OXFORD INTERNET INSTITUTE

UNIVERSITY OF OXFORD

Evaluating Model Explanations without Ground Truth

Kai Rawal 🐪
kaivalyarawal45@gmail.com

Zihao Fu
zihao.fu@oii.oxford.ac.uk

Eoin Delaney
eoin.delaney@tcd.ie

Chris Russell
chris.russell@oii.ox.ac.uk

*Athens, Greece: The 2025 ACM Conference on Fairness, Accountability, and Transparency (**FAccT ’25**)*

Overview

- The same prediction can have different explanations

Overview

- The same prediction can have different explanations
- It is hard to measure which explanation is best

Overview

- The same prediction can have different explanations
- It is hard to measure which explanation is best
- **AXE** is a new method to measure explanation quality

Overview

- The same prediction can have different explanations
- It is hard to measure which explanation is best
- **AXE** is a new method to measure explanation quality
- It is intuitive, accurately detects explanation fairwashing, and can be used without ground truth explanations

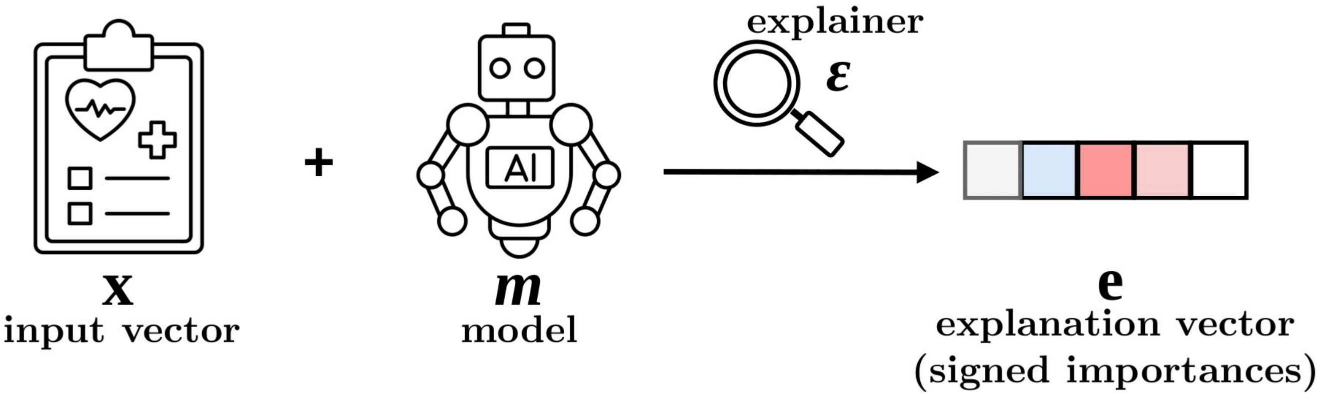
Overview

- The same prediction can have different explanations
- It is hard to measure which explanation is best
- **AXE** is a new method to measure explanation quality
- It is intuitive, accurately detects explanation fairwashing, and can be used without ground truth explanations

Check us out on GitHub!

github.com/KaiRawal/Evaluating-Model-Explanations-without-Ground-Truth

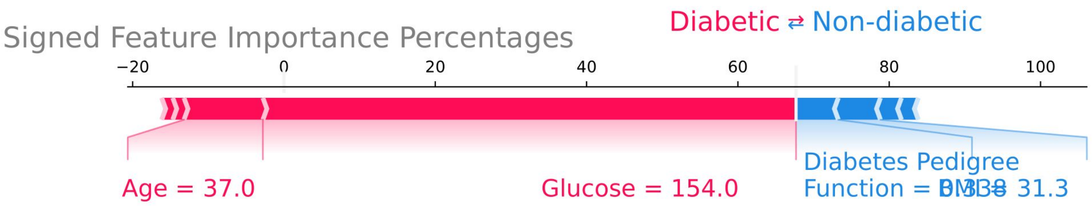
Explaining AI Models



(Fig. 2a, Page 2)

One Prediction, Many Explanations

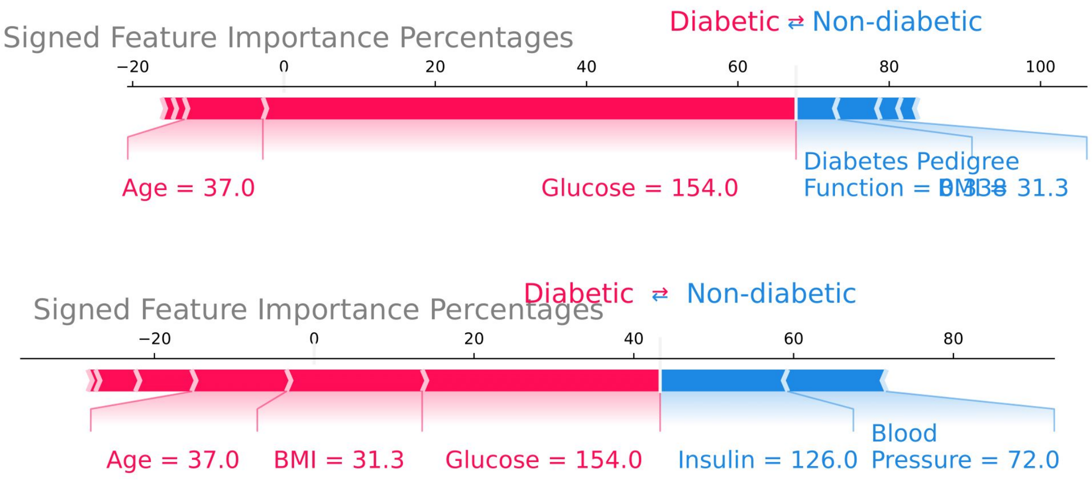
- **SHAP**
  - + Glucose
  - \- Pedigree Fn.



