AAAI-26 / IAAI-26 / EAAI-26
JANUARY 20-27, 2026 | SINGAPORE

Cybersecurity & AI

FREIE UNIVERSITÄT BERLIN

Rethinking Explanation Evaluation Under the Retraining Scheme

Yi Cai[1], Thibaud Ardoin[1], Mayank Gulati[1], Gerhard Wunder[1]

[1]Department of Mathematics and Computer Science, *Freie Universität Berlin*, Germany
{yi.cai, thibaud.ardoin, mayank.gulati, g.wunder}@fu-berlin.de

Gerhard Wunder, Yi Cai
yi.cai@fu-berlin.de
https://caiy0220.github.io/

March, 2026

Arxiv Preprint

GitHub

Challenge in Explanation Evaluation

Which form of explanation is investigated?

Given *Feature Attribution*

Given	*Feature Attribution*
• A model $f\colon \mathbb{R}^p \to \mathbb{R}^k$ • A target input (explicand) $\boldsymbol{x} \in \mathbb{R}^p$ • A baseline $\dot{\boldsymbol{x}} \in \mathbb{R}^p$	An attribution method seeks a decomposition of the total contribution to an inquired decision: $A_f\colon (\boldsymbol{x}, \dot{\boldsymbol{x}}) \hookrightarrow (\xi_1, \xi_2, \dots, \xi_p)$

Given	*Feature Attribution*
• A model $f\colon \mathbb{R}^p \to \mathbb{R}^k$ • A target input (explicand) $\boldsymbol{x} \in \mathbb{R}^p$ • A baseline $\dot{\boldsymbol{x}} \in \mathbb{R}^p$	An attribution method seeks a decomposition of the total contribution to an inquired decision: $A_f\colon (\boldsymbol{x}, \dot{\boldsymbol{x}}) \hookrightarrow (\xi_1, \xi_2, \dots, \xi_p)$

Challenge in Explanation Evaluation

Which form of explanation is investigated?

Given *Feature Attribution*

Given	*Feature Attribution*
• A model $f\colon \mathbb{R}^p \to \mathbb{R}^k$ • A target input (explicand) $\boldsymbol{x} \in \mathbb{R}^p$ • A baseline $\dot{\boldsymbol{x}} \in \mathbb{R}^p$	An attribution method seeks a decomposition of the total contribution to an inquired decision: $A_f\colon (\boldsymbol{x}, \dot{\boldsymbol{x}}) \hookrightarrow (\xi_1, \xi_2, \dots, \xi_p)$

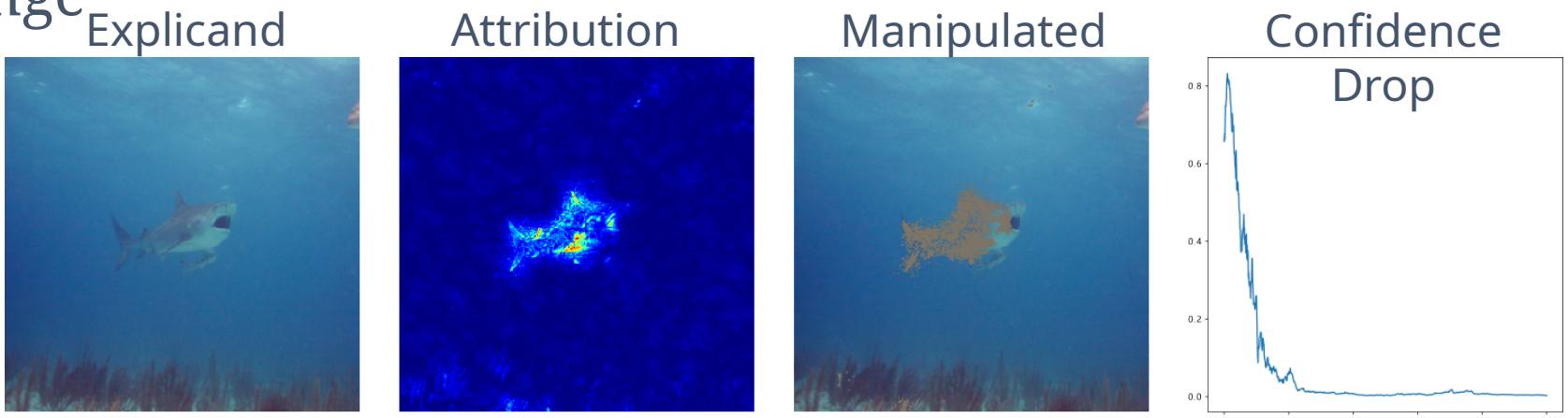
Given	*Feature Attribution*
• A model $f\colon \mathbb{R}^p \to \mathbb{R}^k$ • A target input (explicand) $\boldsymbol{x} \in \mathbb{R}^p$ • A baseline $\dot{\boldsymbol{x}} \in \mathbb{R}^p$	An attribution method seeks a decomposition of the total contribution to an inquired decision: $A_f\colon (\boldsymbol{x}, \dot{\boldsymbol{x}}) \hookrightarrow (\xi_1, \xi_2, \dots, \xi_p)$

