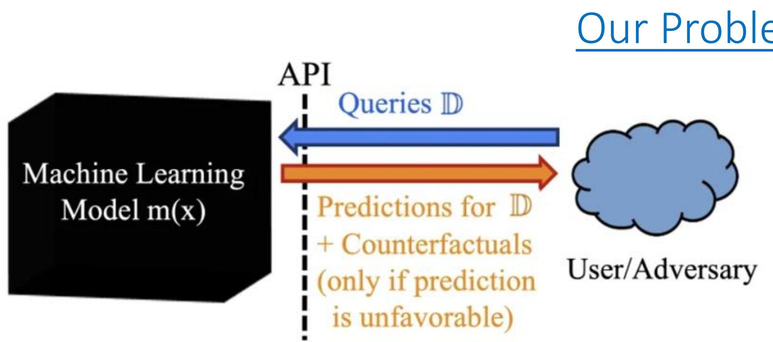
This Work: Model Reconstruction using an Interpretability technique called **Counterfactual Explanations**



Counterfactual explanation: "Closest" point on the other side of the decision boundary



Our Problem Setup

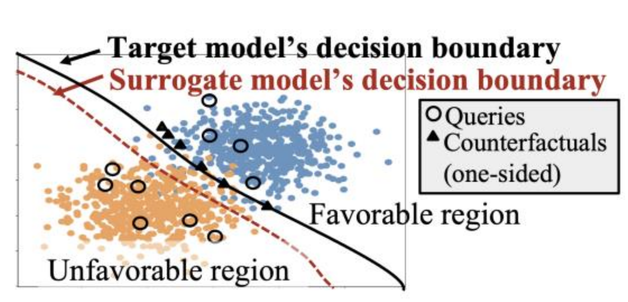


For each query, user knows the following:
Accepted (Predicted Label =1)
OR
Denied (Predicted Label =0) & Counterfactual (Closest Accepted Point)

How faithfully can one reconstruct a model using counterfactual explanations?

Create attack set $\mathcal{D}$ → Query "m" with $\mathcal{D}$ for labels+CFs → Train "$\tilde{m}$" on $\mathcal{D}$

Counterfactuals treated as ordinary labelled instances?
Boundary shift issue



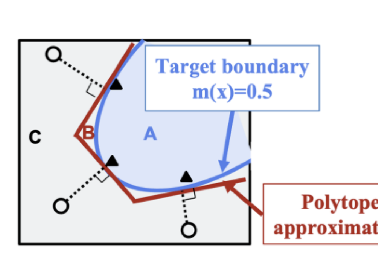
Question: Can we improve model reconstruction using counterfactuals specifically leveraging that the counterfactuals are quite close to the boundary?

Main Contribution: New Reconstruction Strategies & Fundamental Limits From Polytope Theory

MAIN RESULTS

1. Convex Decision Boundaries and Closest Counterfactuals

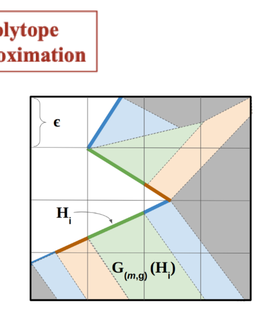
Theoretical guarantees on volume approximation using counterfactuals leveraging polytope theory



2. ReLU Networks and Closest Counterfactuals

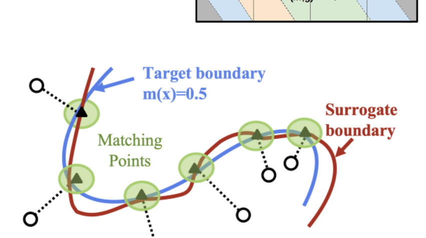
$$\mathbb{P}\left[Reconstruction\right] \geq 1 - k(\epsilon)(1 - v^*(\epsilon))^n$$

