# Explanation Faithfulness is Alignment: A Unifying, and Geometric Perspective on Interpretability Evaluation

**Anonymous authors**
**Paper under double-blind review**

## Abstract

Interpretability researchers face a universal question: without access to ground truth explanation labels, how can the faithfulness of an explanation to its model be determined? Despite immense efforts to develop new evaluation methods, current approaches remain in a pre-paradigmatic state: fragmented, difficult to calibrate, and lacking cohesive theoretical grounding. Observing the lack of a unifying theory, we propose a *Generalised Explanation Faithfulness* (GEF) evaluative criterion centred on alignment that combines existing perturbation-based evaluations, eliminating the need for singular, task-specific evaluations. Complementing this unifying perspective, from a geometric point of view, we reveal a prevalent yet critical oversight in current evaluation practice: the failure to account for the learned geometry, and non-linear mapping present in the model, and explanation spaces. To solve this, we propose a general-purpose, threshold-free faithfulness evaluator that incorporates principles from differential geometry, facilitating evaluation agnostically across tasks, and explanation approaches. Through extensive cross-domain benchmarks on natural language processing, vision, and tabular tasks, we provide first-of-its-kind insights into the comparative performance of local linear approximations, and global feature visualisation methods, and the faithfulness of large language models (LLMs) as post-hoc explainers, and sparse autoencoders (SAEs). Our contributions are of substantial importance to the interpretability community, offering a principled, unified approach to evaluate the faithfulness of explanations. Code is available at `url`.

## 1 Introduction

Explaining the general behaviour, and predictions of machine learning (ML) models, particularly those functioning as black boxes, is critical, especially in domains such as healthcare, finance, and law. Driven by the urgency to comply with regulations like the EU AI Act, and GDPR, the eXplainable AI (XAI) research community has produced a plethora of explainability (or "interpretability") methods in recent years (Baehrens et al., 2010; Zeiler & Fergus, 2014a; Lundberg & Lee, 2017; Bykov et al., 2022; Fel et al., 2024; Lieberum et al., 2024). Simultaneously, the rise of large-scale, multi-tasking large language models (LLMs) (or "foundation models") (OpenAI, 2023; Mesnard et al., 2024) has spurred a significant shift in the interpretability landscape, with the mechanistic interpretability community producing a new generation of methods specifically designed to decompose, and reverse-engineer these increasingly black-box models (Elhage et al., 2022; Conmy et al., 2023; Bykov et al., 2023; Bills et al., 2023; Templeton et al., 2024). Despite this immense activity, consensus is lacking whether existing methods are of sufficient quality or trustworthy (Adebayo et al., 2018; Ghassemi et al., 2021; Bordt & von Luxburg, 2024; Bhattacharjee & von Luxburg, 2024). Since black-box models lack ground truth explanation labels (Bellido & Fiesler, 1993; Benitez et al., 1997), the universal question: "how faithful is the explanation to the model it seeks to explain?" remains difficult to answer. The prevalence of method-level disagreements within the XAI community (Neely et al., 2021; Watson et al., 2022; Krishna et al., 2022; Koenen & Wright, 2024) signals that the challenge of evaluation is still unsolved.

To *approximate* explanation quality (Agarwal et al., 2022b; Hedström et al., 2023b), researchers commonly use perturbation-based evaluations, where robustness (Montavon et al., 2018; Alvarez-Melis & Jaakkola,

2018b; Yeh et al., 2019; Nguyen & Martinez, 2020; Dasgupta et al., 2022), sensitivity (Adebayo et al., 2018; Hedström et al., 2024), and faithfulness methods (Bach et al., 2015; Samek et al., 2017; Ancona et al., 2018; Rieger & Hansen, 2020; Dasgupta et al., 2022; Bhatt et al., 2020; Rong et al., 2022) are well-embraced criteria to examine the relationship between explanation, and model outputs under perturbation, albeit with different emphases. Here, robustness, and sensitivity refer to making small or large perturbations (*e.g.,* adding noise to the input or randomising model parameters), and then measuring corresponding changes in the explanation output. Faithfulness measures how much the model's performance degrades when inputs, such as tokens or pixels, are cumulatively perturbed according to the explanation. Significant changes in model behaviour are interpreted as indicators of explanation faithfulness.

**Lack of Cohesive, Unified Theory.** Despite repeated attempts to define, and measure faithfulness (Montavon et al., 2018; Jacovi & Goldberg, 2020; Bhatt et al., 2020; Turpin et al., 2023; Lanham et al., 2023; Agarwal et al., 2024), fragmented mathematical terminology (Bordt & von Luxburg, 2024) makes it an ongoing, and unresolved matter. What exactly is explanation faithfulness, and how do robustness, and sensitivity evaluations differ from it? From a conceptual standpoint, although these evaluations share common steps—such as perturbing the inputs or the model parameters, measuring the effects, and interpreting the functional outcomes—the overwhelming number of evaluation methods (Lakkaraju et al., 2022), and the absence of a cohesive, unified theory makes it difficult to answer such seemingly straightforward questions. To better understand these evaluations' shared attributes, assumptions, and outcomes, a mathematical discussion is required. In Sec. 2, we propose a *unifying* perspective that formalises robustness, sensitivity, and faithfulness evaluations, providing a principled *Generalised Explanation Faithfulness* (GEF) criterion in Sec. 3, which substitutes singular evaluations.

**Ignoring the Impact of Geometry.** Alongside the lack of a cohesive, unified theory, most perturbation-based evaluations (Sec. 2.1.2)—while well-intended, and intuitive—often rely on overly simplistic assumptions about the underlying geometry of both model, and explanation spaces. When perturbations are introduced, the functional outcomes of models, and explanations are frequently compared using direct distance measures or correlation coefficients (Alvarez-Melis & Jaakkola, 2018b; Yeh et al., 2019; Ancona et al., 2018; Bhatt et al., 2020; Nguyen & Martinez, 2020; Agarwal et al., 2022a), which, from a *geometric* perspective, overlooks a simple yet critical fact: that a uniform perturbation such as input noise or parameter shifts can affect non-linear systems in highly non-uniform ways. Only in a linear system, the perturbation effects would be uniform. By neglecting the geometric differences (*e.g.,* differences in curvatures) between the model, and explanation spaces, current evaluations risk misjudging how faithful the explanation is *w.r.t.* its underlying model. For fair measurements across non-linear systems, perturbation effects must be measured in the context of the distinct geometric structures of the respective manifolds (Lee, 2012). In Sec. 4, we examine these geometric factors, and introduce a solution that accounts for the intrinsic geometry of each space, thereby improving current evaluation practice.