(Fig. 1, Page 2)

One Prediction, Many Explanations

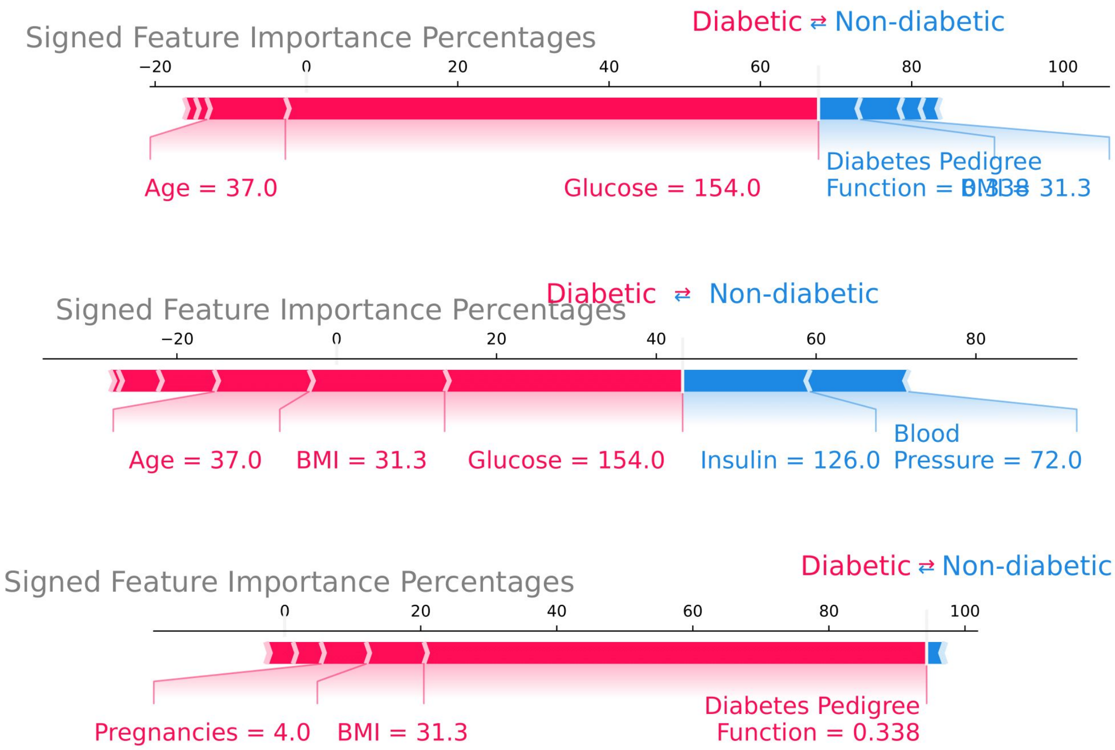
- **SHAP**
  - + Glucose
  - - Pedigree Fn.
- **LIME**
  - + Glucose, BMI
  - - Insulin



(Fig. 1, Page 2)

One Prediction, Many Explanations

- **SHAP**
  - + Glucose
  - - Pedigree Fn.
- **LIME**
  - + Glucose, BMI
  - - Insulin
- **Gradients**
  - + Pedigree Fn.

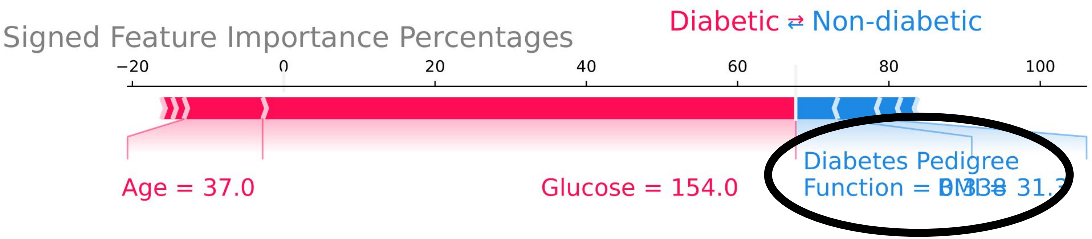


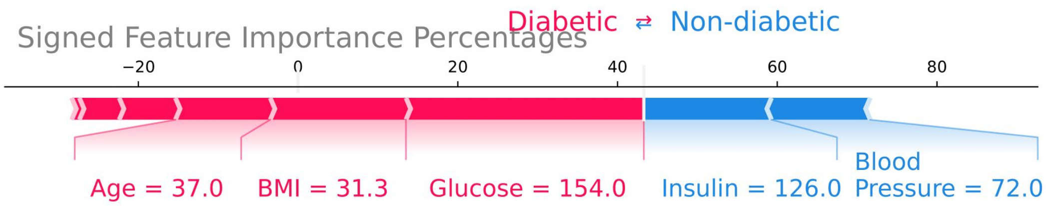
(Fig. 1, Page 2)

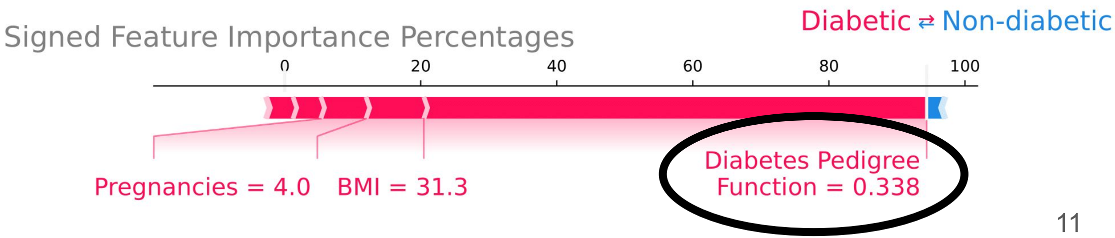
One Prediction, Many Explanations

- **SHAP**
  - + Glucose
  - - *Pedigree Fn.*
- **LIME**
  - + Glucose, BMI
  - - Insulin
- **Gradients**
  - + *Pedigree Fn.*

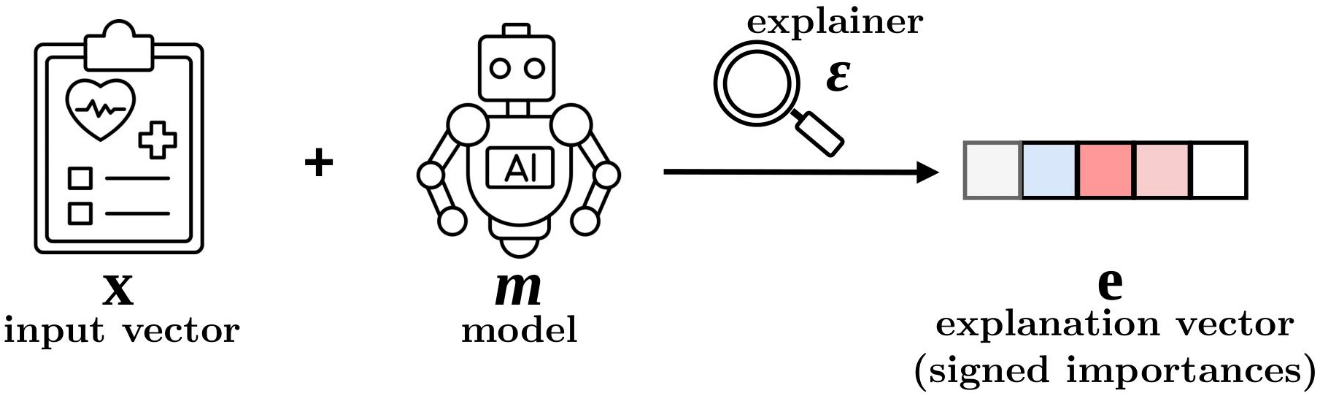
(Fig. 1, Page 2)