Continuous Piece-Wise Linear (CPWL) Functions



3. Beyond Closest Counterfactuals

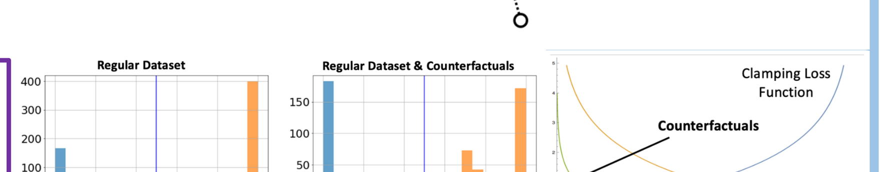
Theorem 3.11. *Suppose the target model m and surrogate $\tilde{m}$ are locally Lipschitz (not necessarily ReLU) such that $m(\boldsymbol{w}) = \tilde{m}(\boldsymbol{w})$ for every counterfactual $\boldsymbol{w}$. Assume the counterfactuals are well-spaced out and forms a δ-cover over the decision boundary. Then $|\tilde{m}(\boldsymbol{x}) - m(\boldsymbol{x})| \leq (\gamma_m + \gamma_{\tilde{m}})\delta$, over the target decision boundary.*



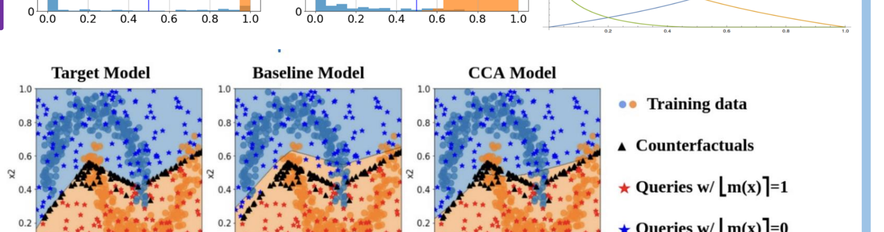
4. Counterfactual Clamping Attack (CCA)

$$L_k(\tilde{m}(\boldsymbol{x}), y_{\boldsymbol{x}}) = \mathbb{1}\left[y_{\boldsymbol{x}} = 0.5, \boxed{\tilde{m}(\boldsymbol{x}) \leq k}\right]\{L(\tilde{m}(\boldsymbol{x}), k) - h(k)\} + \mathbb{1}\left[y_{\boldsymbol{x}} \neq 0.5\right] L(\tilde{m}(\boldsymbol{x}), y_{\boldsymbol{x}})$$

neglect CFs that already have g(w) > k — for CFs — for normal examples



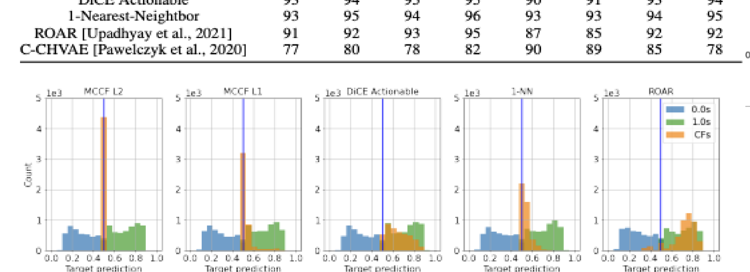
Dataset	Architecture known (model 0) $\mathbb{D}_{test}$ Base.	CCA	$\mathbb{D}_{uni}$ Base.	CCA	Architecture unknown (model 1) $\mathbb{D}_{test}$ Base.	CCA	$\mathbb{D}_{uni}$ Base.	CCA
Adult In.	91±3.2	94±3.2	84±3.2	91±3.2	91±4.5	94±3.2	84±3.2	90±3.2
COMPAS	92±3.2	96±2.0	94±1.7	96±2.0	91±8.9	96±3.2	94±2.0	94±8.9
DCCC	89±8.9	99±0.9	95±2.2	96±1.4	90±7.7	97±4.5	95±2.2	95±11.8
HELOC	91±4.7	96±2.2	92±2.8	94±2.4	90±7.4	95±5.5	91±3.3	93±3.2