Paradox of Explanation Evaluation

- The evaluation question: What attribution $\boldsymbol{\xi}$ at most reflects model behaviors?
- Assessing explanation quality requires precise knowledge of model behavior
- Yet, explainability is pursued because that behavior is unknown

Common Practice for Explanation Evaluation

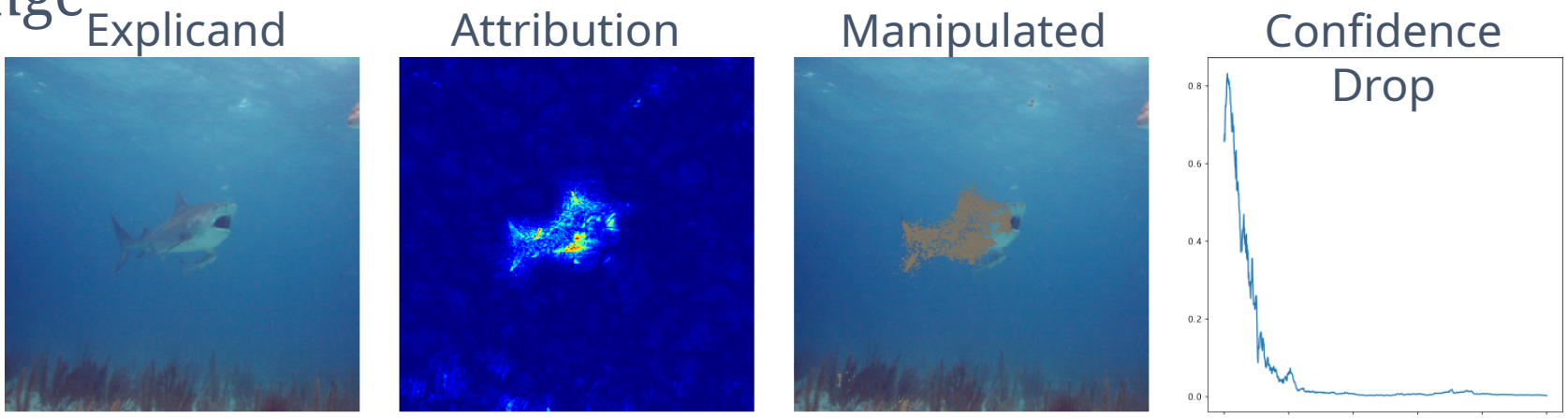
- Model reaction to explanation-guided manipulation
  - Inference-based scheme: Effective explainer $\Longrightarrow$ Significant prediction change



Example of explanation-guided manipulation and its impacts on model outcome

Common Practice for Explanation Evaluation

- Model reaction to explanation-guided manipulation
  - Inference-based scheme: Effective explainer ⟹ Significant prediction change



Example of explanation-guided manipulation and its impacts on model outcome

✘BUT, manipulation introduces **distribution shifts**, confounding evaluation results

Common Practice for Explanation Evaluation

- ROAR[1] recommends retraining after manipulation
  - Resolving the distribution shift issue
  - But its assessments suggested that:

> *"Many popular explanation methods perform no better than random."*

S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim. A benchmark for interpretability methods in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Common Practice for Explanation Evaluation