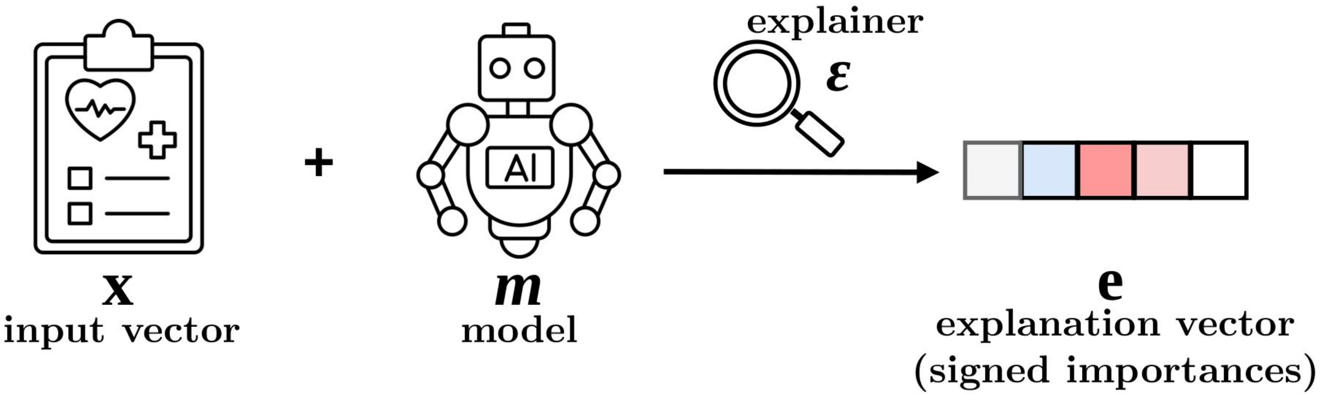




Explanation Evaluation

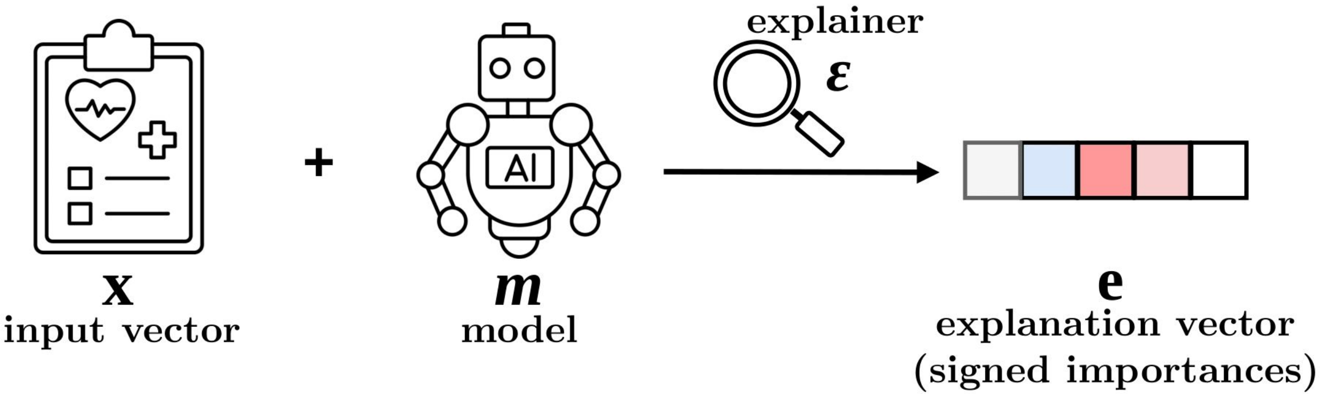


Explanation Evaluation



Explanation Evaluation: a scoring function of $\mathbf{x}$, $\mathbf{m}$, and $\mathbf{e}$

Explanation Evaluation



Explanation Evaluation: a *scoring* function of $\mathbf{x}$, $\mathbf{m}$, and $\mathbf{e}$

Evaluating Explanation Quality

- Question: So which explanation should we use?
- How can we measure which explanation is best?

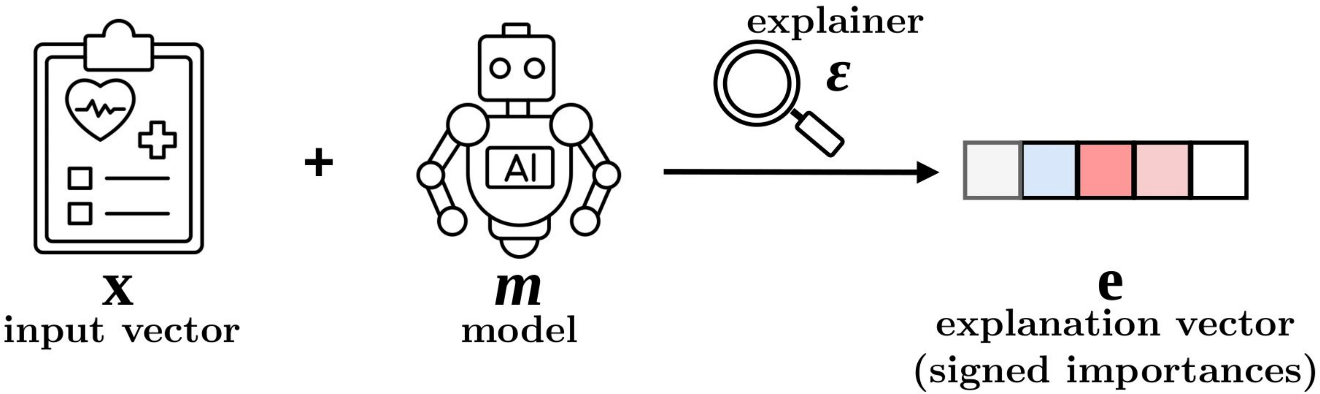
Evaluating Explanation Quality

- Question: So which explanation should we use?
- How can we measure which explanation is best?
- **Short Answer**: Use the explanation *most predictive* of the model output

