Groundless Faithfulness: Structural Shortcomings of Explainable AI Evaluation Metrics for Feature Importance

Marco Zullich[1] , **Emily Schiller**[2,3] , **Ivan Gentile**[4] , **David Dembinsky**[5,6]
Adriano Lucieri[6] , **Kilian Göller**[7] , **Steffen Seitz**[7,8]
[1]TU Delft, [2]XITASO GmbH, [3]University College Cork, [4]IFAB Foundation, [5]RPTU University Kaiserslautern-Landau, [6]DFKI, [7]TU Dresden, [8]SECAI
All authors in no particular order.
Correspondence: m.zullich@tudelft.nl

Abstract

Explainable Artificial Intelligence (XAI) tools have often been presented as methods to increase trust in machine learning models by making aspects of their algorithms more human-understandable. However, evidence accumulated over the last decade points to an overall failure of this objective, instead promoting overreliance on incorrect predictions, which can subsequently lead to algorithmic aversion. Rather, XAI appears to function primarily as a debugging tool, useful for obtaining anecdotal insights into, for example, features learned by the model. One of the main issues underlying this limitation is that XAI methods often oversimplify the inner workings of the model, resulting in misalignment between the explanation and the model itself—a behavior referred to as *unfaithfulness* or *infidelity*. *Faithfulness* can be interpreted as the degree of approximation between an explanation and the model. Computing faithfulness in a rigorous manner would thus make it possible to quantify one of the various elements of trustworthiness of an explanation. However, existing metrics present significant limitations, making this quantification either infeasible or computationally expensive. In this position paper, we outline five major shortcomings of current faithfulness metrics and propose four possible avenues for research to bolster their reliability.

1 Introduction

With the increasing power of artificial intelligence (AI) models, their internal algorithmic processes have become more difficult to interpret, as exemplified in the notorious interpretability-accuracy trade-off [James *et al.*, 2021]. This opacity conflicts with transparency requirements such as the GDPR "Right to Explanation" [Goodman and Flaxman, 2017]. Explainable AI (XAI) has emerged as a field to enable transparency of *black-box* models by providing *explanations* that simplify *black-box* models for human understanding. Haufe *et al.* [2026] indicate ideal goals of XAI, including the recognition of model and data diagnostics (i.e., discovering incorrect model behaviors, such as "Clever Hans effects" [Lapuschkin *et al.*, 2019], or bias in datasets), the discovery of unknown relationships between variables, and the identification of intervention targets for modifying said relationships. However, evidence from roughly ten years of XAI research shows these tools often fall short on these promises, as recently discussed by Afroogh *et al.* [2026]. Beyond highlighting misalignment in explanation recipients, they report that XAI can foster user overreliance on incorrect AI predictions that would otherwise be questioned. In safety-critical applications, this may cause serious consequences and potential algorithmic aversion. Both Haufe *et al.* [2026] and Afroogh *et al.* [2026] indicate *infidelity* (or *unfaithfulness*) of XAI outputs to the original model as a source of said downfalls.

However, these works do not examine the technical issues of faithfulness metrics, nor explore faithfulness beyond feature importance (FI). In their survey on the evaluation of XAI, Dembinsky *et al.* [2026] cite 19 faithfulness metrics, applicable to different types of XAI: FI, concepts, example explanations, white-box surrogates, and natural language explanations. Faithfulness for FI is by far the type with the most approaches, driven also by the fact that FI is currently the most common XAI task [Saarela and Podgorelec, 2024]. Nevertheless, techniques for the computation of faithfulness, which mostly rely on (iterative) feature removal, are problematic at best: they are unreliable—e.g., due to issues with out-of-distribution (OoD) data, require arbitrary baselines, or require re-training phases which alter the nature of the original model. Some other fidelity metrics do not even measure proper faithfulness, or conflate it with other XAI properties. This unreliability raises the question whether FIs can be truly considered *unfaithful*, or whether faithfulness metrics themselves are dubious.

In this position paper, we analyze recent studies on faithfulness metric unreliability, arguing that modern XAI should prioritize more trustworthy approaches. The field should move beyond established but unreliable metrics like pixel flipping [Bach *et al.*, 2015] and remove-and-retrain (ROAR) [Hooker *et al.*, 2019] toward techniques considering (i) what is the real goal behind faithfulness, (ii) the model's uncertainty estimation (UE) capabilities, and (iii) expected properties of FI coefficients with respect to model predictions. We conclude with research avenues for advancing faithfulness evaluation in FI, restricting our claims to this domain given the substan-

tial available research.

2 Background

2.1 Feature Importance

FI is one of the foundational types of XAI [Dembinsky *et al.*, 2026]. Let us consider a model $f : \mathbb{R}^d \to \mathcal{Y}$, where $\mathcal{Y}$ is the output space. We can define a FI technique as a mapping $e : \mathbb{R}^d \to \mathbb{R}^d$ whose output we term *explanation*. The FI explanation provides a value (which we can term importance *coefficient* or *score*) for each feature[1], each value indicating the *importance* of the feature in the model's predictive dynamics. FI can be *global* (thus, the importance coefficients are valid for the model as a whole) or *local* (i.e., the importance scores are valid only for one specific point of the input space). Importance scores with a high magnitude indicate that the model relies greatly on the corresponding features for outputting the prediction; the sign instead indicates whether the feature contributes in increasing the model prediction (if positive) or in decreasing it (if negative). Some techniques might only present positive coefficients: for instance Permutation Feature Importance [Fisher *et al.*, 2019] only outputs importance in term of magnitude, while other techniques, such as Grad-CAM [Selvaraju *et al.*, 2020] suppresses negative importance scores for simplifying the interpretation of the explanation. A faithful FI should correctly indicate how important individual features are in the model's predictive dynamics. In the case of some FI techniques, such as SHAP [Lundberg and Lee, 2017], this is often defined in terms of *feature removal*: the deletion of an important feature (leaving the others intact) should cause a change in the model prediction, and that this change should be proportional to the importance coefficient. However, other tools, such as gradient-based FI for neural networks just report (aggregations of) gradients as scores, without referring to feature removal. FI can also be called *feature attribution*; when working with image data, these explanations can also be termed *(importance) heatmaps* or *saliency maps*.