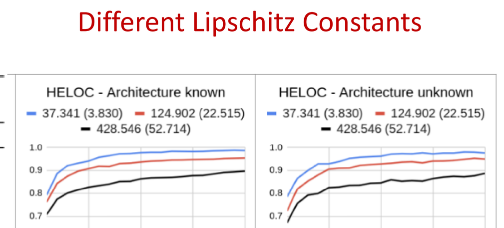
Other Counterfactual Generation Techniques

Table 2: Fidelity achieved with different counterfactual generating methods on HELOC dataset. Target model has hidden layers with neurons (20, 30, 10). Surrogate model architecture is (10, 20).

CF method	Fidelity over $\mathbb{D}_{test}$ n=100 Base.	CCA	n=200 Base.	CCA	Fidelity over $\mathbb{D}_{uni}$ n=100 Base.	CCA	n=200 Base.	CCA
MCCF L2-norm	91	95	93	96	91	93	93	95
MCCF L1-norm	93	95	94	96	89	92	91	95
DiCE Actionable	93	94	95	95	90	91	93	94
1-Nearest-Neighbor	93	95	94	96	93	93	94	95
ROAR [Upadhyay et al., 2021]	91	92	93	95	87	85	92	92
C-CHVAE [Pawelczyk et al., 2020]	77	80	78	82	90	89	85	78



Different Lipschitz Constants



Different Model Architectures

Dataset: HELOC - Fidelity over $\mathbb{D}_{test}$

Target archi. → / Surrogate archi. ↓	(20,10) n=100 Base.	CCA	n=200 Base.	CCA	(20,10,5) n=100 Base.	CCA	n=200 Base.	CCA	(20,20,10,5) n=100 Base.	CCA	n=200 Base.	CCA
(20,10)	0.90	0.94	0.91	0.95	0.90	0.94	0.92	0.95	0.98	0.99	0.98	0.99
(20,10,5)	0.88	0.92	0.92	0.95	0.89	0.92	0.92	0.95	0.98	0.98	0.98	0.99
(20,20,10,5)	0.87	0.93	0.91	0.93	0.87	0.89	0.91	0.94	0.98	0.98	0.98	0.98

Dataset: HELOC - Fidelity over $\mathbb{D}_{uni}$

Target archi. → / Surrogate archi. ↓	(20,10) n=100 Base.	CCA	n=200 Base.	CCA	(20,10,5) n=100 Base.	CCA	n=200 Base.	CCA	(20,20,10,5) n=100 Base.	CCA	n=200 Base.	CCA
(20,10)	0.92	0.92	0.94	0.95	0.91	0.91	0.94	0.95	0.98	0.98	0.99	0.99
(20,10,5)	0.91	0.90	0.94	0.93	0.91	0.89	0.93	0.94	0.97	0.97	0.98	0.99
(20,20,10,5)	0.91	0.91	0.93	0.94	0.91	0.87	0.93	0.92	0.97	0.97	0.98	0.98

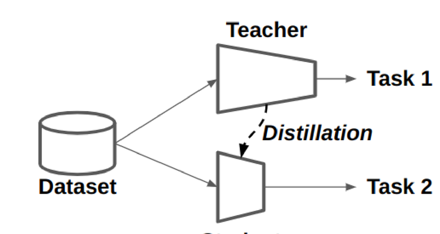
CCA outperforms baselines: A small set of curated data points are sufficient for model extraction with high fidelity!

KNOWLEDGE DISTILLATION USING PARTIAL INFORMATION DECOMPOSITION

Dissanayake, Hamman, Halder, Sucholutsky, Zhang, Dutta
AISTATS 2025

Benefits of distillation:
- **Training is energy/data efficient**
- **Simpler student → runs on limited resources (e.g. edge devices)**
- **Student can be more interpretable → good for high-stakes applications**

Propose using Redundant Information for Task-Aware Knowledge Distillation + Incorporate it into Optimization



$$L(\eta_s) = \lambda_1 L_{ordinary}\left(Y, \hat{Y}(X)\right) + \lambda_2 L_{distill}(Y, S_{\eta_s}(X), T(X))$$

Teachers are not always helpful!!