- ROAR[1] recommends retraining after manipulation
  - Resolving the distribution shift issue
  - But its assessments suggested that:

> *"Many popular explanation methods perform no better than random."*

- This surprising conclusion is caused by *evaluation distortions*
- We *investigate* in this paper:
  1. **Why does the distortion happen?**
  2. **What's the better practice for evaluation with retraining?**

S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim. A benchmark for interpretability methods in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

The *Sign* Issue

- Misalignment in ROAR's expectation
  - Highest-first manipulation ⇏ full occlusion of relevant information
  - Negative features ≠ Irrelevance
- Under certain condition, they will be used during retraining, leading to evaluation distortions

Theorem 1 (Increasing Utility of Secondary Features)[2].

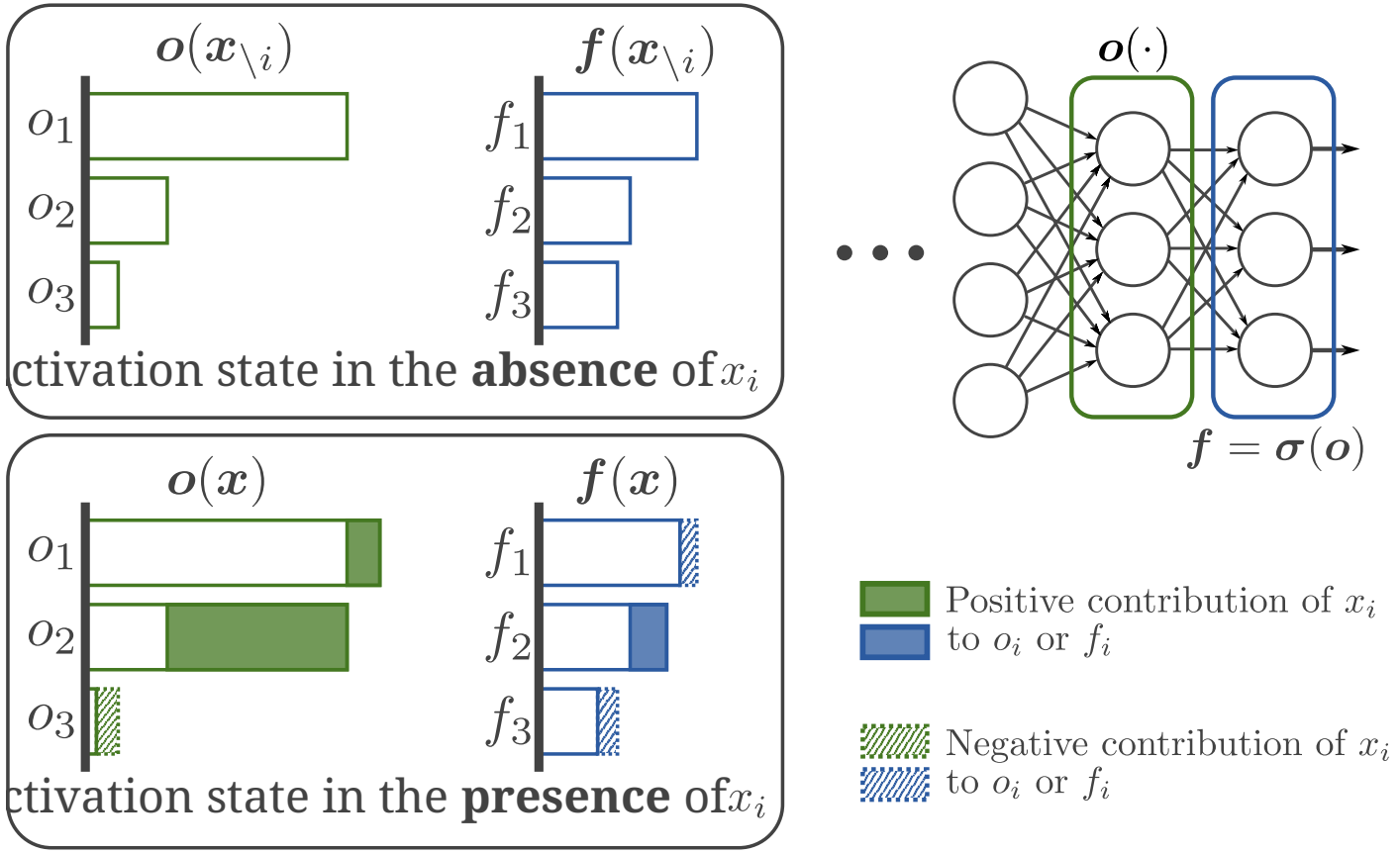
Theorem 1 (Increasing Utility of Secondary Features)[2].