2.2 Evaluation of Feature Importance

According to Nauta *et al.* [2023], explanations can be evaluated from the perspectives of their content, their presentations, or their users. The assessment of the content can be further grouped in different categories—alignment between model and explanation (which the authors call *correctness* and *completeness* and that overall makes up the faithfulness [Dembinsky *et al.*, 2026]), continuity (robustness) of the explanation to perturbations (to input, output, or model), and compactness, among others. The presentation point of view assesses the mean through which the explanation is presented, regardless of its content. The user viewpoint describes instead the conditions (context, previous knowledge, etc.) in which the user interacts with the explanation. A subgroup of the user category worth mentioning is *coherence*, which other works also call *plausibility*: the alignment between an explanation and human prior knowledge.

Faithfulness in Feature Importance

Faithfulness (also called *fidelity* or *correctness* in the literature) describes how truthful the explanation is to the original model f [Nauta *et al.*, 2023]. For FI, this requires alignment between features the model truly considers important and those the explanation identifies as important. In their survey, Dembinsky *et al.* [2026] identify input perturbation as the primary methodology for evaluating FI faithfulness. The rationale follows from the feature removal considerations presented in Section 2.1: perturbing (i.e., *removing*) important features should substantially change model predictions, with magnitude proportional to feature importance. The sign should be the opposite of the sign of the FI output. The authors distinguish between *unguided* and *guided* perturbation approaches. Unguided approaches operate a perturbation without considering the hierarchy dictated by the FI output. Guided approaches, conversely, perturb a data point sequentially targeting features by importance, either most-to-least in most relevant first (MoRF) or least-to-most in least relevant first (LeRF). Nauta *et al.* [2023] additionally distinguish between *incremental deletion* or *incremental addition* approaches: in the former, the features are sequentially perturbed starting from an unperturbed data point; in the latter the initial data point is completely perturbed, and the original features are *revealed* in MoRF or LeRF order. Plotting model output versus perturbation iteration yields a *perturbation curve*. In the case of classification, the model *confidence* (probability value associated to the predicted class) is often used as output in the plot. We showcase a possible example for classification in Figure 1—MoRF removal causes a sharp decrease in confidence at first, while removing less important features has a small effect on the output; with LeRF removal the confidence drops at an increasing rate, proportional to the importance of features removed. We point out that the behavior presented in Figure 1 is ideal—often, for reasons explored in the next sections, the situation can be very noisy. Other factors, such as *gradient starvation* [Pezeshki *et al.*, 2021], can also modify the behavior of the curve: if multiple features are associated with a given prediction, models may concentrate originally only on a small subset of these. After perturbation of important features, the models may shift the focus on the other features, while still outputting the same prediction, modifying the *ideal* scenario in Figure 1. To quantify faithfulness for an (f, e) pair, the technique is applied over n data points and aggregated via:

i) Correlation-based metrics: compute the Pearson's or Spearman's correlation, or Kendall's τ, between the importance of the perturbed features and the opposite of the change in model prediction. A positive correlation should be observed.

ii) Area-based metrics: plot the model prediction over the importance of perturbed features and measure faithfulness as the area under the perturbation curve (AUPC) or area over the perturbation curve (AOPC) (or a combination of both [Knoll *et al.*, 2025]). The metric can be reported either absolute or normalized. Depending

[1]For simplicity, we consider only FIs that generate one importance value per raw feature. This is not necessarily true, e.g., when considering feature interaction [Fumagalli *et al.*, 2023], supersets of features [Ribeiro *et al.*, 2016], or explanations in different domains than the input's [Mishra *et al.*, 2020].

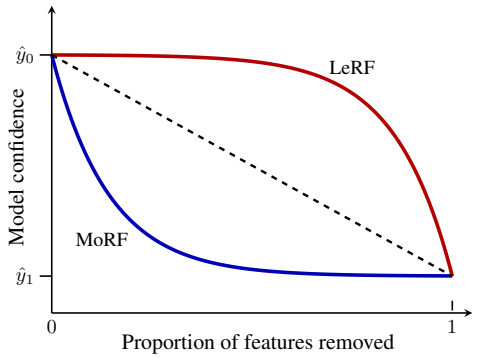


Figure 1: Example of possible perturbation curves obtained in a classification setting for an incremental deletion faithfulness evaluation. The model starts by predicting a confidence value $\hat{y}_0$ for the predicted class on the unperturbed data. As more features are removed, the model confidence drops until reaching a minimum $\hat{y}_1$ when the data point is completely perturbed. The curve depicted in blue shows an optimal MoRF confidence trend, while the red curve is expected to be attained for a LeRF perturbation.

on the approach adopted—incremental deletion versus addition, MoRF versus LeRF—a highly faithful explanation should lead to a normalized area value close to 1 or 0 (if the metric is normalized).