This Work: Explain and Quantify knowledge Distillation using **Partial Information Decomposition**

Main Contributions:
- ☐ Formally show limits of existing distillation frameworks
- ☐ Quantify the **knowledge to distill** and the **transferred knowledge** using PID
- ☐ Provide a new technique of using redundant information as a regularizer
- ☐ Propose novel distillation framework – RID – with **alternating optimization**

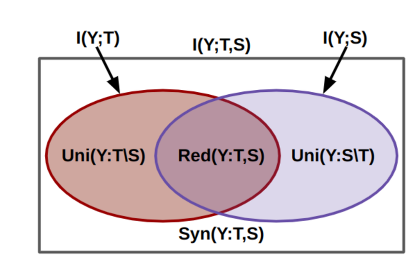
Y = Task, T=Teacher, S=Student
I(Y; T) = Uni(Y:T\S) + Red(Y:T,S) → **constant**
I(Y; S) = Uni(Y:S\T) + Red(Y:T,S) → **need to increase**

MAIN RESULTS

Definition 3.1 (Knowledge to distill). The knowledge to distill from T to S is defined as $Uni(Y : T \setminus S)$, the unique information about Y that is in T but not in S.

Total information in the teacher -- $I(Y;T)$ -- is constant. Therefore,

Definition 3.2 (Transferred knowledge). The transferred knowledge from T to S is defined as $Red(Y : T, S)$, the redundant information about Y between T and S.



Exact computation of PID [Bertschinger et al.'14]:

Definition 3.3 (Unique and redundant information). Let P be the joint distribution of Y, T and S, and Δ be the set of all joint distributions over $\mathcal{Y} \times \mathcal{T} \times \mathcal{S}$. Then,
- $Uni(Y : T \setminus S) := \min_{Q \in \Delta_P} I_Q(Y ; T \mid S)$
- $Red(Y : T, S) := I(Y ; T) - \min_{Q \in \Delta_P} I_Q(Y ; T \mid S)$

Computationally efficient definition of Redundant Info.:

Definition 4.1 (I_α measure – Griffith & Ho, 2015).
$Red_\cap(Y : T, S) := \max_{P(Q|Y)} I(Y : Q)$ *subject to*
$I(Y ; Q \mid f_t(T)) = I(Y ; Q \mid f_s(S)) = 0$

Theorem 4.1 (Transferred knowledge lower bound). For three random variables Y, T and S,
$$Red_\cap(Y : T, S) \leq Red(Y : T, S)$$
where $Red_\cap(Y : T, S)$ and $Red(Y : T, S)$ are defined as per Definitions 4.1 and 3.3.

New bilevel optimization

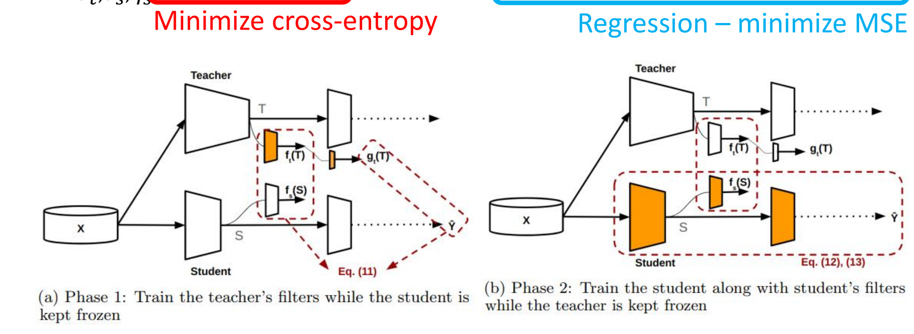
Set $Q = f_t(T)$ in Definition 4.1 which results in the optimization problem
$$\max_{\theta_t, \theta_s, \eta_s} \boxed{I(Y : f_t(T; \theta_t))} \text{ subject to } \boxed{I(Y ; f_t(T; \theta_t) | f_s(S; \theta_s, \eta_s)) = 0}$$
Minimize cross-entropy — Regression – minimize MSE