To address these research gaps in unified theory (Sec. 3), and methodological neglect of the impact of geometry (Sec. 4), our work offers a fourfold contribution.

- **(C1)** In the absence of cohesive theory, we systematise common steps in numerous perturbation-based evaluation algorithms (Sec. 2), and provide a unifying criterion for robustness, sensitivity, and faithfulness evaluations (Sec. 3).
- **(C2)** To account for geometric discrepancies in many evaluation methods, we propose a solution based on differential geometry that ensures fair measurements across non-linear mappings (Sec. 4).
- **(C3)** Recognising the need for a general-purpose, threshold-free, task-agnostic faithfulness evaluator, we provide `GEF` and `Fast-GEF`, serving different compute budgets (Sec. 5)[1]
- **(C4)** Observing the lack of cross-domain comparative insights on the faithfulness across explanation approaches such as local, global, LLM-as-explainer, and Sparse Autoencoder (SAE)—we perform extensive experiments across vision, tabular, and natural language processing (NLP) tasks (Sec. 6).

[1]Code to be released upon publication of the non-anonymous version of the paper.

Our contributions carry substantial importance to the interpretability (and related) communities. The reliability of individual explanation methods, and XAI as a field is already under hot debate, thus it is not only timely but relevant to provide clarity on the matter of explanation faithfulness. As we enter a new era of interpretability, it is of utmost importance to revisit, and revise existing evaluation approaches: both from a *unifying*, and *geometric* perspective. We hope this work will pave the way for a more reliable selection, and development of both existing, and upcoming interpretability methods.

## 2 Explainable AI Evaluation: Where Are We Now?

In this section, we present the scope of this work. We begin by outlining preliminaries to estimate explanation quality, followed by a description of the general workflow of perturbation-based evaluation. Finally, we mathematically formalise robustness, sensitivity, and faithfulness evaluation, revealing critical assumptions essential for their validity. Complete notation tables are provided in Appendix A.9.

### 2.1 Preliminaries

Let $f_\theta : \mathcal{X} \to \mathcal{Y}$ be a differentiable neural network (NN) that maps inputs $\boldsymbol{x} \in \mathbb{R}^D$ to predictions $\boldsymbol{y} \in \mathbb{R}^C$ of $C$ classes. By functionally mapping $\boldsymbol{x} \in \mathcal{X}$ to $\boldsymbol{y} \in \mathcal{Y}$ with parameters $\theta$ such that $\boldsymbol{y} = f(\boldsymbol{x}; \theta)$, a trained model $f_\theta$ is obtained, which we refer to as $f$. Here, $\theta$ includes weights, and biases, and exists in parameter space $\Theta \in \mathbb{R}^W$ for a fixed architecture in function space $f_\theta \in \mathcal{F}$. Here, $f$ may represent NN architectures ranging from simple feedforward MLPs, CNNs to highly parameterised transformers.

**Local Explanations.** To interpret a specific model prediction (*i.e.,* logit) $y := y_c$ of a class $c \in [1, 2, \dots C]$, we may employ a *local* method. Let $\phi_L : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^V$ be a local explanation function that takes an input, and logit pair, and assigns importance scores to a subset (or all) of its input features such that $\boldsymbol{e} = \phi_L(f, \boldsymbol{x}, y; \lambda)$, where $\boldsymbol{e} \in \mathbb{R}^V$ is the explanation output, parameterised by $\lambda$. A broad variety of local explanation approaches fall within the scope of our work, *e.g.,* gradient-based (Simonyan & Zisserman, 2015; Smilkov et al., 2017; Sundararajan et al., 2017; Bykov et al., 2022; Krishna et al., 2023; Selvaraju et al., 2020), back-propagation-based (Bach et al., 2015; Shrikumar et al., 2017), model-agnostic (Zeiler & Fergus, 2014a; Lundberg & Lee, 2017), local surrogate (Ribeiro et al., 2016a), attention-based (Chefer et al., 2021; Covert et al., 2022), or prototypical explanation methods (Simonyan & Zisserman, 2015). More recent approaches (Krishna et al., 2023; Kroeger et al., 2023) that leverage separate LLMs as the explanation function $\phi$ to interpret local predictions in a post-hoc manner, are also within the scope of this work.

**Global Explanations.** To study the model $f$ from a *global* point of view, producing an explanation independent of a specific instance $\boldsymbol{x}$. Here, a global explanation method $\phi_G : \mathcal{F} \times \mathcal{Y} \to \mathbb{R}^V$ takes a trained model $f$, and generate an explanation $\boldsymbol{e} \in \mathbb{R}^V$ for specific neural activation associated with a target class $c$, represented by logit $y$ such that $\boldsymbol{e} = \phi_G(f, y; \kappa)$, where $\phi_G$ is parameterised by $\kappa$. Here $\phi_G$ may be variants of activation-maximisation (or "feature visualisation") which provide either natural, or synthetic data points of maximal activation (Berkes & Wiskott, 2006; Erhan et al., 2009; Olah et al., 2017; Nguyen, 2020; Fel et al., 2024). Recently, trained sparse autoencoders (SAEs) Bricken et al. (2023); Lieberum et al. (2024); Huben et al. (2024) have emerged as an alternative formulation for $\phi_G$ to provide interpretable "monosemantic" feature encodings of intermediate model representations.

For convenience, we let $\phi \in \mathcal{E}$ denote $\phi_L$, and $\phi_G$ although they formally reside in different spaces. Furthermore, to avoid label leakage (Jethani et al., 2023), we use the predicted class (and not the true class) to generate the explanation $\boldsymbol{e}$.

#### 2.1.1 Estimate Explanation Quality

Without ground truth explanation labels, the task of estimating the quality of an explanation $\phi$ is non-trivial. To approximate explanation quality, researchers rely on metric-based heuristics (or "metrics"). Following Hedström et al. (2023a), we define a general evaluation function $\Psi_\tau : \mathcal{E} \times \mathcal{X} \times \mathcal{F} \times \mathcal{Y} \to \mathbb{R}$:

$$q = \Psi(\phi, \boldsymbol{x}, f, y; \tau) \tag{1}$$

which returns a quality estimate $q \in \mathbb{R}$, indicating the quality of a given explanation, parameterised by $\tau$. When global explanations $\phi_G$ are evaluated, $\boldsymbol{x}$ is omitted from Eq. 1. Unless required, we omit hyperparameters $\tau, \lambda, \kappa, \zeta$ for notational convenience.