3 Issues with existing faithfulness metrics

3.1 Faithfulness computation is almost exclusively operated on classification

Most of the literature on faithfulness computation is limited to classification settings (e.g., [Bhatt *et al.*, 2021; Zheng *et al.*, 2024]), while regression tasks remain insufficiently addressed. This limitation is not merely a gap in coverage but reflects incompatibilities between metric design assumptions and regression. Most metrics assume that removing (or inserting) important features should induce monotonic degradation (or improvement) in model confidence, specifically measured through classification scores [Samek *et al.*, 2016] such as softmax probabilities or logits. This assumption is well-grounded in classification, where the relevant quantity, class probability, is naturally bounded between 0 and 1, and where degradation has a clear interpretation of decreased confidence. In regression, the target quantity itself is ambiguous: should faithfulness measure changes in the predicted value or in deviation from the ground truth [Demetriou *et al.*, 2025]? These probe fundamentally different qualities. Because both quantities are typically unbounded, normalization becomes necessary but problematic: faithfulness scores become incomparable across datasets and even across instances within the same dataset, undermining their utility as evaluation metrics. Additionally, perturbations may cause arbitrary shifts in predictions without clear monotonic relationships, making the "degradation" criterion ill-defined. Focusing on prediction error conflates model accuracy with explanation quality, measuring whether the model fails rather than whether the explanation faithfully represents the model's reasoning. Altogether, the literature reflects a deep bias of faithfulness metrics for FI toward classification tasks that cannot be trivially adapted to regression contexts.

3.2 Some faithfulness metrics conflate fidelity with other properties of explanations

The very definition of faithfulness which we presented in Section 2.2 is often misrepresented in some mainstream approaches for faithfulness computation presented in the literature. We first note that sometimes the term *correctness* is confused for *plausibility*, as in the cases of Rony and Hassan [2025], and Maathuis *et al.* [2026].

On a more foundational level, we notice that, additionally, some metrics for faithfulness conflate faithfulness itself with the notion of plausibility. Since faithfulness is aimed at measuring the alignment between the explanation and the model, its evaluation should not depend on external notions such as human-generated labels or ground-truth annotations. In line with the observations on fidelity computation from Jacovi and Goldberg [2020], we posit that faithfulness should be assessed only on the basis of the model internals. Similarly, Dembinsky et al. [2026] argue that ground-truth dataset evaluation assesses faithfulness only when a uniquely defined rationale can be assumed beyond reasonable doubt; otherwise, it merely evaluates plausibility. However, some highly-adopted approaches for faithfulness computation target changes in performance metrics of the model instead of prediction values. This is the case, e.g., for ROAR [Hooker *et al.*, 2019] and ROAD [Rong *et al.*, 2022]. ROAR and ROAD measure faithfulness in terms of performance difference after perturbation (and re-training for the former). Measuring faithfulness, e.g., in terms of accuracy drops is probing the alignment between explanation and ground truth, not between explanation and model.

Additionally, since ROAR requires a re-training of the model on perturbed input, we posit that this metric is not even assessing the quality of the explanations according to the original model, but according to a new model, which can be completely distinct in its internals from the original—an issue which Dembinsky *et al.* [2026] call *different contextuality*. For these reasons, we will avoid including ROAR in future discussions, even though it is often cited as being a state-of-the-art faithfulness metric [Song *et al.*, 2023; Klein *et al.*, 2024].

3.3 Perturbation as a stand-in for feature removal requires arbitrary baselines

As stated in Section 2.2, perturbations are involved in the faithfulness computation procedure as a stand-in for feature removal, which is often impossible to reproduce with most machine learning models unless a retraining procedure is operated. The main problem with this methodology is that there does not exist a single best solution for reproducing feature removal—the baseline for this procedure is thus arbitrary and it is thus impossible to define what perturbation would encode this *missingness of information*. The problem is equivalent to the one faced by FI tools that work by feature removal, the most important one being SHAP. Strumfels *et al.* [2020] already ran an extensive study on the effect of the baseline selection (constant baseline, blurring, random uniform and

gaussian) for image-based FI generation, which led to the conclusion of there not being a "best" solution. For what concerns faithfulness computation, different approaches are presented in the literature and largely depend on the data type: in image data, the most common approach is *fixed imputation*, where a patch of a uniform color (e.g., white, gray, black, or the dataset mean) is applied to the image to *occlude* important features [Samek *et al.*, 2016; Sundararajan *et al.*, 2017]; other approaches use blurring or noising [Zheng *et al.*, 2024] or use generative models for image inpainting [Agarwal and Nguyen, 2020]. For text data, hard token deletion can be applied [DeYoung *et al.*, 2020], or tokens can be replaced with specific mask tokens [Atanasova *et al.*, 2020], fixed embeddings (e.g., zero-embedding [Li *et al.*, 2016]) or with tokens generated by another language model [Kim *et al.*, 2020; Wu *et al.*, 2021]. In tabular data, common approaches include fixed imputation with a constant or marginal baseline or noising.

3.4 Many perturbation-based approaches cause OoD issues with models

Besides the difficulty with identifying what information missingness means for a specific model or type of data, the search for *good* perturbation baselines is also driven by another noticeable issue with artificial neural networks (ANNs): their overconfidence on OoD data. Indeed, these models have been largely known to output random, overconfident predictions when presented with data which are too *distant* from their original training data distribution [Nguyen *et al.*, 2015]. All the perturbation strategies insofar presented are somewhat sensitive to the OoD issue [Hase *et al.*, 2021], and some [Rong *et al.*, 2022; Zheng *et al.*, 2024] have specifically been developed to tackle this problem. Another connected issue is that classification ANNs, unless explicitly trained on data with missing information, or correctly calibrated on the uncertainty profile [Guo *et al.*, 2017], these models do not possess the ability to say "*I do not know*" (expressed as a flat output distribution). The way the OoD issue reflects on the perturbation curve is the fact that the model response becomes erratic, with highly-perturbed data often having a higher confidence than the previous iterations. We display a real-life example in Figure 4. However, the faithfulness metrics introduced in Section 2.2 all work under the assumption that the perturbation curve is monotonically decreasing, which clearly conflicts from the "clean" examples from Figure 1.