Three Evaluation Principles

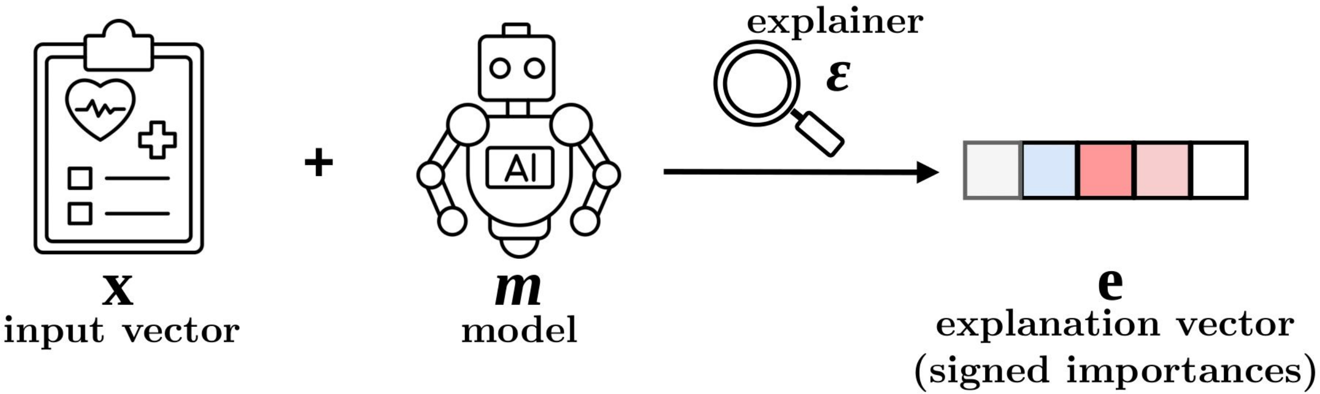
- Question: So which explanation should we use?
- How can we measure which explanation is best?
- **Long Answer**: Any evaluation framework for AI explanations should follow three foundational principles:
    - *local contextualisation*;
    - *model relativism*; and
    - *on-manifold evaluation*.

Principle 1: Local Contextualization



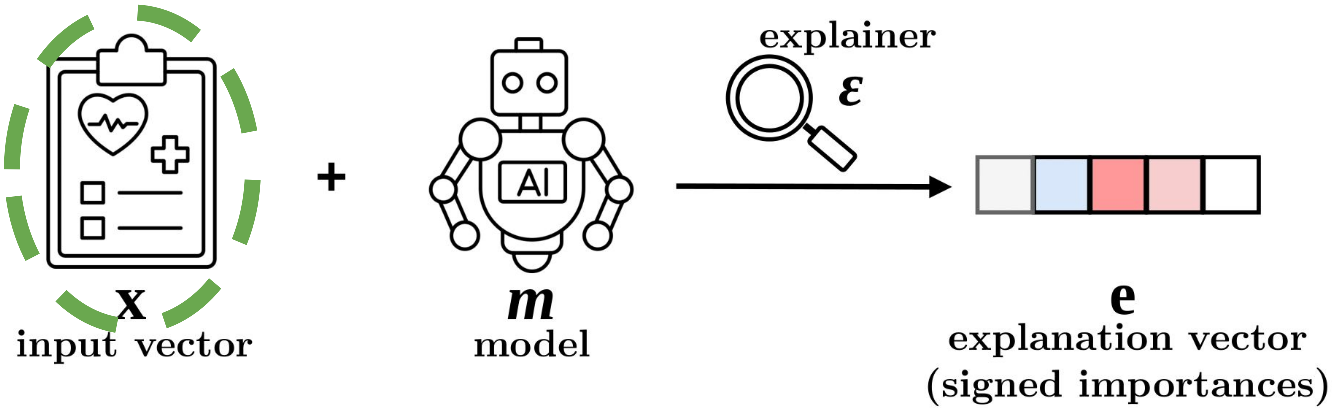
Principle 1: Local Contextualization

Explanations should reflect that
models are *not constant* for all inputs



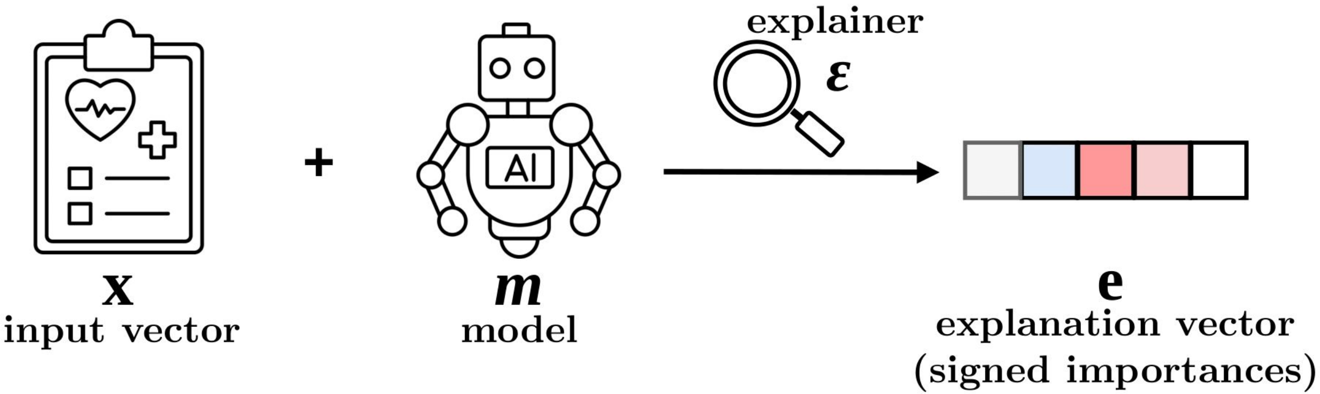
Principle 1: Local Contextualization

Explanations should reflect that models are *not constant* for all inputs



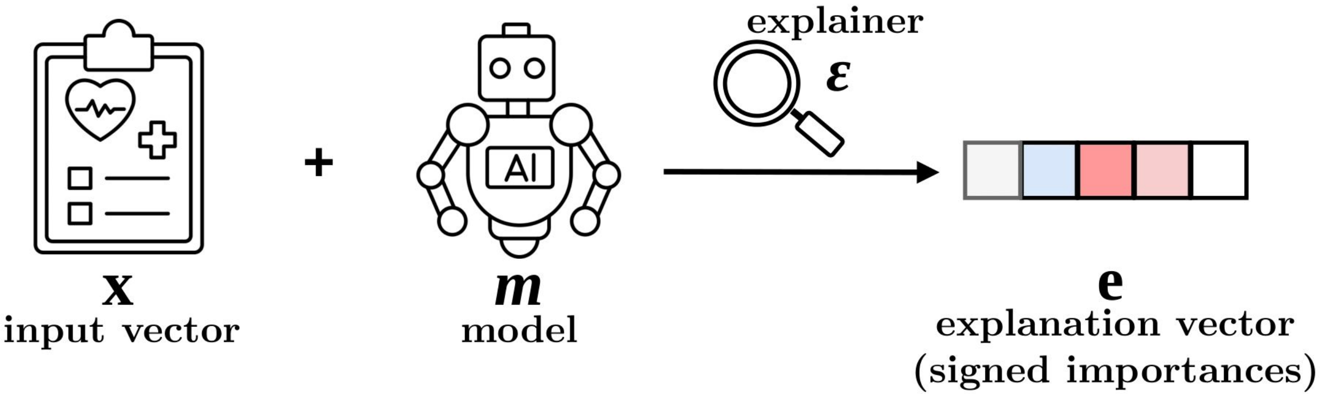
"When input **x** changes, the evaluation *might* change."