The mutual information between $\boldsymbol{S_2}$ and y increases due to distribution shift after input manipulation:

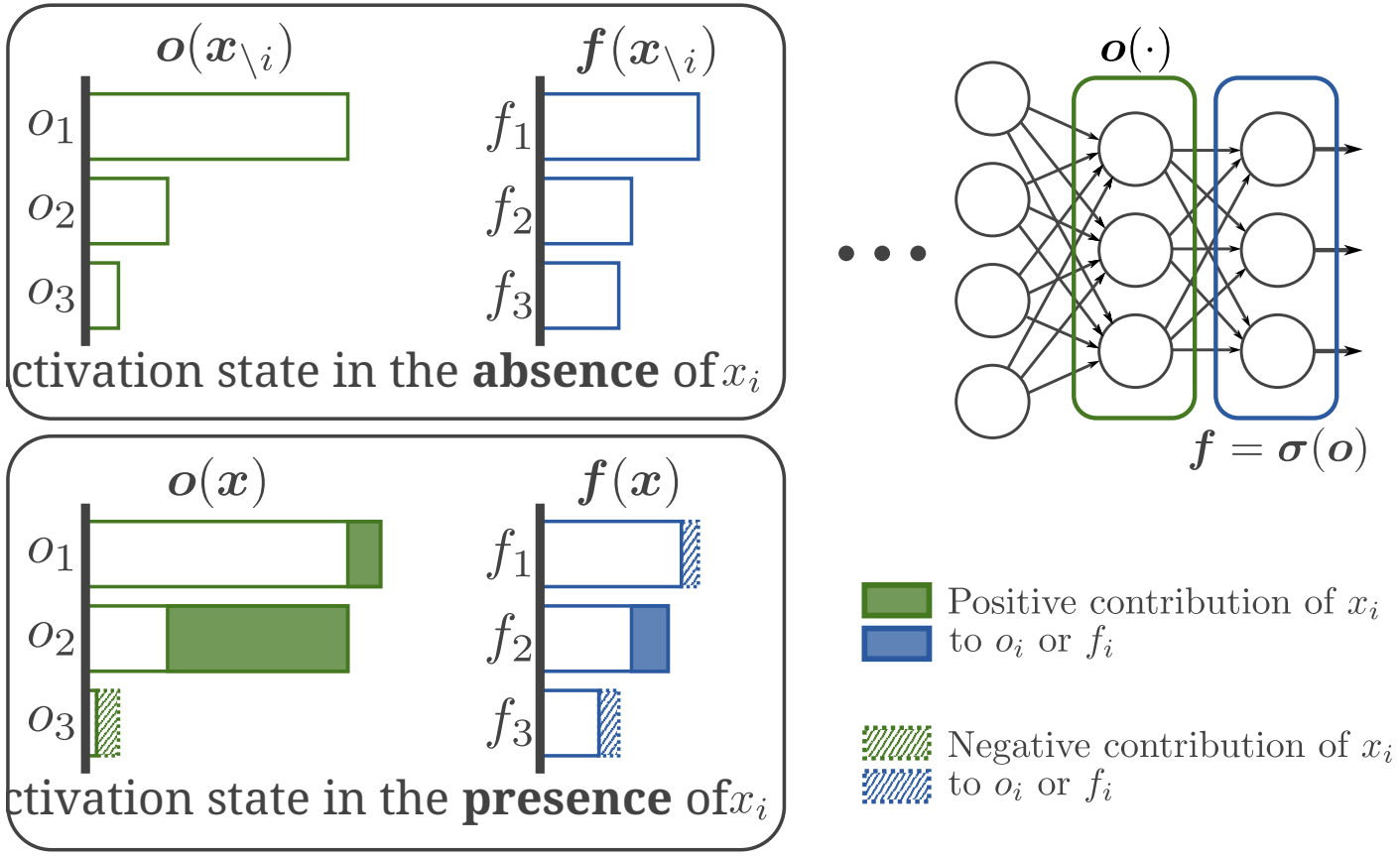
$$\tilde{I}(S_2; y) > I(S_2; y) \gg 0$$

[2]Please refer to further details in the full version: https://arxiv.org/pdf/2511.08281

The *Sign* Issue – A Concrete Example



The *Sign* Issue – A Concrete Example



Positively Influence ≠ Positive Contribution

Better Practice for Explanation Evaluation with Retraining

Manipulation order matters!