The main approaches for combating the OoD issue in faithfulness are: (i) augmentative training [Hase *et al.*, 2021] (inlcuding the F-fidelity framework [Zheng *et al.*, 2024]), where the training data is augmented by reproducing the perturbation patterns used in the faithfulness assessment, and (ii) generative inpainting [Agarwal and Nguyen, 2020], where the features are removed using an inpainting model that should still produce feasible data. However, these two approaches do not solve the problem: augmentative training only works until a reasonable level of perturbation. High levels of perturbation will likely cause a complete erasing of the original features, which, in the case of supervised learning, will potentially force the model to learn an incorrect relationship between data and response. This potentially causes confident predictions even with a large portion of important features removed. Generative inpainting causes absence of information in the image, which a non-calibrated model does not know how to interpret—a behavior that leads to overconfident OoD predictions, thus impairing an approach which in principle should be valid. Notice also that these approaches are mainly aimed at image data, and may not be easily translated to other data modalities. Recent advances in XAI-aware model calibration [Sridhar *et al.*, 2026; Cai *et al.*, 2026] seem to go in the right direction: address the model calibration, without proposing other workarounds for OoD detection.

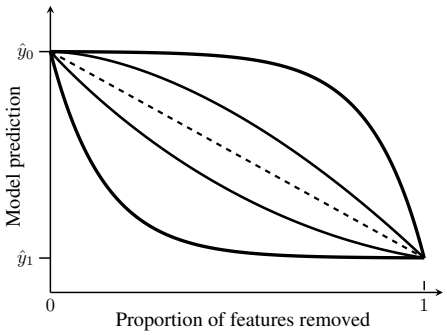


Figure 2: Example of perturbation curves whose Spearman's correlation coefficient is always -1. The dotted line represents a case in which Pearson's correlation coefficient is -1. In either case (MoRF or LeRF), this would lead to a perfect faithfulness, although these configurations clearly represent different cases. In the MoRF case, the lower curves would represent an ideal scenario, whereby the most important features have a higher impact on the model prediction; the dotted line and the two upper curves represent instead an unsound scenario for faithfulness, with highly-important feature having close to no effect on the prediciton, while features with a lower importance contribute more. This shows the fault in correlation coefficients as a faithfulness metric.

3.5 The Formal Divergence of Correlation and Area-Based Metrics from Explanatory Faithfulness

This point represents the fundamental criticism we raise to current FI faithfulness metrics—the existing correlation-based and area-based metrics do not really measure faithfulness correctly. Miró-Nicolau *et al.* [2025] already showed, on simple experiments, that existing faithfulness metrics seem to be misaligned to the actual goal of measuring faithfulness. In this section, we expand on their experiments by providing additional criticism on the current approaches. For simplicity, we will work with an example of classification model, on which we evaluate explanations using guided perturbation fidelity obtained via MoRF, incremental deletion feature perturbation. We additionally assume that the explanation only presents positive importance scores. We expect the behavior of the curve to follow a trend similar to the one depicted in Figure 1: there should be a proportionality between the importance of features removed and the drop in confidence observed. We can observe that, in this case, neither correlation-based, nor area-based metrics are aligned with measuring

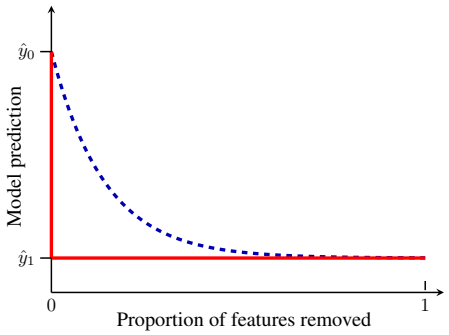


Figure 3: Example perturbation curves in a MoRF, incremental deletion setting. Despite the dashed blue curve representing an desirable case of faithfulness, AOPC rewards the red curve with a perfect score (normalized AOPC of 1). However, the latter is a degenerate situation, whereby only the feature with the highest importance affects the model's prediction, while lower-importance features have no effect. On the other hand AOPC should reward the blue case, abiding to the aforementioned assumption of proportionality.

faithfulness since they break this assumption:

- Pearson's correlation assumes a linear relationship between the two axes. Thus, a perfectly faithful explanation (see Figure 2) would require a uniform drop in confidence for each iteration. This would mean that the contribution of each feature is the same, regardless of its importance.
- Nonlinear correlation metrics (e.g., Spearman's ρ) are just measuring the monotonicity of the trend, they do not consider its convexity. Thus, the proportionality assumption could be violated, since any monotonic trend would lead to a maximum faithfulness score, as depicted in Figure 2. High nonlinear correlation is hence more of a *necessary* property of a faithful explanation than a *sufficient* one.
- AOPC would require a perfectly faithful explanation to follow the trend highlighted in Figure 3. This conflicts with the expected trend from Figure 1 and violates the proportionality assumption: only the most important feature(s) would lead to a drop in confidence, and all the relevant features with lower importance would not have any effect on the model prediction. The only situation in which AOPC would be a valid faithfulness metric is when the explanation is binarized, thus possibly leading to the behavior in Figure 3.

In addition, there is one more point to be made: the assumption of confidence drop in classification can only be reliably met when the faithfulness metrics are run on the predicted class. In case the explanation is run on a non-predicted class, it is likely that the confidence would be very low, possibly close to 0. In the ideal case of a perfectly calibrated model, removing important features would possibly cause an *increase* in confidence prediction (towards $1/K$, with K being the number of classes). In the uncalibrated case, this confidence increase could be even higher, due to the existence of a so called *catch-all category* [Miró-Nicolau *et al.*, 2025] that the model designates as a default class for OoD examples. This makes the current approaches infeasible for measuring the faithfulness of non-predicted classes. Figure 4 shows a real-life example of perturbation curves in a model with poor OoD capabilities, where one class is designated as a high-confident *catch-all*.