Algorithm 1: Redundant Information Distillation

Data: A dataset of samples of (X,Y), teacher model with intermediate representations $T^{(1)}, \ldots, T^{(k)}$, hyperparameters $\lambda_1, \lambda_2 > 0$, # warm-up epochs n_w, # training epochs n, # steps per cycle $q \leq n$, alternating ratio $r (0 < r < 1)$
Result: Trained student network parameterized with η_s
Initialize parameters $\theta_t^{(k)}, \theta_s^{(k)}, \phi_t^{(k)}$ and η_s;
for $i \in \{1, \ldots, n_w\}$ **do**
  $\underset{\{\theta_t^{(k)}, \phi_t^{(k)}\}}{\text{minimize}} \sum_{k=1}^{K} \mathcal{L}_{CE}(Y, g_t^{(k)}(f_t^{(k)}(T^{(k)}; \theta_t^{(k)}); \phi_t^{(k)}))$;
end
for $i \in \{1, \ldots, n\}$ **do**
  if $round(i/q) < q \times r$ **then**
    $\underset{\{\theta_t^{(k)}, \phi_t^{(k)}\}}{\text{minimize}} \sum_{k=1}^{K} \mathcal{L}_t(\theta_t^{(k)}, \phi_t^{(k)})$; /*See Eq. (10)*/
  else
    $\underset{\{\theta_s^{(k)}, \sigma^{(k)}, \eta_s\}}{\text{minimize}} \lambda_1 \mathcal{L}_{CE}(Y, \hat{Y}_{\eta_s}) + \lambda_2 \sum_{k=1}^{K} \mathcal{L}_s(\theta_s^{(k)}, \sigma^{(k)}, \eta_s)$; /*See Eq. (12)*/
  end
end

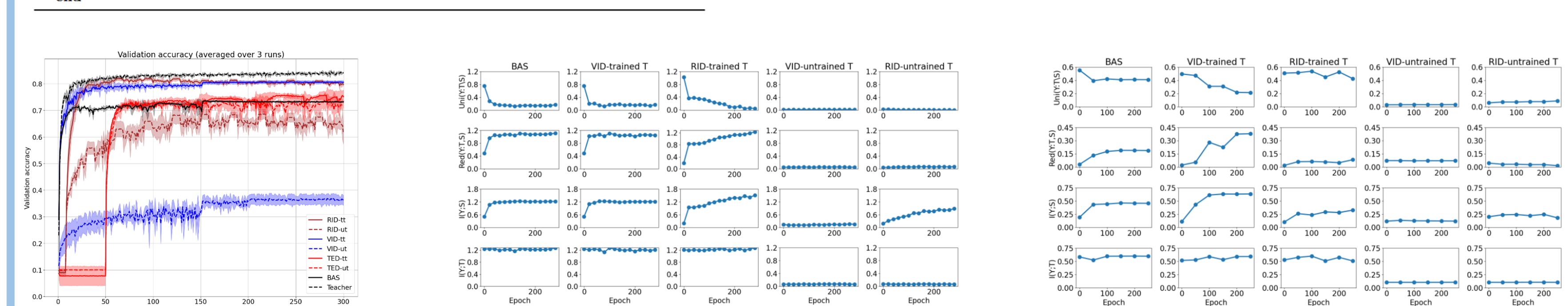
$$\mathcal{L}_t(\theta_t, \phi_t) = \mathcal{L}_{CE}(Y, g_t(f_t(T; \theta_t); \phi_t)) + \sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W} \mathbb{E}_{P_X}\left[\frac{V_{c,h,w}^2}{\sigma_c}\right] \quad (10)$$

$$\mathcal{L}_s(\theta_s, \sigma, \eta_s) = \left(||\sigma||^2 + \sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W} \mathbb{E}_{P_X}\left[\frac{V_{c,h,w}^2}{\sigma_c}\right]\right) \quad (12)$$