#### 2.1.2 Related Works

Within approaches that evaluate explanation quality by approximation, we concentrate on those that examine the *functional relationship* between the explanation, and the model through means of perturbation, *i.e.,* assessing qualities such as robustness, sensitivity, and faithfulness. These are briefly introduced below, and mathematically formalised in Sec. 2.3.

**Robustness.** Robustness (also referred to as "continuity", and "stability") methods evaluate the explanation function's resilience to infinitesimal input noise, and is a widely used evaluation technique (Yeh et al., 2019; Montavon et al., 2018; Alvarez-Melis & Jaakkola, 2018b; Nguyen & Martinez, 2020; Agarwal et al., 2022a; Dasgupta et al., 2022). Most commonly, robustness is evaluated by first perturbing an input sample, then generating the explanation for the perturbed input, and finally comparing this explanation to the original explanation. Higher similarity between the original, and perturbed explanation indicates higher quality. Existing robustness measures differ in how noise is applied to the input (*e.g.,* using a Gaussian (Alvarez-Melis & Jaakkola, 2018b; Yeh et al., 2019) or a uniform distribution (Agarwal et al., 2022a)), and how explanation similarity is measured (*e.g.,* Yeh et al. (2019) computes difference with *Monte-Carlo* sampling, and Alvarez-Melis & Jaakkola (2018b); Agarwal et al. (2022a) rely on variants of *Lipschitz* constant).

**Sensitivity.** Sensitivity (or "randomisation") methods (Adebayo et al., 2018; Hedström et al., 2024) act complementary to robustness, and assesses a critical, indisputable evaluative quality: that the explanation function $\phi$ should be sensitive to randomisation of model parameters. Existing sensitivity measures differ in how the change in the explanation outputs is measured (*e.g.,* Adebayo et al. (2018) relies on *Structural Similarity Index* (SSIM), and Hedström et al. (2024) uses discrete entropy calculations), and how perturbation is applied (*e.g.,* Adebayo et al. (2018) randomises model parameters layer-by-layer in a *top-down* fashion, and Hedström et al. (2024) uses *bottom-up* or *full* parameter randomisation). The sensitivity criterion asks that the explanation should change significantly when the model parameters are randomised, whether layer-by-layer (Adebayo et al., 2018) or entirely (Hedström et al., 2024).

**Faithfulness.** Faithfulness (or "fidelity") methods (Bach et al., 2015; Samek et al., 2017; Montavon et al., 2018; Ancona et al., 2018; Rieger & Hansen, 2020; Dasgupta et al., 2022; Bhatt et al., 2020; Rong et al., 2022; Atanasova et al., 2023; Blücher et al., 2024; Chuang et al., 2024) evaluate explanations by gradually perturbing the input based on the importance of pixels or tokens indicated by the explanation, and observing the resulting degradation in model performance. Mehthods differ in how model responses are reported (with logits (Alvarez-Melis & Jaakkola, 2018a; Yeh et al., 2019; Bhatt et al., 2020) or softmax probabilities (Montavon et al., 2018; Ancona et al., 2018; Rieger & Hansen, 2020; Nguyen & Martinez, 2020; Dasgupta et al., 2022; Rong et al., 2022)), how perturbations are ordered (ascending (Arya et al., 2019; Nguyen & Martinez, 2020) or descending (Bach et al., 2015; Samek et al., 2017; Rong et al., 2022)), and in the general approach to perturbation (whether using single-pixel changes (Bach et al., 2015), patch-based masking with a constant value (Samek et al., 2017), or linear interpolation (Rong et al., 2022)). Faithfulness methods typically aggregate model responses into a single quality estimate, such as AUC (Bach et al., 2015; Samek et al., 2017; Rong et al., 2022). For faithfulness to be considered fulfilled, the model's performance should rapidly decrease as perturbations are applied—the steeper the degradation, the higher the explanation quality.

Beyond approximation techniques, interpretability researchers have explored alternative ways to evaluate explanation quality, such as using human judgment (Zeiler & Fergus, 2014b; Ribeiro et al., 2016b), restricting tasks to synthetic or toy environments (Guidotti, 2021; Carmichael & Scheirer, 2023). Both approaches lack scalability, and generalisability to real-world scenarios, and are not covered in this work.

### 2.2 Perturbation-Based Evaluation

A key observation is that robustness, sensitivity, and faithfulness evaluations generally rely on three common steps. First, a perturbation is applied to either the input (*e.g.,* by adding infinitesimal noise) or the model parameters (*e.g.,* by randomisation). Second, the effect of the perturbation is measured on the output of either the explanation function $\phi$ or the model $f$. Third, an interpretation is made to assess whether this change in functional outputs is acceptable given a criterion, such as requiring the distance in explanation outputs to be small when the perturbation is small. We refer to Fig. 1 for an illustration.

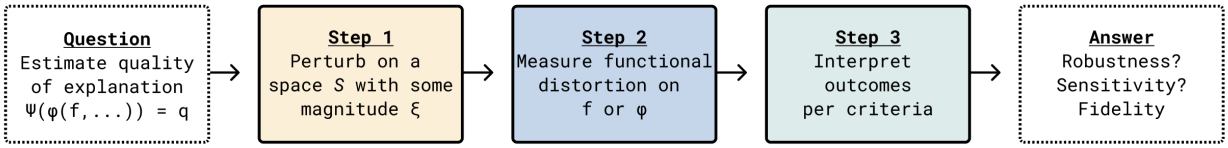



Figure 1: An overview of the "perturb, measure, and interpret" evaluation methodology (Sec. 2.3).

To facilitate mathematical unification (Sec. 4), and further insights (Sec. 5), we next formalise the three steps of perturbation-based evaluation. Therefore, some general notation for perturbation (Eqs 2-3), and measurement (Def. 1) is introduced. By systematising XAI evaluation, we can advance our conceptual understanding, especially in clarifying how existing methods differ, and what attributes are shared.