Principle 2: Model Relativism



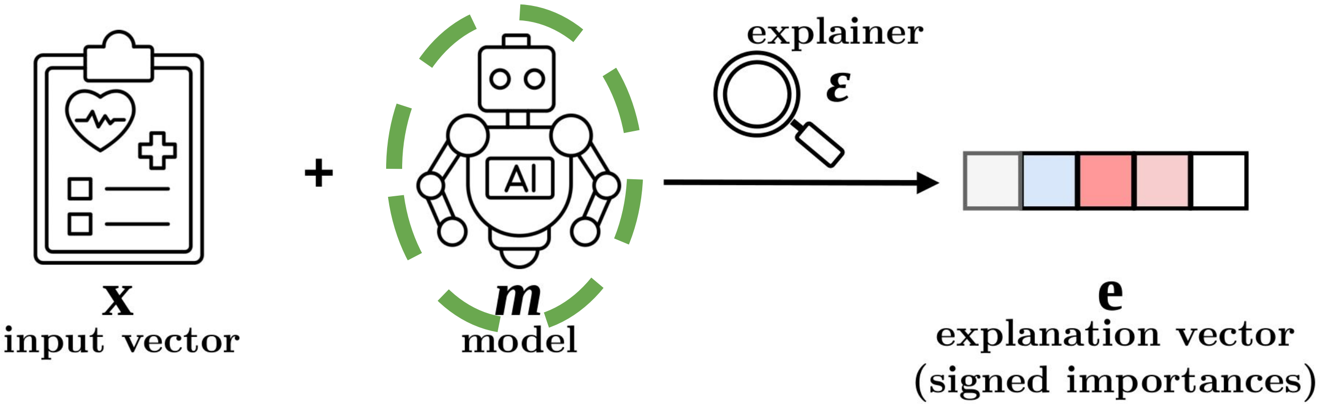
Principle 2: Model Relativism

Explanations should depend on the AI model



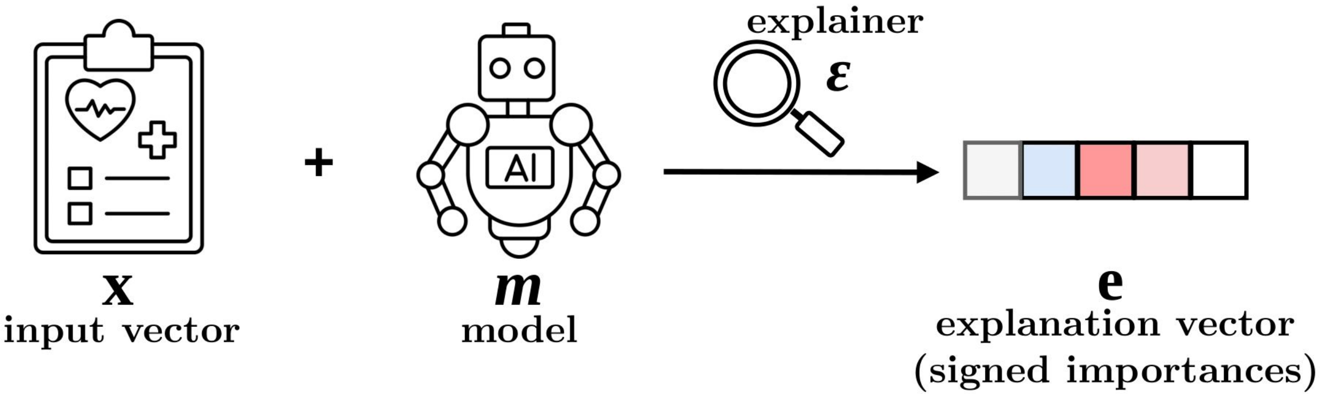
Principle 2: Model Relativism

Explanations should depend on the AI model



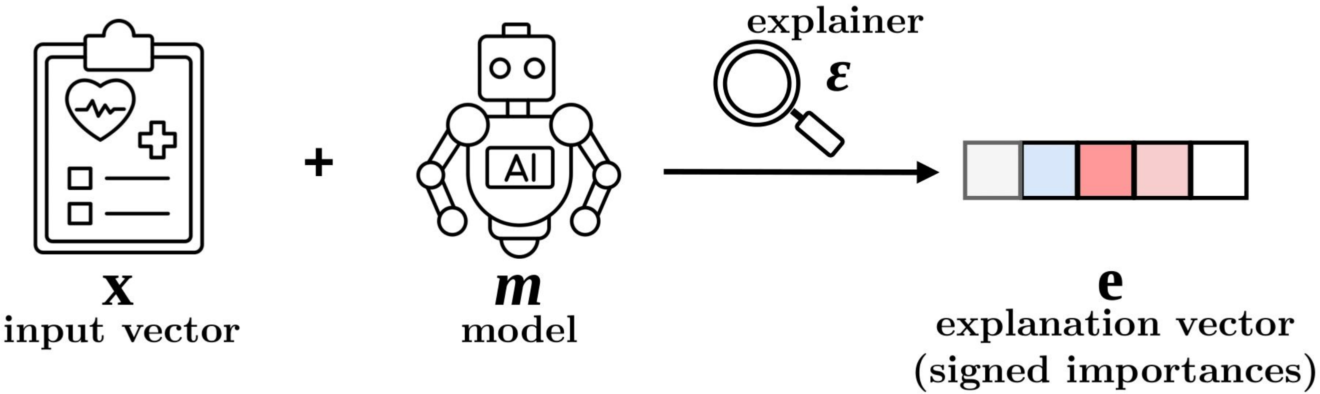
"When model m changes, the evaluation *should* change."