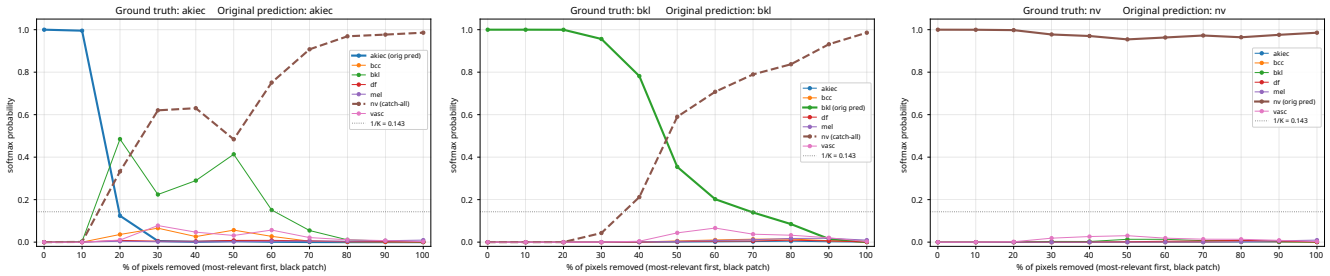
4 Our proposal for the direction of FI faithfulness research

Having outlined key issues with current faithfulness computation, we propose research directions towards more solid and trustworthy faithfulness evaluation. This aims to restore confidence in XAI methods, which, per Section 1, has declined due to unfulfilled promises regarding model interpretability.

Proposal I – faithfulness should be evaluated on well-calibrated models robust to OoD data: faithfulness relies on the assumption that models should reduce the confidence in their predictions as the data gets progressively perturbed, as explored in Section 2.2. This collides with the fact that most ANNs are largely overconfident in their predictions, especially for what concerns OoD data, as seen in Section 3.4. We hence posit that faithfulness should be computed on well-calibrated, OoD-robust models. In that regard, we believe the recent research on XAI-aware model calibration [Sridhar *et al.*, 2026; Cai *et al.*, 2026] to be a valid direction, favoring a more reliable faithfulness computation. Another underexplored line of research that needs to be expanded upon is XAI for Bayesian neural networks [Valdenegro-Toro and Mulye, 2024]: these models tend to display a more natural robustness to OoD data. In addition to these benefits, calibration has other positive effects on the model trustworthiness: a model whose uncertainty estimates are well calibrated enhances human oversight, allowing stakeholders to immediately get an indication on the possible (in)correctness of a prediction—outputs with low confidence are more likely to be inaccurate than highly-confident ones [Guo *et al.*, 2017].

Proposal II – faithfulness metrics for classification needs to consider the cases of non-predicted classes: as outlined in Section 3.5, faithfulness currently lacks an evaluation approach for explanations generated on non-predicted classes. The generation of explanations on classes other than the predicted ones could be necessary in some cases, e.g., for contrastive explanations [Sarti *et al.*, 2024; Zhang *et al.*, 2025]. Thus, we posit that approaches for faithfulness computation for classification tasks should generalize to all classes, not only the predicted one. Until such approaches are formulated, we posit that practitioners should concentrate only on generating explanations on the predicted category.

Proposal III – faithfulness for regression needs more research efforts: as discussed in Section 3.1, current faithfulness metrics are predominantly tailored to classification tasks and cannot be trivially adapted to regression settings without introducing conceptual problems. Unlike classification, regression requires metrics that assess whether FI explanations correctly predict both the direction and magnitude of changes in predicted values. We argue that addressing this gap requires dedicated research efforts that explicitly account for the unique characteristics of regression tasks. Specifically, the community should develop formalizations of faith-



(a) (b) (c)

Figure 4: Example of perturbation curves obtained on a ResNet-18 ANN [He *et al.*, 2016] trained on the HAM10000 dataset [Codella *et al.*, 2019]. The dataset contains images from seven classes of skin cancer (`akiec`=Actinic keratoses, `bcc`=basal cell carcinoma, `bkl`=benign keratosis-like lesions, `df`=dermatofibroma, `mel`=melanoma, `nv`=melanocytic nevi, `vasc`=vascular lesions). The plots are obtained by incrementally deleting features in MoRF order with black patches. In all cases, the model designs the `nv` class as a *catch-all category*. (a) and (b) show images from ground truth classes different than `nv`, while (c) shows an example from the `nv` class, where the model confidence remains extremely high when the image is completely black. This demonstrates that faithfulness computation is flawed for models with poor OoD capabilities. The $1/K$ line represents the optimal case for a well-calibrated model, which should output a flat distribution for completely black images. None of the three figures showcase anything close to this behavior.

fulness metrics designed specifically for regression, rather than adapting classification-based approaches that rely on incompatible assumptions. New metrics such as TIMING [Jang *et al.*, 2025] represent a solid avenue, although the reliance on area-based metrics may still be problematic, due to the issues raised in the present paper. We additionally highlight that this proposal may pose issues in conjunction with Proposal I, since calibration for regression models less well-established than for classification [Levi *et al.*, 2022].