#### 2.2.1 Step 1. Perturbation

First, a perturbation is initiated. This is typically done either on the model parameter space in large magnitudes, *e.g.,* by randomising weights, or on the input in small magnitudes, *e.g.,* by adding Gaussian noise. Alternatively, perturbations can be applied cumulatively, such as by masking pixels or regions of pixels, or by replacing tokens in textual inputs. To accommodate diverse evaluation methods across different data modalities, we follow Hedström et al. (2023a), and define a general perturbation function that can be applied on any real-valued space $\mathcal{S} \subseteq \{\mathcal{X}, \Theta, \mathcal{Y}\}$. Let $\mathcal{P}_\mathcal{S} : \mathcal{S} \to \mathcal{S}$ be a perturbation function of $\boldsymbol{s} \in \mathcal{S}$ with parameters $\omega \in \mathbb{R}$:

$$\mathcal{P}_\mathcal{S}(\boldsymbol{s}; \omega) = \hat{\boldsymbol{s}}, \tag{2}$$

where $\forall \hat{\boldsymbol{s}}, \boldsymbol{s} \in \mathcal{S}$, and $\hat{\boldsymbol{s}} \neq \boldsymbol{s}$. For brevity, we may omit $\omega$ such that $\mathcal{P}_\mathcal{S}(\boldsymbol{s}) := \mathcal{P}_\mathcal{S}(\boldsymbol{s}; \omega)$. With Eq. 2, we may, *e.g.,* generate a perturbed instance $\hat{\boldsymbol{s}}$ with input perturbation, , *i.e.,* $\hat{\boldsymbol{x}} = \mathcal{P}_\mathcal{X}(\boldsymbol{x})$ or model parameter randomisation, *i.e.,* $\hat{\theta} = \mathcal{P}_\Theta(\theta)$. Since robustness, sensitivity, and faithfulness evaluations require distinct perturbation magnitudes, we let $\xi$ denote the difference between $\boldsymbol{s}$, and $\hat{\boldsymbol{s}}$:

$$\delta(\boldsymbol{s}, \hat{\boldsymbol{s}}) = \xi, \tag{3}$$

where $\delta : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ is a general discrepancy function, *e.g.,* an $\ell_p$-norm, cosine distance or Pearson correlation.

#### 2.2.2 Step 2. Measurement

Following perturbation, as a second step, perturbation impacts are measured on relevant functions. Common approaches include measuring the distance between explanation outputs or recording the change in model responses under random or cumulative masking guided by the explanation output. We define a general approach to measure the perturbation impact on a separate function (*e.g.,* the impact of input perturbation on the model function) below.

**Definition 1 (Functional Distortion)** *Let $\boldsymbol{s}, \hat{\boldsymbol{s}} \in \mathcal{S}$ denote instances in space $\mathcal{S} \subseteq \{\mathcal{X}, \Theta, \mathcal{Y}\}$, before, and after perturbation, respectively. Let $k : \mathcal{S} \to \mathcal{H}$ denote a separate function that maps $\boldsymbol{s}, \hat{\boldsymbol{s}}$ to a distinct space $\mathcal{H} \subseteq \{\mathcal{F}, \mathcal{E}\}$ from $\mathcal{S}$. Then, perturbation impact in function $k$ is measured by functional distortion $\boldsymbol{D}_k : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ as follows:*

$$\boldsymbol{D}_k(\boldsymbol{s}, \hat{\boldsymbol{s}}) = \delta(k(\boldsymbol{s}), k(\hat{\boldsymbol{s}})), \tag{4}$$

*where $k(\boldsymbol{s}) = h$ with $h \in \mathcal{H}$, and $\delta : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$.*

**Model, and Explanation Distortion.** With Def. 1, we can flexibly apply perturbation in one space, and then evaluate the effect in a different space[2]. For example, assume we have applied perturbation on the input space, *i.e.,* $\hat{\boldsymbol{x}} = \mathcal{P}_{\mathcal{X}}(\boldsymbol{x})$ (Eq. 2), and therefore have two instances $\boldsymbol{x}$, and $\hat{\boldsymbol{x}}$. Then, to measure the perturbation impact on the model function $f$, we follow Def. 1, and set $k = f$ where $h = \boldsymbol{y}$. Evaluating $\boldsymbol{D}_f(\boldsymbol{x}, \hat{\boldsymbol{x}})$ from Eq. 4 effectively means that we compare model evaluations on perturbed, and non-perturbed inputs, *i.e.,* $\delta(y, \hat{y})$ with $\hat{y} = f_c(\hat{\boldsymbol{x}}; \theta)$ for the same class $c$. Alternatively, to measure perturbation impacts on the explanation function $\phi$, we set $k = \phi$ where $h = \boldsymbol{e}$. Evaluating $\boldsymbol{D}_\phi(\boldsymbol{x}, \hat{\boldsymbol{x}})$, practically means that we compute $\delta(\boldsymbol{e}, \hat{\boldsymbol{e}})$ where $\hat{\boldsymbol{e}} = \phi(\hat{\boldsymbol{x}}, \ldots)$ is the explanation *w.r.t.* perturbed input $\hat{\boldsymbol{x}}$. For comparability, $\hat{\boldsymbol{e}}$ is generated *w.r.t.* the same class $c$ as its non-perturbed counterpart $\boldsymbol{e}$. Similarly, to compute functional distortion after parameter perturbation, *i.e.,* $\hat{\theta} = \mathcal{P}_\Theta(\theta)$, we compute $\boldsymbol{D}_f(\theta, \hat{\theta})$, and $\boldsymbol{D}_\phi(\theta, \hat{\theta})$ using logit $\hat{y} = f(\boldsymbol{x}; \hat{\theta})$, and explanation $\hat{\boldsymbol{e}} = \phi(f_{\hat{\theta}}, \ldots)$, respectively. To generalise the notation across different perturbation types, we let $\boldsymbol{D}_f$, and $\boldsymbol{D}_\phi$ denote the model, and explanation distortion quantities, respectively.

#### 2.2.3 Step 3. Interpretation

In the final step of the evaluation workflow (Fig. 1), the distortion quantities are examined separately according to their evaluative criteria. For example, if robustness is evaluated, generally low values for $\boldsymbol{D}_\phi$ are expected, assuming perturbation magnitude $\xi$ is small. Conversely, if sensitivity is evaluated, high values for $\boldsymbol{D}_\phi$ are expected, assuming perturbation magnitude $\xi$ is large. If faithfulness is evaluated, model distortions $\boldsymbol{D}_f$ are anticipated to increase as perturbation is cumulatively applied according to the explanation function output. Notably, a key limitation of this step is the need for researchers to thresholds for distinguishing between low-, and high-quality evaluation outcomes, which can be manipulated (Wickstrøm et al., 2024).