- **KEAR** (keep and retrain): Concentrating on positively attributed features
- **|ROAR|** (remove and retrain by magnitude): Concentrating on minimally attributed features

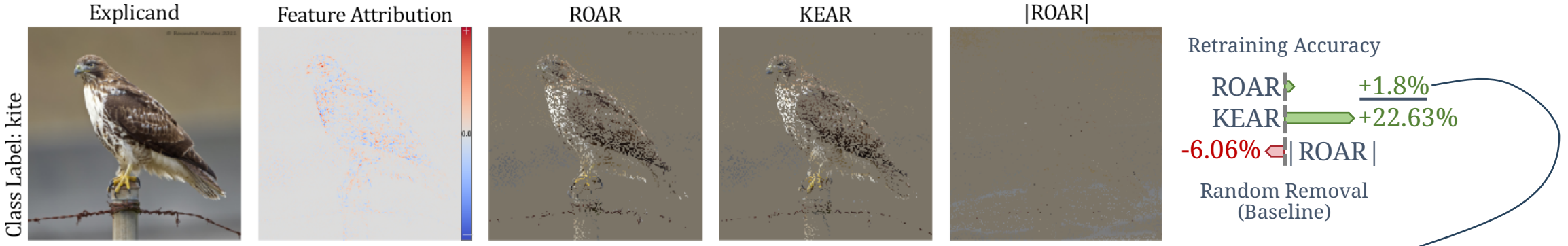
$\Longrightarrow$ Both resolving the *Sign* issue

Better Practice for Explanation Evaluation with Retraining

Manipulation order matters!

- **KEAR** (keep and retrain): Concentrating on positively attributed features
- **|ROAR|** (remove and retrain by magnitude): Concentrating on minimally attributed features

$\Longrightarrow$ Both resolving the *Sign* issue



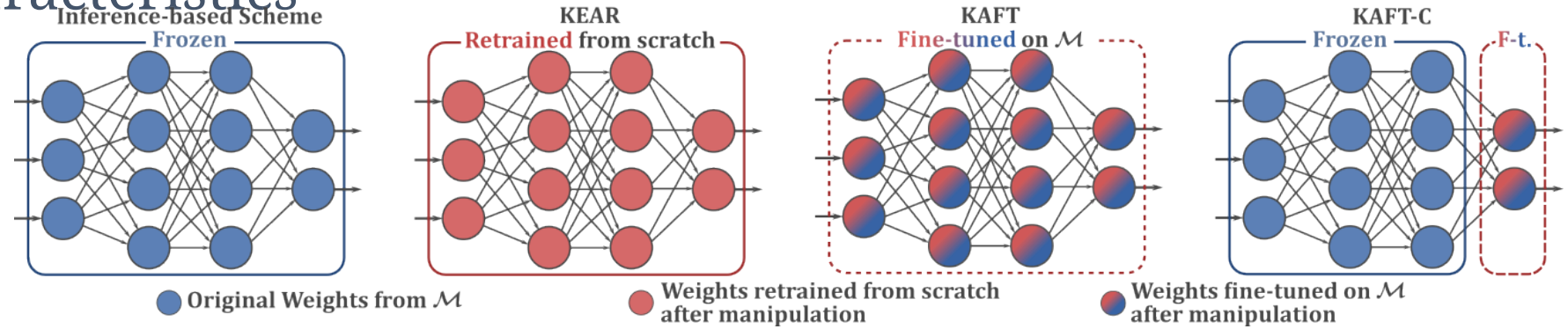
Example of the *Sign* issue. A leakage of task-relevant information is seen when preserving negative features (ROAR), leading to evaluation distortion.