Proposal IV – faithfulness metrics should be more aligned to their goal: with regards to what has already been outlined in Section 3.5, we believe the current approaches based on correlation and area-based metrics (and related work-arounds) should be scraped, favoring instead approaches that consider what are the *expectations* from a faithful explanation, and that reliable metrics be developed abiding to these expectations. We consider the following points:

i. The x-axis of the perturbation curve should represent the importance of the features removed, while many approaches use the iteration number instead. The latter does not connect to a direct value of FI, which does not allow to get a sense of the prediction change proportional to said importance.

ii. If an optimal shape of the prediction curve is established (e.g., similar to the one in Figure 1), then an inspiration for the faithfulness computation could be coming from the field of uncertainty estimation. Specifically, the expected calibration error (ECE) already offers an example of a metric considering a difference with an expected curve shape—see Figure 5. We stress out that currently there does not exist an agreed-upon definition of *optimality* for a perturbation curve, and we posit more research should be devoted to eliciting this. The choice of perturbation and phenomena such as gradient starvation (see Section 3.5) may pose a big hurdle in determining an optimal shape of the curve.

iii. Since faithfulness metrics should foster model trustworthiness, these should necessarily be relative in order to allow for direct interpretation of the faithfulness value. Absolute metrics calculated, e.g., on the output logits, such as in the case of IROF [Rieger and Hansen, 2020] do not offer a relative value (e.g., between 0 and 1) and logits can easily differ between models, making comparisons impossible and faithfulness scores uninterpretable.

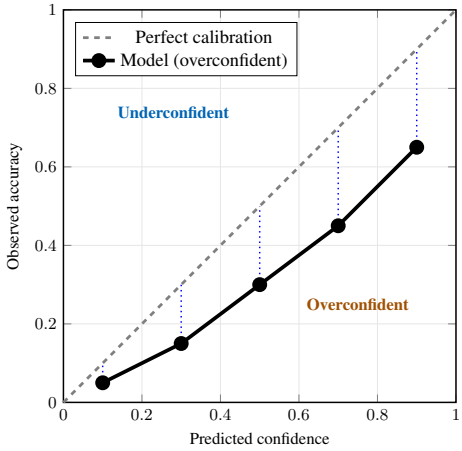


Figure 5: Example of calibration plot for a toy classification model. The model is evaluated on a validation dataset. The data are then split into bins depending on the confidence and the average accuracy per bin is computed and plotted (solid black line). The ECE is calculated as the magnitude of differences (dotted blue lines) between these points and the diagonal (gray dashed line), which represents perfect calibration where confidence equals accuracy.

5 Contextualizing Faithfulness Assessment

Issues with faithfulness metrics are well-known in the XAI literature. Ever since the still-used pixel/feature flipping experiment [Bach *et al.*, 2015], subsequent methods have tried to address its shortcomings, mainly on OoD data created

in the perturbation process: we have cited the examples of ROAR [Hooker *et al.*, 2019]—which proposes a retraining process—and the derivative ROAD [Rong *et al.*, 2022]—which instead proposes blurring to avoid the expensive retraining. More recently, works like F-fidelity [Zheng *et al.*, 2024] have proposed the usage of augmentative finetuning to curb the OoD issues. However, there are papers which have targeted a more comprehensive critical approach to faithfulness evaluation: the previously cited work by Miró-Nicolau *et al.* [2025] indicates how the main approaches seem to be misaligned with the goal of measuring faithfulness in models trained on simple tasks, where the important features are known in advance. The authors frame their findings on a critical level, not listing concrete proposals for future directions. Additionally, Chan *et al.* [2022], in their comparative study, notice inconsistencies between different faithfulness metrics on natural language processing tasks. They argue in favor of correlation-based metrics; however, they miss the issues behind these approaches which we instead outline in Section 3.5. Seth and Sankarapu [2025] highlight the importance of a sound evaluation of XAI and criticize the current field because of (i) fragmentation—a lack of standard in metrics—and (ii) manipulability—metrics are not robust and can be manipulated easily. However, their work is designed at a governance level and does not address the technical details behind these shortcomings. Finally, the previously mentioned Haufe *et al.* [2026] also address XAI shortcomings. Their perspective encompasses several points, including the lack of formalization in XAI, the fragility of the current metrics for XAI evaluation, the lack of verification that explanations abide by the formal requirements, and the scarcity of real-world data for performing empirical evaluations. The authors also mention that current XAI tools and faithfulness metrics do not take into consideration the data generating distribution; however, they do not propose how to integrate this into existing methods. Our proposals go in the direction highlighted by the authors; specifically, we also call for higher formality in XAI evaluation, e.g., by soliciting research to define a desired form of perturbation curve (Proposal IV) and by posing more stringent requirements on the model calibration (Proposal I).

6 Discussion and conclusion

In the present paper, we have summarized several issues connected with the current metrics for the computation of faithfulness of FI. Faithfulness is defined as the level of alignment between the FI explanation and the internals of the model on which said explanation is generated. We highlighted how the main metrics approach the computation from the process of *data intervention* [Dembinsky *et al.*, 2026]: perturbing specific features of a given input should yield variations in the model output which are proportional to the importance of the features. This is often done in an iterative fashion, by incrementally *removing* (or *revealing*) features by means of perturbations operated in a MoRF or LeRF fashion. The output of the model is tracked for each perturbed input and then plotted against the fraction of features removed, as shown in Figure 1. Faithfulness is then summarized with a scalar by either computing the correlation of the perturbation curve or its area under (or over) the curve.