### 2.3 Formalising Robustness, Sensitivity, and Faithfulness

Equipped with a general perturbation function $\mathcal{P}_{\mathcal{S}}$ (Eq. 2), and its magnitude $\xi$ (Eq. 3) as well as a measure to compute functional distortion of the explanation, and model functions (Def. 1), we combine a wide variety of existing evaluation techniques into general formalisations of robustness, sensitivity, and faithfulness evaluation methods (*cf.* Sec. 2.1.1) Based on these three main criteria (Defs. 2-4), we show that the validity of each explanation criterion critically depends on fulfilling a separate, implicit model assumption (Ass. 1-3).

We proceed by presenting a definition of explanation robustness, absorbing the spirit of numerous existing robustness methods[3] (Yeh et al., 2019; Montavon et al., 2018; Alvarez-Melis & Jaakkola, 2018b; Nguyen & Martinez, 2020; Agarwal et al., 2022a).

**Definition 2 (Explanation Robustness)** *Let $\hat{\boldsymbol{x}} = \mathcal{P}_{\mathcal{X}}(\boldsymbol{x})$ be a perturbed input, and $\Psi^{RO}$ be a quality estimator to yield robustness estimates $q^{RO} \in \mathbb{R}$ such that $q^{RO} = \boldsymbol{D}_\phi(\boldsymbol{x}, \hat{\boldsymbol{x}})$. Given thresholds $\alpha, \varepsilon^{RO}_{\boldsymbol{D}_\phi} \in \mathbb{R}^+$, an explanation function $\phi$ is robust if the perturbation magnitude $\xi^{RO} \leq \alpha$:*

$$q^{RO} \leq \varepsilon^{RO}_{\boldsymbol{D}_\phi}. \tag{5}$$

For an explanation function $\phi$ to be considered robust, the estimator $\Psi^{RO}$ should yield low values, *i.e.,* $q^{RO} \leq \varepsilon^{RO}_{\boldsymbol{D}_\phi}$, reflecting minor differences between the original explanation $\boldsymbol{e}$, and the perturbed explanation $\hat{\boldsymbol{e}}$. Since the stability expectations of the explanation function $\phi$ are dictated by the robustness of $f$ (Yeh et al., 2019; Chalasani et al., 2020; Agarwal et al., 2022a; Tan & Tian, 2023), it would be false to expect $\phi$ to exhibit robustness if its underlying model is not robust. Consequently, the validity of Eq. 5 depends on the fulfillment of model robustness (Ass. 1).

**Assumption 1 (Model Robustness)** *Given an input perturbation $\mathcal{P}_{\mathcal{X}}$ of magnitude $\xi^{RO}$, and thresholds $\alpha, \varepsilon^{RO}_{\boldsymbol{D}_f} \in \mathbb{R}^+$, $\xi^{RO} \leq \alpha$, the model distortion (Eq. 4) is bounded by $\boldsymbol{D}_f(\boldsymbol{x}, \hat{\boldsymbol{x}}) \leq \varepsilon^{RO}_{\boldsymbol{D}_f}$.*

[2]While both the perturbation magnitude $\xi$ (Eq. 3), and the distortion $\boldsymbol{D}_k$ (Eq. 4) use the discrepancy function $\delta(\cdot, \cdot)$, their outputs differ. Notably, $\xi$ expresses the discrepancy between the original, and perturbed instance, and $\boldsymbol{D}_k$ measures the discrepancy in a distinct space from the perturbation space.

[3]Some algorithmic details are omitted in the definition. For completeness, mathematical definitions are provided for each evaluation method in Appendix A.4.5.

In line with works of Adebayo et al. (2018); Hedström et al. (2024), we define explanation sensitivity in the following.

**Definition 3 (Explanation Sensitivity)** *Let $\hat{\theta} = \mathcal{P}_\Theta(\theta)$ create a model $f_{\hat{\theta}}$ with perturbed parameters, and $\Psi^{SE}$ be a quality estimator that yields sensitivity estimates $q^{SE} \in \mathbb{R}$ such that $q^{SE} = \boldsymbol{D}_\phi(\theta, \hat{\theta})$. Given thresholds $\alpha, \varepsilon^{SE}_{\boldsymbol{D}_\phi} \in \mathbb{R}^+$, an explanation function $\phi$ is sensitive if the perturbation magnitude $\xi^{SE} > \alpha$:*

$$q^{SE} > \varepsilon^{SE}_{\boldsymbol{D}_\phi}. \tag{6}$$

For $\phi$ to be considered sensitive to randomness, the differences between explanations should be substantial, meaning $\Psi^{SE}$ yields high estimates, *i.e.,* $q^{SE} > \varepsilon^{SE}_{\boldsymbol{D}_f}$, reflecting significant discrepancies between $\boldsymbol{e}$, and $\hat{\boldsymbol{e}}$. This expectation that $q^{SE}$ should be large is based on the assumption that the model responded strongly to the perturbation. Similar to how explanation robustness depends on the stability of $f$, the emphasis on a large $q^{SE}$ assumes a different model response. Therefore, the validity of the sensitivity evaluation (Eq. 6) depends on model sensitivity (Ass. 2).

**Assumption 2 (Model Sensitivity)** *Given a parameter perturbation $\mathcal{P}_\Theta$ of magnitude $\xi^{SE}$, and thresholds $\alpha, \varepsilon^{SE}_{\boldsymbol{D}_f} \in \mathbb{R}^+$, $\xi^{SE} > \alpha$, the model distortion (Eq. 4) is bounded by $\boldsymbol{D}_f(\theta, \hat{\theta}) > \varepsilon^{SE}_{\boldsymbol{D}_f}$.*

With various existing interpretations of explanation faithfulness (Sec. 2.1.2), we focus on common criteria to combine these interpretations into a single definition below.