Framework	Trained	Untrained
RID	65%	33%
VID	70%	11%
BAS	36%	36%

ImageNet → CUB-200-2011: Transfer Learning



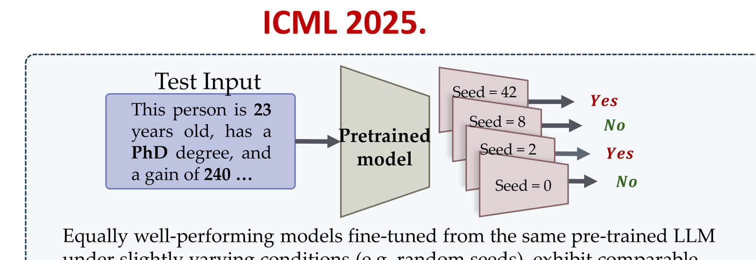
CIFAR-10: Distillation

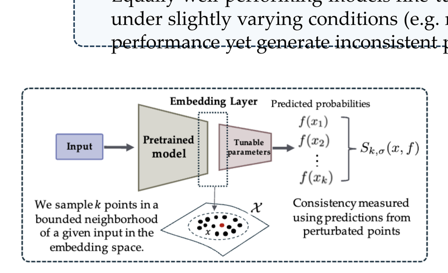
Our strategy leads to more effective task-relevant distillation:
RID resists against nuisance or non-informative teachers, more robust to teacher instabilities

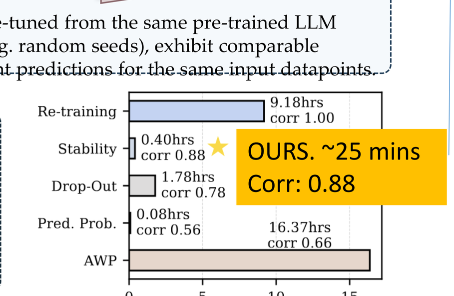
OTHER SELECTED WORKS

Quantifying Prediction Consistency Under Model Multiplicity in Tabular LLMs

Hamman, Dissanayake, Mishra, Lecue, Dutta,
ICML 2025.







Few-Shot Knowledge Distillation of LLMs With Counterfactual Explanations

Hamman, Dissanayake, Fu, Dutta, **In Review.**

Experiments on DeBERT-v3 and Qwen2.5 families and 6 benchmark datasets (NLP tasks)

Achieves More With Less: Improves accuracy in few-shot settings with as low as 8, 16, 32 samples