We posit that these approaches present several flaws. Some of these criticisms are novel, while some others are already well known in the literature. We present five main shortcomings: (1) disproportionate focus on classification for faithfulness research, (2) conflation of faithfulness with other XAI evaluation facets, (3) necessity of an arbitrary baseline for simulating feature removal, (4) emergence of OoD data (which most ANNs handle poorly) as effect of the perturbation, and (5) misalignment between the main metrics and the proper computation of faithfulness. These drawbacks severely limit the assessment of faithfulness—a metric whose importance is vital in XAI evaluation, since it essentially quantifies the *level of approximation* of an explanation with regard to the model, and is hence fundamental for fostering reliability of FI as the main XAI task. Increasing the trustworthiness of XAI is hence necessary, in a time in which, after roughly a decade of research, the field is being severely criticized for its shortcomings in fulfilling its initial commitments in improving model transparency for its users [Afroogh *et al.*, 2026]. We conclude our paper by proposing four avenues for faithfulness research, namely: (I) contemplate model calibration when FI needs formal evaluation, (II) for classification, consider that explanations can be generated either on predicted or non-predicted classes (while current methods concentrate solely on predicted classes), (III) direct more effort to faithfulness for regression, and (IV) align metrics to the actual goal of computing faithfulness. We hope our manuscript can provide a clear and comprehensive summary of the current problems with the computation of faithfulness, at the same time giving directions to practitioners and researchers on how the research in the field can be improved. We close our paper by shifting back to Section 1—we posit Afroogh *et al.*'s [2026] point on infidelity to be inaccurate: it is not that XAI is unfaithful, but we are currently lacking a principled way for assessing faithfulness. Even though a more formal approach to faithfulness will not solve all of the shortcomings of XAI, as outlined by Haufe *et al.* [2026], this will still help with bolstering its credibility.

Limitations: despite providing a comprehensive critical outline of the current metrics for FI faithfulness, our paper does not consider other popular XAI tasks, such as counterfactual examples or concepts, that all require faithfulness computation [Dembinsky *et al.*, 2026]. Concepts, for instance, present similar challenges to FI in this regard. Additionally, if we exclude the example from Figure 4, the present paper lacks experiments that summarize the issues we highlight in Section 3, although many of these problems are already well-known in the literature. Finally, some of our directions from Section 4 (e.g., Proposals II and III) act more as recommendations for research and do not contain any concrete idea on how the limitations should be tackled. Similarly, the limitation on the arbitrary nature of perturbation baselines from Section 3.3 is not tackled by our proposals and remains currently unresolved. Notwithstanding these limitations, we believe our position paper still effectively summarizes shortcomings of current approaches and provides valuable insights for directing future research on faithfulness.

Acknowledgment

This work is partly supported by BMFTR (Federal Ministry of Research, Technology and Space) in DAAD project 57616814 (SECAI, School of Embedded Composite AI, https://secai.org/).

References

[Afroogh *et al.*, 2026] Saleh Afroogh, Syed Ishtiaque Ahmed, Petra Ahrweiler, David Alvarez-Melis, Mansur Maturidi Arief, Emilia Barakova, Falco J Bargagli-Stoffi, Erdem Biyik, Hanjie Chen, Xiang'Anthony' Chen, et al. Beyond Explainable AI (XAI): An Overdue Paradigm Shift and Post-XAI Research Directions. *arXiv preprint arXiv:2602.24176*, 2026.

[Agarwal and Nguyen, 2020] Chirag Agarwal and Anh Nguyen. Explaining image classifiers by removing input features using generative models. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[Atanasova *et al.*, 2020] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 3256–3274, 2020.

[Bach *et al.*, 2015] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[Bhatt *et al.*, 2021] Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021.

[Cai *et al.*, 2026] Yi Cai, Thibaud Ardoin, Mayank Gulati, and Gerhard Wunder. Rethinking Explanation Evaluation Under the Retraining Scheme. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 19826–19834, 2026.

[Chan *et al.*, 2022] Chun Sik Chan, Huanqi Kong, and Liang Guanqing. A comparative study of faithfulness metrics for model interpretability methods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 5029–5038, 2022.

[Codella *et al.*, 2019] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, and Michael Marchetti. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1902.03368*, 2019.

[Dembinsky *et al.*, 2026] David Dembinsky, Adriano Lucieri, Stanislav Frolov, Hiba Najjar, Ko Watanabe, and Andreas Dengel. Unifying VXAI: A Systematic Review and Framework for the Evaluation of Explainable AI. *Transactions on Machine Learning Research*, 2026.

[Demetriou *et al.*, 2025] Xenia Demetriou, Sophie Sananikone, Vojislav Tobias Westmoreland, Matthia Sabatelli, and Marco Zullich. A Study on the Faithfulness of Feature Attribution Explanations in Pruned Vision-Based Multi-Task Learning. In *Joint of the xAI 2025 Late-Breaking Work, Demos and Doctoral Consortium, LB/D/DC@ xAI 2025*, pages 121–128. CEUR Workshop Proceedings, 2025.

[DeYoung *et al.*, 2020] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4443–4458, 2020.

[Fisher *et al.*, 2019] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

[Fumagalli *et al.*, 2023] Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. SHAP-IQ: Unified approximation of any-order shapley interactions. *Advances in Neural Information Processing Systems*, 36:11515–11551, 2023.

[Goodman and Flaxman, 2017] Bryce Goodman and Seth Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

[Guo *et al.*, 2017] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[Hase *et al.*, 2021] Peter Hase, Harry Xie, and Mohit Bansal. The out-of-distribution problem in explainability and search methods for feature importance explanations. *Advances in neural information processing systems*, 34:3650–3666, 2021.

[Haufe *et al.*, 2026] Stefan Haufe, Rick Wilming, Benedict Clark, Rustam Zhumagambetov, Ahcène Boubekki, Jörg Martin, and Danny Panknin. Explainable AI needs formalization. *npj Artificial Intelligence*, 2(1):42, 2026.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Hooker *et al.*, 2019] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

[Jacovi and Goldberg, 2020] Alon Jacovi and Yoav Goldberg. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, 2020.