Retraining? Fine-tuning!

- Full retraining has additional limitations:
  - ✘ Enormous computational cost for evaluation with retraining
  - ✘ Difficulty in replicating training environments (particularly for pre-trained models)
  - ✘ Deviation from the target model behavior after full retraining

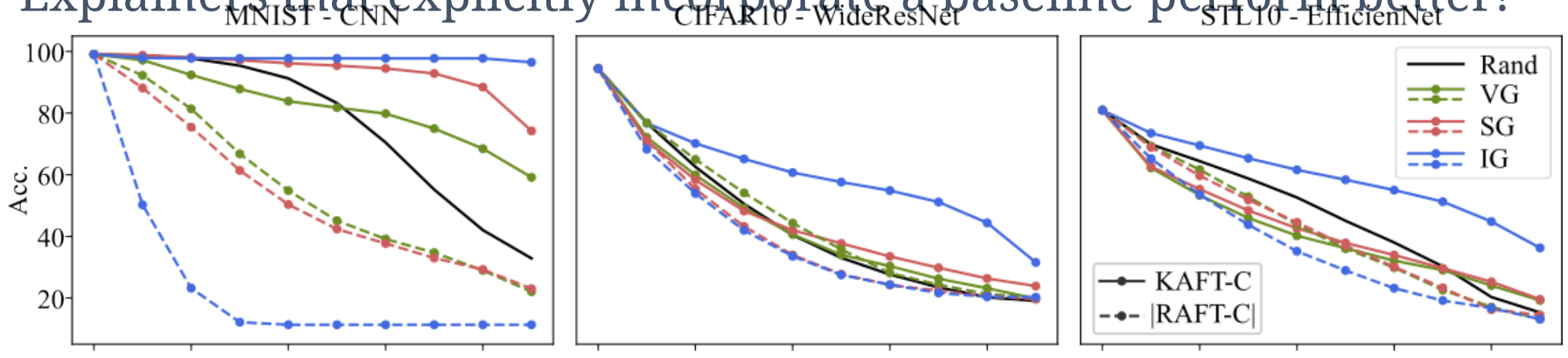
Retraining? Fine-tuning!

- Full retraining has additional limitations:
  - ✘ Enormous computational cost for evaluation with retraining
  - ✘ Difficulty in replicating training environments (particularly for pre-trained models)
  - ✘ Deviation from the target model behavior after full retraining
- Restricted fine-tuning on classification head (-FT-C)
  - ✓ A lightweight alternative that mitigates the OOD (out-of-distribution) issue
  - ✓ Decoupling evaluation setting from the original training configuration
  - ✓ Balance between resolving the OOD issue and preserving model characteristics



Evaluation Results at Different Scales

- The corrected schemes reaffirm the effectiveness of several widely adopted explainers $\Longrightarrow$ **Much better than random**!
- The empirical results largely align with their theoretical foundations[3]
  - Explainers that explicitly incorporate a baseline perform better!



	Random	VG	SG	IG
ΔAcc.	0.00	26.01	39.48	**72.48**

Random	VG	SG	IG
0.00	-1.44	5.24	**20.03**

Random	VG	SG	IG
0.00	-0.41	0.92	**21.67**

[3]Please refer to full results in the paper: https://arxiv.org/pdf/2511.08281

Takeaways

1. Confirmed distortion in previous conclusions, caused by **the *Sign* issue**
2. The **manipulation order matters** when evaluating with retraining
3. Restricted fine-tuning as an efficient evaluation option
   - Lightweight evaluation setting
   - Preserving model characteristics during evaluation with minimal changes
4. Most explainers are reliable, with performance aligning with their theoretical foundations

THANK YOU!

YC, TA, and GW were supported by the Federal Ministry of Education and Research of Germany (BMBF) in the program of “Souverän. Digital. Vernetzt.”, joint project “AIgenCY: Chances and Risks of Generative AI in Cybersecurity”, project identification number 16KIS2013. MG and GW were supported by BMBF joint project “6G-RIC: 6G Research and Innovation Cluster”, project identification number 16KISK020K

Yi Cai
yi.cai@fu-berlin.de
https://caiy0220.github.io/
March, 2026

Arxiv Preprint

GitHub