Principle 3: On-Manifold Evaluation



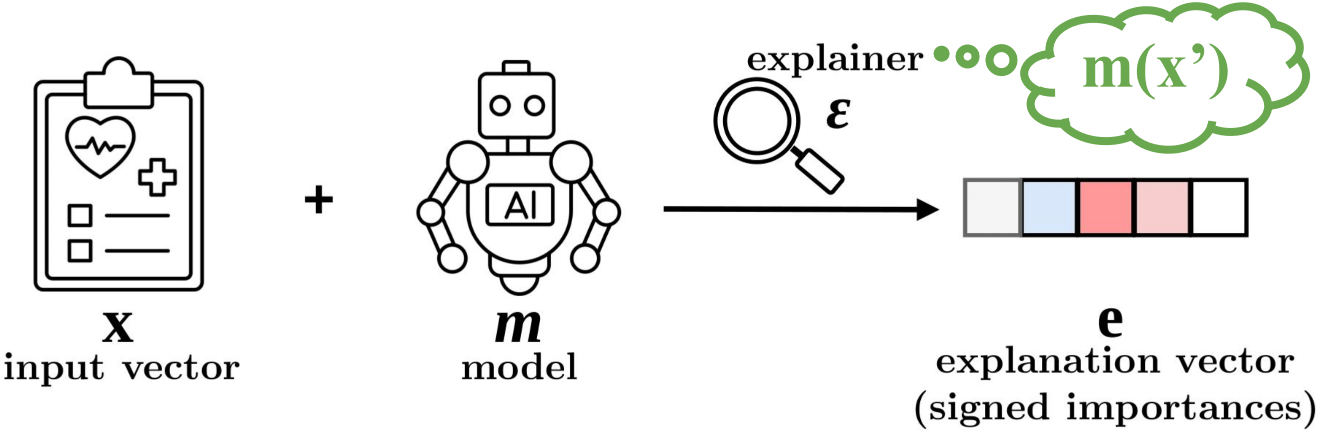
Principle 3: On-Manifold Evaluation

Explainers that sample points in synthetic neighbourhoods should not be sensitive to off-point model behaviour



Principle 3: On-Manifold Evaluation

Explainers that sample points in synthetic neighbourhoods should not be sensitive to off-point model behaviour



"The evaluation should *not* depend on model output **m(x')**."

Key Idea

- Recall “Short” Answer: select explanation that predicts model output.

Key Idea

- Recall "Short" Answer: select explanation that predicts model output.

- Use a K-NN model on the subset of important features to predict model output.

Key Idea

- Recall "Short" Answer: select *explanation* that predicts model output.

- Use a K-NN model on the *subset of important features* to predict model output.

Key Idea

- Recall "Short" Answer: select *explanation* that predicts model output.

- Use a K-NN model on the *subset of important features* to predict model output.

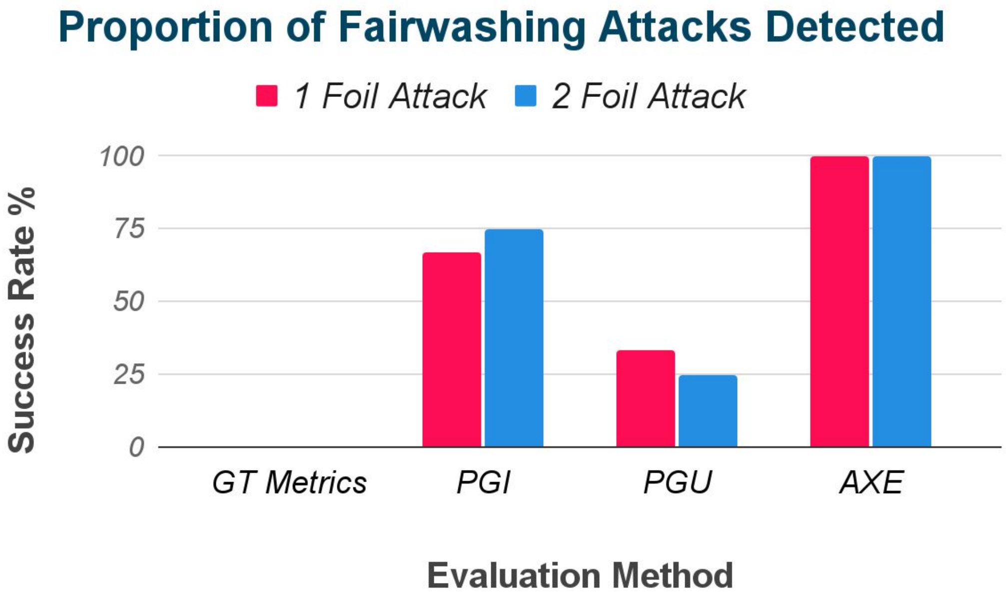
(Algorithm 1, Page 5)

Algorithm 1 Evaluating Explanation Quality with AXE_n^k

Require: Number of Features n, Number of Neighbors k
Dataset $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{\nu}$, Predictions $Y_{\text{preds}} = \{y_i\}_{i=1}^{\nu}$, and Explanations $E = \{\mathbf{e}_i\}_{i=1}^{\nu}$

1: Initialize an empty list: $\hat{Y} \leftarrow []$
2: **for** each datapoint $\mathbf{x}_i$ and explanation $\mathbf{e}_i$ in $(\mathcal{X}, E)$ **do**
3: Find n most important features: $f_{\text{imp}} \leftarrow \text{ImpFeatures}(\mathbf{e}_i, n)$
4: Create $\mathcal{X}_f$ with subset of features f_{imp} from $\mathcal{X}$
5: Train K-NN model M_i^k with inputs $\mathcal{X}_f$ and target Y_{preds}
6: Obtain prediction $\hat{y}_i$ from M_i^k for datapoint $\mathbf{x}_i$
7: Append $\hat{y}_i$ to $\hat{Y}$
8: **end for**
9: Return performance measure: $\text{Accuracy}(\hat{Y}, Y_{\text{preds}})$