**Definition 4 (Explanation Faithfulness)** *Let $\hat{\boldsymbol{x}}^z = \mathcal{P}_\mathcal{X}(\boldsymbol{x}; z)$ denote the input after the $z^{th}$ perturbation for $z \in [1, Z]$, where $\mathcal{P}_\mathcal{X}$ progressively masks the top-$z$ features according to the indices given by argmax$(\boldsymbol{e})$, with perturbation magnitudes $\xi_z$ satisfying $\xi_1 \leq \xi_2 \leq \ldots \leq \xi_Z$. A quality estimator $\Psi^{FA}$ yields a vector of faithfulness estimates $\boldsymbol{q}^{FA} \in \mathbb{R}^Z$ with entries $q^{FA}_z = f(\hat{\boldsymbol{x}}^z, \theta)$. The overall faithfulness score $q^{FA} \in \mathbb{R}$ is obtained by aggregating these estimates via a function $\nu : \mathbb{R}^Z \to \mathbb{R}$:*

$$q^{FA} = \nu(\hat{\boldsymbol{q}}^{FA}). \tag{7}$$

When $\nu$ is defined using AUC, a faithful explanation is expected to produce low aggregated scores $q^{FA}$ (Eq. 7). The conventional expectation in faithfulness evaluation (Bach et al., 2015; Samek et al., 2017; Rong et al., 2022) is that significant distortions should occur early, as the "more important features" are removed first. To ensure that the faithfulness score is solely driven by the quality of $\phi$, and not by other factors, such as out-of-distribution samples (OOD) (Hase et al., 2021; Hesse et al., 2024), non-linear feature effects or artefacts introduced by cumulative perturbations (Hooker et al., 2019; Brunke et al., 2020; Hase et al., 2021; Rong et al., 2022; Brocki & Chung, 2022), the model distortion to these perturbations should be monotonically non-decreasing, *i.e.,* satisfies model faithfulness (Ass. 3).

**Assumption 3 (Model Faithfulness)** *Given $Z$ cumulative perturbations $\mathcal{P}_\mathcal{X}$ of magnitudes $\xi_z$ with $\xi_1 \leq \xi_2 \leq \cdots \leq \xi_Z$ the corresponding model distortions (Eq. 4) are: $\boldsymbol{D}^1_f \leq \boldsymbol{D}^2_f \leq \cdots \leq \boldsymbol{D}^Z_f$ with $\boldsymbol{D}^z_f = \boldsymbol{D}_f(\boldsymbol{x}, \hat{\boldsymbol{x}}^z)$.*

### 2.4 Model Assumptions in Practice

Evaluations under Defs. 2-4 typically assume that model distortions are proportional to perturbation magnitudes, *i.e.,* that larger perturbations lead to greater distortions, and smaller perturbations result in lesser distortions. This prompts a natural question: with commonly used perturbation techniques for evaluating robustness (*e.g.,* additive Gaussian noise), sensitivity (*e.g.,* layer-wise randomisation), and faithfulness (*e.g.,* cumulative input masking, is this assumption valid in practice? In Appendix A.5, we extensively analyse the extent to which Ass. 1-3 hold versus fail across various explanation methods, and NN models. Notably, we find that Ass. 1-3 are systematically violated in practice. While this is expected due to the inherent non-linearity of the models, it has significant consequences for the validity of existing evaluations (Defs. 2-4). Evaluation outcomes may be misleading when explanation robustness is enforced for models that fundamentally lack it (Chalasani et al., 2020; Tan & Tian, 2023; Agarwal et al., 2022a), or when faithfulness scores are attributed to explanation quality without considering OOD scenarios (Hase et al., 2021; Hesse et al., 2024). In Sec. 3.3, we propose a mitigation strategy to address this issue.

# 3 A Unifying Perspective

With clear definitions of robustness, sensitivity, and faithfulness evaluations (Sec. 2.3), we may now explore their shared attributes, and outcomes. In the following, we discuss the unifying aspects of these evaluations, and introduce a novel definition of GEF, which integrates the underlying concepts of these three singular evaluations into a comprehensive criterion of explanation quality.

## 3.1 Unifying Attributes

Upon formalising the evaluation criteria (Defs. 3-4), a notable observation is that robustness, sensitivity, and faithfulness exhibit common attributes. Each of the evaluative criteria (1) introduces *perturbations* of a specific magnitude $\xi$, (2) *measures* the functional effects, and (3) *interprets* these effects, *i.e.,* the quality estimate $q$, considering an explicit or implicit boundary. The evaluation is performed under the respective *assumptions* about the model responses $\boldsymbol{D}_f$. We refer to Table 1 for a summary of these findings.

Table 1: A concise overview of the attributes of the *robustness*, *sensitivity*, and *faithfulness* evaluations. The last row presents our *unified* evaluation (Def 5), whose theory and practical implementation are described in Sec.3 and Sec.5, respectively.

| Evaluation ($\Psi$) (Defs. 2-4) | Step 1. Perturbation; Magnitude (Eq. 2-3) | Step 2. Measurement (Def. 4) | Step 3. Interpretation (Eq. 5-7) | Model Assumptions (Ass. 1-3) |
|---|---|---|---|---|
| Robustness ($\Psi^{RO}$) | $\mathcal{P}_{\mathcal{X}}(\boldsymbol{x})$; $\xi^{RO} \leq \alpha$ | $\boldsymbol{D}_\phi(\boldsymbol{x}, \hat{\boldsymbol{x}})$ | $q^{RO} \leq \varepsilon^{RO}_{\boldsymbol{D}_\phi}$ | $\boldsymbol{D}_f \leq \varepsilon^{RO}_{\boldsymbol{D}_f}$ |
| Sensitivity ($\Psi^{SE}$) | $\mathcal{P}_{\Theta}(\theta)$; $\xi^{SE} > \alpha$ | $\boldsymbol{D}_\phi(\theta, \hat{\theta})$ | $q^{SE} > \varepsilon^{SE}_{\boldsymbol{D}_\phi}$ | $\boldsymbol{D}_f > \varepsilon^{SE}_{\boldsymbol{D}_f}$ |
| Faithfulness ($\Psi^{FA}$) | $\mathcal{P}_{\mathcal{X}}(\boldsymbol{x}, z)$; $\xi^{FA}_1 \leq \xi^{FA}_2 \leq \cdots \leq \xi^{FA}_Z$ | $f(\hat{\boldsymbol{x}}^z, \theta)$ | $q^{FA} = \nu(\hat{\boldsymbol{q}}^{FA})$ | $\boldsymbol{D}^1_f \leq \boldsymbol{D}^2_f \leq \cdots \leq \boldsymbol{D}^Z_f$ |
| Unified ($\Psi^{GEF}$) | $\mathcal{P}_{\Theta}(\theta, z)$; $\xi^{GEF}_1 \leq \xi^{GEF}_2 \leq \cdots \leq \xi^{GEF}_Z$ | $\boldsymbol{D}_\phi(\theta, \hat{\theta})$, and $\boldsymbol{D}_f(\theta, \hat{\theta})$ | $\rho(\boldsymbol{d}_f, \boldsymbol{d}_\phi) \approx 1$ | None |