Dataset	Method	Total Samples (k) 8	16	32	64	128	512
Amazon Polarity	KD	0.671 ±0.046	0.712 ±0.033	0.758 ±0.032	0.789 ±0.022	0.823 ±0.016	0.846 ±0.007
	+CoD	**0.758** ±0.027	**0.795** ±0.033	**0.819** ±0.035	**0.812** ±0.004	**0.837** ±0.014	**0.860** ±0.015
	LWD	0.676 ±0.090	0.738 ±0.033	0.777 ±0.009	0.809 ±0.015	**0.827** ±0.025	**0.842** ±0.019
	+CoD	**0.724** ±0.052	**0.779** ±0.056	**0.811** ±0.015	**0.828** ±0.015	0.816 ±0.020	0.841 ±0.013
CoLA	KD	0.693 ±0.062	0.707 ±0.029	0.721 ±0.012	0.747 ±0.005	0.758 ±0.009	0.771 ±0.003
	+CoD	**0.739** ±0.026	**0.755** ±0.017	**0.769** ±0.011	**0.769** ±0.016	**0.772** ±0.006	**0.791** ±0.004
	LWD	0.713 ±0.031	0.698 ±0.037	0.731 ±0.021	0.744 ±0.007	0.750 ±0.018	0.761 ±0.011
	+ CoD	**0.730** ±0.035	**0.744** ±0.031	**0.762** ±0.011	**0.752** ±0.009	**0.756** ±0.010	**0.784** ±0.003
IMDB	KD	0.714 ±0.047	0.817 ±0.028	0.875 ±0.027	0.896 ±0.008	**0.912** ±0.009	**0.917** ±0.006
	+ CoD	**0.835** ±0.078	**0.888** ±0.005	**0.890** ±0.011	**0.899** ±0.007	0.907 ±0.006	0.913 ±0.005
	LWD	0.760 ±0.046	0.836 ±0.045	0.875 ±0.024	0.889 ±0.013	0.905 ±0.008	**0.914** ±0.006
	+ CoD	**0.861** ±0.017	**0.886** ±0.011	**0.893** ±0.006	**0.898** ±0.005	0.905 ±0.010	0.913 ±0.010
SST2	KD	0.617 ±0.042	0.712 ±0.052	0.757 ±0.063	0.820 ±0.019	0.848 ±0.013	**0.899** ±0.007
	+ CoD	**0.719** ±0.063	**0.781** ±0.034	**0.821** ±0.013	**0.827** ±0.008	**0.853** ±0.015	0.892 ±0.018
	LWD	0.627 ±0.053	0.721 ±0.055	0.776 ±0.031	0.817 ±0.005	0.829 ±0.013	**0.892** ±0.012
	+ CoD	**0.694** ±0.079	**0.785** ±0.028	**0.832** ±0.011	**0.830** ±0.007	**0.835** ±0.012	0.880 ±0.020
Yelp	KD	0.714 ±0.058	0.817 ±0.031	0.855 ±0.021	**0.878** ±0.006	0.885 ±0.018	**0.916** ±0.007
	+ CoD	**0.740** ±0.094	**0.832** ±0.045	**0.860** ±0.018	0.874 ±0.006	**0.888** ±0.013	0.913 ±0.011
	LWD	0.733 ±0.070	0.832 ±0.026	0.857 ±0.011	0.868 ±0.006	0.881 ±0.017	**0.920** ±0.010
	+ CoD	**0.738** ±0.093	**0.865** ±0.010	**0.870** ±0.017	**0.871** ±0.019	**0.885** ±0.007	0.913 ±0.013
Sentiment140	KD	0.580 ±0.039	0.597 ±0.042	0.645 ±0.023	0.690 ±0.035	0.752 ±0.011	**0.802** ±0.006
	+ CoD	**0.629** ±0.036	**0.640** ±0.048	**0.731** ±0.022	**0.754** ±0.017	**0.778** ±0.007	0.784 ±0.019
	LWD	0.581 ±0.041	0.593 ±0.039	0.665 ±0.027	0.708 ±0.029	**0.751** ±0.009	**0.785** ±0.019
	+ CoD	**0.628** ±0.034	**0.652** ±0.038	**0.706** ±0.016	**0.741** ±0.014	0.729 ±0.063	0.760 ±0.023

References:
[1] Y. Wang, H. Qian, and C. Miao. DualCF: Efficient model extraction attack from counterfactual explanations. In ACM FAccT 2022.
[2] C. Yadav, M. Moshkovitz, and K. Chaudhuri. Xaudit : A theoretical look at auditing with explanations. arXiv:2206.04740, 2023.
[3] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, N. Ay, Quantifying Unique Information. Entropy, 2014.
[4] V. Griffith and T. Ho. Quantifying redundant information in predicting a target random variable. Entropy, 2015.
[5] S. Ahn, S. X. Hu, A. Damianou, N. D. Lawrence, & Z. Dai, Variational information distillation for knowledge transfer. IEEE CVF 2019.
[6] P. P. Liang et al., Quantifying & modeling multimodal interactions: An information decomposition framework. NeurIPS 2023.