[James *et al.*, 2021] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An Introduction to Statistical Learning: with Applications in R. Springer Texts in Statistics. Springer US, 2021.

[Jang *et al.*, 2025] Hyeongwon Jang, Changhun Kim, and Eunho Yang. TIMING: Temporality-Aware Integrated Gradients for Time Series Explanation. In *Forty-second International Conference on Machine Learning*, 2025.

[Kim *et al.*, 2020] Siwon Kim, Jihun Yi, Eunji Kim, and Sungroh Yoon. Interpretation of NLP models through input marginalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3154–3167, 2020.

[Klein *et al.*, 2024] Lukas Klein, Carsten Lüth, Udo Schlegel, Till Bungert, Mennatallah El-Assady, and Paul Jäger. Navigating the maze of explainable ai: A systematic approach to evaluating methods and metrics. *Advances in Neural Information Processing Systems*, 37:67106–67146, 2024.

[Knoll *et al.*, 2025] Carsten Knoll, Julian Ullrich, Thomas Manjooran, Kilian Göller, Haadia Amjad, Ronald Tetzlaff, and Steffen Seitz. Xaiev–a framework for the evaluation of xai-algorithms for image classification. In *World Conference on Explainable Artificial Intelligence*, pages 250–263. Springer, 2025.

[Lapuschkin *et al.*, 2019] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1096, 2019.

[Levi *et al.*, 2022] Dan Levi, Liran Gispan, Niv Giladi, and Ethan Fetaya. Evaluating and calibrating uncertainty prediction in regression tasks. *Sensors*, 22(15):5540, 2022.

[Li *et al.*, 2016] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, 2016.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[Maathuis *et al.*, 2026] Henry Maathuis, Marcel Stalenhoef, Sieuwert Van Otterloo, Raymond Zwaal, Kees Van Montfort, and Danielle Sent. Designing effective explainable AI: a human-centered evaluation of explanation formats in financial decision-making. *Frontiers in Artificial Intelligence*, 9:1668029, 2026.

[Miró-Nicolau *et al.*, 2025] Miquel Miró-Nicolau, Antoni Jaume-i Capó, and Gabriel Moyà-Alcover. A comprehensive study on fidelity metrics for XAI. *Information Processing & Management*, 62(1):103900, 2025.

[Mishra *et al.*, 2020] Saumitra Mishra, Emmanouil Benetos, Bob LT Sturm, and Simon Dixon. Reliable local explanations for machine listening. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[Nauta *et al.*, 2023] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023.

[Nguyen *et al.*, 2015] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

[Pezeshki *et al.*, 2021] Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[Rieger and Hansen, 2020] Laura Rieger and Lars Kai Hansen. Irof: a low resource evaluation metric for explanation methods. *arXiv preprint arXiv:2003.08747*, 2020.

[Rong *et al.*, 2022] Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. *arXiv preprint arXiv:2202.00449*, 2022.

[Rony and Hassan, 2025] Md Main Uddin Rony and Naeemul Hassan. Between Consensus and Ambiguity: Expert Evaluation of LLM Explanations for Misleading Headlines. In *Proceedings of the Computation + Journalism Symposium (C+J '25)*, Miami, FL, December 2025. ACM.

[Saarela and Podgorelec, 2024] Mirka Saarela and Vili Podgorelec. Recent applications of explainable AI (XAI): A systematic literature review. *Applied Sciences*, 14(19):8884, 2024.

[Samek *et al.*, 2016] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

[Sarti *et al.*, 2024] Gabriele Sarti, Grzegorz Chrupała, Malvina Nissim, and Arianna Bisazza. Quantifying the plausibility of context reliance in neural machine translation. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024.

[Selvaraju *et al.*, 2020] Ramprasaath R Selvaraju, Michael Cogswell, Das Abhishek, Vedantam Ramakrishna, Parikh Devi, and Batra Dhruv. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2):336–359, 2020.

[Seth and Sankarapu, 2025] Pratinav Seth and Vinay Kumar Sankarapu. Bridging the Gap in XAI-Why Reliable Metrics Matter for Explainability and Compliance. In *EurIPS 2025 Workshop on Private AI Governance*, 2025.

[Song *et al.*, 2023] Junhwa Song, Keumgang Cha, and Junghoon Seo. On Pitfalls of RemOve-And-Retrain: Data Processing Inequality Perspective. *arXiv preprint arXiv:2304.13836*, 2023.

[Sridhar *et al.*, 2026] Shailesh Sridhar, Anton Xue, and Eric Wong. Missingness Bias Calibration in Feature Attribution Explanations. In *The Fourteenth International Conference on Learning Representations*, 2026.

[Sturmfels *et al.*, 2020] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the Impact of Feature Attribution Baselines. *Distill*, 2020. https://distill.pub/2020/attribution-baselines.

[Sundararajan *et al.*, 2017] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[Valdenegro-Toro and Mulye, 2024] Matias Valdenegro-Toro and Mihir Mulye. Sanity checks for explanation uncertainty. In *European Conference on Computer Vision*, pages 255–270. Springer, 2024.

[Wu *et al.*, 2021] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, 2021.

[Zhang *et al.*, 2025] Ziheng Zhang, Jianyang Gu, Arpita Chowdhury, Zheda Mai, David Carlyn, Tanya Berger-Wolf, Yu Su, and Wei-Lun Chao. Finer-cam: Spotting the difference reveals finer details for visual explanation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9611–9620, 2025.

[Zheng *et al.*, 2024] Xu Zheng, Farhad Shirani, Zhuomin Chen, Chaohao Lin, Wei Cheng, Wenbo Guo, and Dongsheng Luo. F-fidelity: A robust framework for faithfulness evaluation of explainable ai. *arXiv preprint arXiv:2410.02970*, 2024.