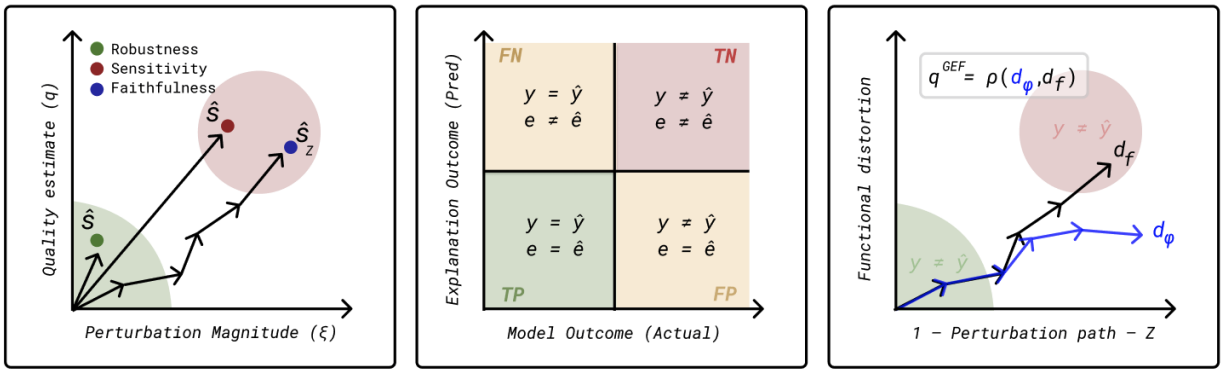
In Fig. 2 (A), we illustrate these theoretical similarities on a graph, with axes corresponding to the shared attributes $\xi$, and $q$. Here, we can observe that robustness evaluation (*green*) involves minimal perturbation with a small difference in expected explanation output (or low $q$). Sensitivity (*red*) employs substantial perturbation, expecting a significant difference in explanation output (or high $q$). Faithfulness (*blue*) uses cumulative perturbation of $Z$ steps, evaluating the corresponding variations in model output. By placing the different perspective of explanation quality onto Fig. 2 (A), and thereafter examining the positions of the post-perturbed instances $\hat{\boldsymbol{s}} \in \mathcal{S}$, we can advance our understanding of how the criteria relate to one another: specifically, that diverse evaluation methods can be unified under a shared conceptual framework.

## 3.2 Unifying Outcomes

Another point of unification emerges when considering the outcomes of these evaluation criteria, and how they interact in practice. In Fig. 2 (B), similar to the traditional confusion matrix (in ML) or contingency table (in statistics), we provide a visual representation of the possible model, and explanation outcomes, post-perturbation. Although model, and explanation outcomes are typically continuous in reality—for clarity conceptually, we classify them into four distinct quadrants: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). A key benefit of *discretising* evaluation outcomes in this way, is that we crucially can distinguish between *aligned*, and *misaligned* explanation behaviour:

- *Aligned outcomes (TP + TN).* The green quadrant represents outcomes where the explanation, and model agree, indicating explanation robustness, *i.e.,* $\boldsymbol{e} = \hat{\boldsymbol{e}}$, and $y = \hat{y}$, and satisfying Ass. 1. Conversely, the red quadrant contains outcomes where both explanation, and model outputs differ, reflecting explanation sensitivity, *i.e.,* , $\boldsymbol{e} \neq \hat{\boldsymbol{e}}$, and $y \neq \hat{y}$, and satisfying Ass. 2. Explanation faithfulness is achieved when evaluation outcomes are aligned over $Z$ steps (Ass. 3).

- *Misaligned outcomes (FP + FN).* The orange quadrants highlight misalignment between $\phi$, and $f$. The top-left quadrant shows explanation dissimilarity despite prediction stability (*i.e.,* , $y = \hat{y}$, and $\boldsymbol{e} \neq \hat{\boldsymbol{e}}$), failing Ass. 2. The bottom-right quadrant shows explanation similarity despite a prediction change (*i.e.,* , $y \neq \hat{y}$, and $\boldsymbol{e} = \hat{\boldsymbol{e}}$), failing Ass. 1.




(A) Unifying Attributes (B) Unifying Outcomes (C) Unifying Criteria

Figure 2: Intuition behind the relationship between robustness, sensitivity, and faithfulness evaluations. (A) illustrates the shared attributes, *i.e.,* perturbation magnitude $\xi$, and quality estimate $q$ that unifies robustness (*green*), sensitivity (*red*), and faithfulness (*blue*) evaluations. (B) displays a confusion matrix of *discretised* model, and explanation outcomes, with green, and red quadrants indicating aligned behaviour, and orange quadrants showing misalignment. (C) shows our proposed GEF criteria (Def. 5) which measures explanation to model alignment over diverse evaluation perspectives.

**Explanation Faithfulness is Alignment.** Our analysis reveals that evaluation using Defs. 2-4, fundamentally concerns the *alignment* between the explanation, and the model's behaviour, whether across single (Defs. 2-3) or multiple (Def. 4) perturbation steps. A key observation is that existing robustness, and sensitivity measures provide a limited view of isolated model conditions: robustness evaluates alignment when the model's predictions remain stable (TP quadrant), while sensitivity evaluates alignment when predictions change (TN quadrant). Faithfulness (Def. 4) evaluates alignment over $Z$ steps, assuming non-decreasing, monotonic model responses under cumulative perturbations (Ass. 3). These *singular* perspectives require strict adherence to specific model conditions, and consequently fail to evaluate the full behaviour of the explanation function. Next, we propose a more comprehensive criterion for faithfulness evaluation.