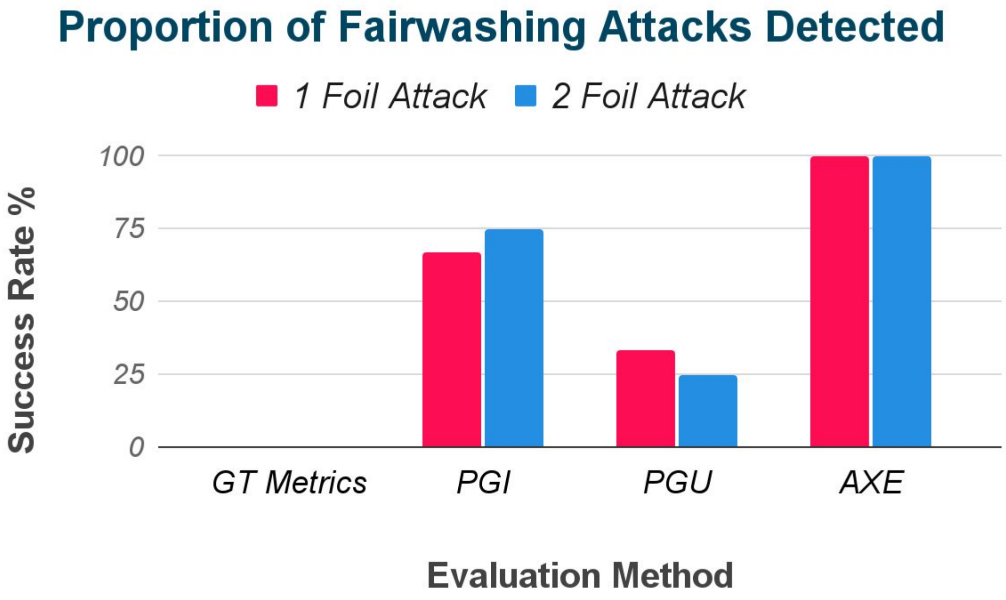
Detecting Explanation Fairwashing



(Table 2, Page 8)

Detecting Explanation Fairwashing

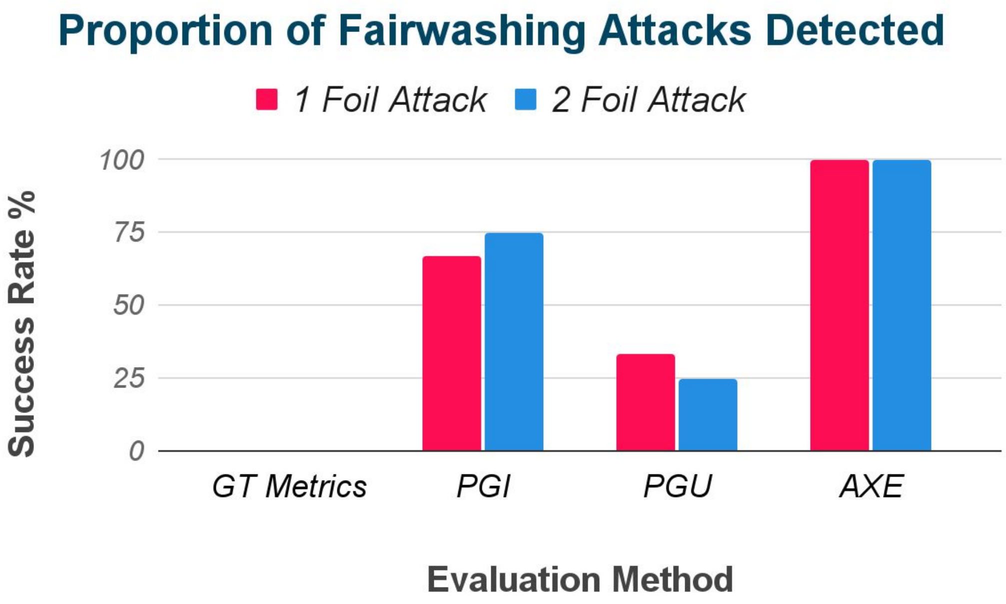
- Ground-Truth based metrics are unusable



(Table 2, Page 8)

Detecting Explanation Fairwashing

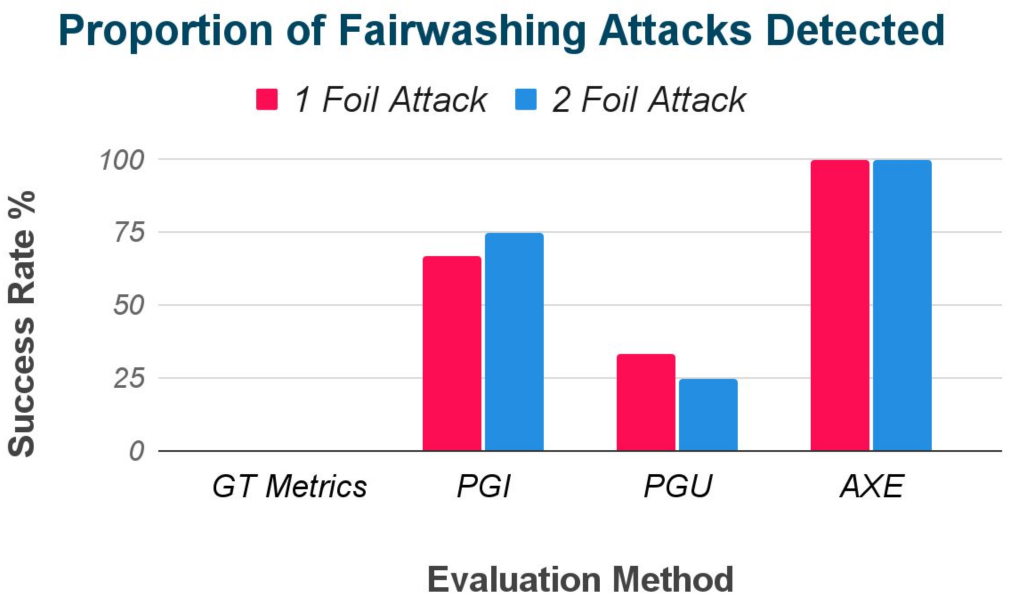
- Ground-Truth based metrics are unusable
- PGI and PGU are susceptible to adversarial attacks



(Table 2, Page 8)

Detecting Explanation Fairwashing

- Ground-Truth based metrics are unusable
- PGI and PGU are susceptible to adversarial attacks
- AXE is invulnerable and perfectly detects fairwashing



(Table 2, Page 8)

Why AXE?

- **Is principled**, following local contextualisation, model relativism, and on-manifold evaluation (sec. 2.4, 2.5);

Why AXE?

- **Is principled**, following local contextualisation, model relativism, and on-manifold evaluation (sec. 2.4, 2.5);
- **Provides robustness** to hyperparam variation (sec. 3.2);

Why AXE?

- **Is principled**, following local contextualisation, model relativism, and on-manifold evaluation (sec. 2.4, 2.5);
- **Provides robustness** to hyperparam variation (sec. 3.2);
- **Detects fairwashing** in explanations (sec. 4.1);

Why AXE?

- **Is principled**, following local contextualisation, model relativism, and on-manifold evaluation (sec. 2.4, 2.5);
- **Provides robustness** to hyperparam variation (sec. 3.2);
- **Detects fairwashing** in explanations (sec. 4.1);
- **Has high cogency** in its evaluations (sec. 4.2); &

Why AXE?

- **Is principled**, following local contextualisation, model relativism, and on-manifold evaluation (sec. 2.4, 2.5);
- **Provides robustness** to hyperparam variation (sec. 3.2);
- **Detects fairwashing** in explanations (sec. 4.1);
- **Has high cogency** in its evaluations (sec. 4.2); &
- Public code on GitHub!

Thank You

Kaivalya **Rawal**, Zihao Fu, Eoin Delaney, and Chris Russell. "**Evaluating Model Explanations without Ground Truth**" In The 2025 ACM Conference on Fairness, Accountability, and Transparency (**FAccT '25**), June 23–26, 2025, Athens, Greece. ACM, New York, NY, USA, 12 pages. **https://doi.org/10.1145/3715275.3732219**

https://github.com/KaiRawal/Evaluating-Model-Explanations-without-Ground-Truth

Thank You

(paper, code, & slides)

Kaivalya **Rawal**, Zihao Fu, Eoin Delaney, and Chris Russell. “**Evaluating Model Explanations without Ground Truth**” In The 2025 ACM Conference on Fairness, Accountability, and Transparency (**FAccT ’25**), June 23–26, 2025, Athens, Greece. ACM, New York, NY, USA, 12 pages. **https://doi.org/10.1145/3715275.3732219**

https://github.com/KaiRawal/Evaluating-Model-Explanations-without-Ground-Truth

Thank You

Questions?

(paper, code, & slides)

Kaivalya **Rawal**, Zihao Fu, Eoin Delaney, and Chris Russell. "**Evaluating Model Explanations without Ground Truth**" In The 2025 ACM Conference on Fairness, Accountability, and Transparency (**FAccT '25**), June 23–26, 2025, Athens, Greece. ACM, New York, NY, USA, 12 pages. **https://doi.org/10.1145/3715275.3732219**

https://github.com/KaiRawal/Evaluating-Model-Explanations-without-Ground-Truth

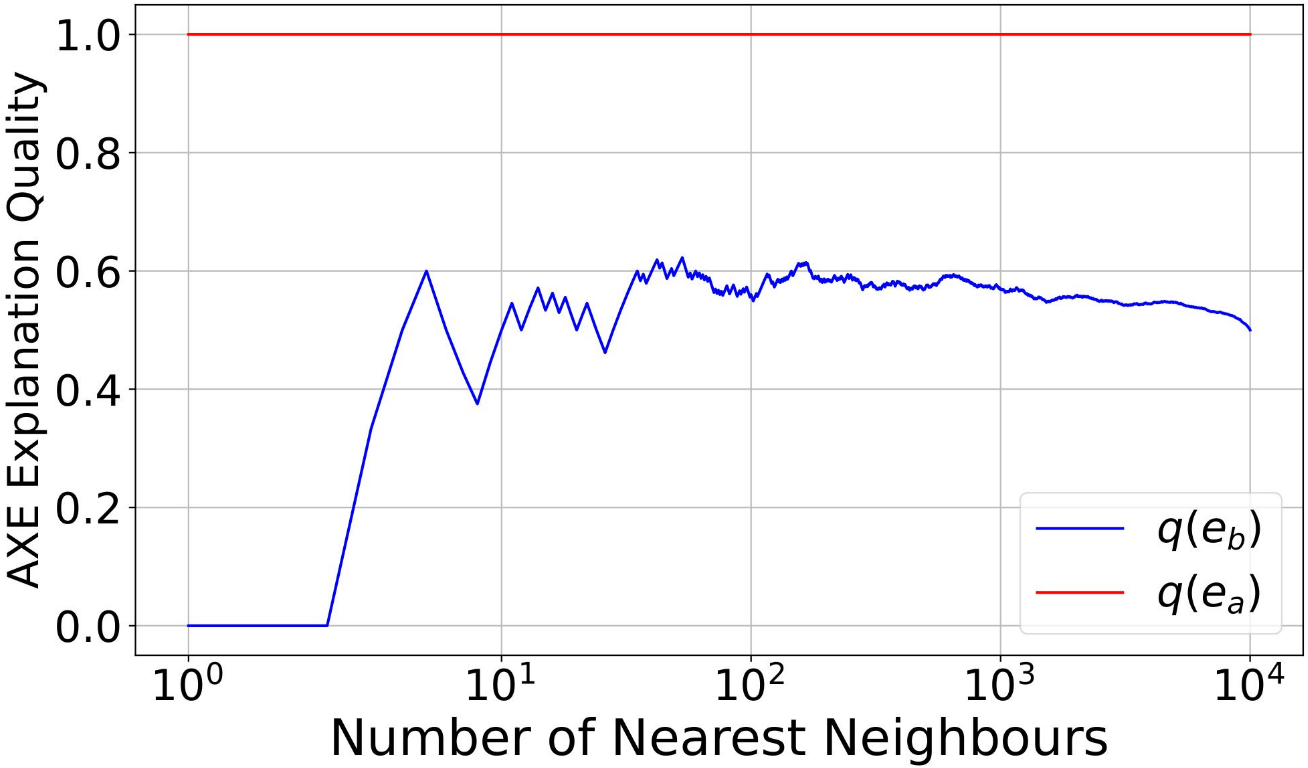
OXFORD INTERNET INSTITUTE

UNIVERSITY OF OXFORD

AXE vs Prior Evaluation Metrics

Metric	Definition	Local Contextualization	Model Relativism	On-Manifold Evaluation
FA: Feature Agreement	Fraction of top-n features common between **e** and **e***.	✗	✗	✔
RA: Rank Agreement	Fraction of top-n features common between **e** and **e*** with the same position in respective rank orders.	✗	✗	✔
SA: Sign Agreement	Fraction of top-n features common between **e** and **e*** with the same sign.	✗	✗	✔
SRA: Signed Rank Agreement	Fraction of top-n features common between **e** and **e*** with the same sign and rank.	✗	✗	✔
RC: Rank Correlation	Spearman's rank correlation coefficient for feature rankings from **e** and **e***.	✗	✗	✔
PRA: Pairwise Rank Agreement	Fraction of feature pairs for which relative ordering in **e** and **e*** is the same.	✗	✗	✔
PGI: Prediction-Gap on Important Feature Perturbation	Mean absolute change in model output upon perturbing top-n most important inputs.	✔	✔	✗
PGU*: Prediction-Gap on Unimportant Feature Perturbation	Mean absolute change in model output upon perturbing top-n most unimportant inputs.	✔	✔	✗
AXE: (ground-truth) Agnostic eXplanation Evaluation	Predictiveness of the top-n most important inputs in recovering model output. **Defined in section 3.1.**	✔	✔	✔

AXE across Hyperparameter Settings



Evaluation Cogency with AXE