### 3.3 Unifying Criteria

We extend, and generalise the current faithfulness criterion (Def. 4) by integrating the robustness, and sensitivity evaluations into a combined criterion, independent of restrictive model assumptions. Using a series of $Z$ perturbations, GEF (Def. 5) measures explanation alignment across a spectrum of model outcomes—from cases where model predictions remain consistent, *i.e.,* $y = \hat{y}$ to those where predictions diverge, *i.e.,* $y \neq \hat{y}$. In this way, a comprehensive, *generalised* definition of explanation faithfulness is obtained.

**Definition 5 (Generalised Explanation Faithfulness)** *Let* $\boldsymbol{d}_f = [\boldsymbol{D}_f^1, \boldsymbol{D}_f^2, \ldots, \boldsymbol{D}_f^Z]$ *and* $_\phi = [\boldsymbol{D}_\phi^1, \boldsymbol{D}_\phi^2, \ldots, \boldsymbol{D}_\phi^Z]$ *be the model, and explanation distortion vectors, where* $\boldsymbol{D}_f^z$*, and* $\boldsymbol{D}_\phi^z$ *are distortion quantities of the* $z^{th}$ *step along a perturbation path* $z \in [1, Z]$*, from robustness at* $z = 1$ *to sensitivity at* $z = Z$ *such that* $\forall y, \hat{y} \in Y$*:*

$$(z = 1 : y = \hat{y}) \quad \textit{and} \quad (z = Z : y \neq \hat{y}),$$

*where* $\hat{y}$*, and* $y$ *are perturbed versus unperturbed model outputs, respectively. Let* $\Psi^{GEF}$ *be a quality estimator that yields estimates* $q^{GEF} \in \mathbb{R}$ *via the correlation coefficient* $\rho : \mathbb{R}^Z \times \mathbb{R}^Z \to \mathbb{R}$ *such that* $q^{GEF} = \rho(\boldsymbol{d}_{f,\phi})$*. An explanation function* $\phi \in \mathcal{E}$ *is faithful to* $f \in \mathcal{F}$ *if:*

$$q^{GEF} \approx 1. \tag{8}$$

With Eq. 8, we define a quality estimator $\Psi^{GEF}$ that yields values ranging between $[-1, 1]$, with a value of 1 implying perfect generalised faithfulness, 0 suggesting an absence of it, and $-1$ an inverse relationship. GEF estimation is therefore *threshold-free* in the sense that the correlation coefficient directly indicates the quality of the explanation, eliminating the need for arbitrary cut-offs. Moreover, Def. 5 applies to a wide range of

explanation functions, as discussed in Sec. 2.1. The choice of $\rho$, and perturbation applied to construct the distortion vectors depends on the practical implementation (Sec. 5.2). Note that in Def. 5, we implicitly rely on predicted class $c$ to generate the perturbed logit $\hat{y}$ as the target for the explanation, and model distortion. In Appendix A.1.2, we discuss a broader application of GEF where the targets $\hat{y}$, and $y$ are replaced by any $c$-th neuron within a layer $l \in [1, L]$ of a feed-forward model.

**Remarks.** Our definition shares similarities with faithfulness estimation (Def. 4) in that it assesses explanation quality along a perturbation path. However, it fundamentally differs by focusing on *general alignment* rather than a specific scenario of measuring the *magnitude* of model response to *cumulative input* perturbation. A key benefit of our proposal is that we use the model distortion to *anchor* the expectations of the explanation distortion, and as such, eliminate the need to rely on arbitrary thresholds. In this way, the evaluation will be grounded in the exact functional response of the model, and thus resilient to OOD scenarios: expecting small explanation distortions only when model distortions are small, and vice-versa.

**Theoretical Benefits.** A good faithfulness measure should assign low scores to unfaithful explanations, and high scores to faithful explanations. In Appendix A.1.3, we prove that a linear model $f = \boldsymbol{\theta}\boldsymbol{x} + c$ where $\boldsymbol{\theta}$ acts as the explanation, attains a perfect faithfulness score, *i.e.,* $q^{GEF} = 1$ with GEF. Conversely, unfaithful explanations are penalised by GEF. For instance, constant explanations that generate no distortion, *i.e.,* $\boldsymbol{D}_\phi(\boldsymbol{e}, \hat{\boldsymbol{e}}) = 0$, pass the conventional robustness test $q^{RO} \leq \varepsilon^{RO}_{\boldsymbol{D}_\phi}$ (Def. 2), but fails in the GEF estimate. Similarly, random explanations (*e.g.,* generated by uniform sampling, *i.e.,* $\hat{\boldsymbol{e}}_i \sim \mathcal{U}(0, 1)$) produce maximal distortion (Binder et al., 2022), and thus generally pass the sensitivity test, *i.e.,* $q^{SE} > \varepsilon^{SE}_{\boldsymbol{D}_\phi}$ (Def. 3) but fails in the GEF estimate. We prove both cases in Appendix A.1.1, and provide empirical evidence in Sec. A.8.3.

## 4 A Geometric Perspective

With a advanced understanding of faithfulness evaluation (Sec. 3), we can perform a more systematic, and rigorous study of explanation behaviour. Without assuming the restricted model conditions (Ass. 1-3) are met, which are often violated in practice (Appendix A.5), we gain a clearer view of the *true* explanation function behaviour. We can formalise questions such as: do explanation functions in real-world evaluation scenarios align or misalign with their model? In the following, we empirically examine this question across common explanation methods (Sec. 2.1) for different NN models. Guided by differential geometry, we provide theoretical considerations on the impact of geometry.

### 4.1 Explanation Alignment Patterns

To empirically analyse whether explanation functions are aligned with their underlying model, we study how the distortions of various local, and global explanation functions, and models change under perturbation. Here, we use additive Gaussian noise, *i.e.,* , $\nu_i \sim \mathcal{N}(0, \sigma)$ to generate perturbed inputs $\hat{\boldsymbol{x}}_i = \boldsymbol{x} + \nu_i$, with $\sigma$ increasing until the model behaves randomly (*i.e.,* accuracy $= 1/C$) with $Z = 10$ perturbation steps. We refer to Tab. 2 in Sec. 6 for details regarding datasets, and models, and to Appendix A.6 for extended results.

Fig. 3 (top, and bottom) presents the results, which can be interpreted as a continuous analogue of the confusion matrix presented in Sec. 3.2. The scatter points, coloured by perturbation magnitude, reveal that $\phi$, and $f$ rarely align fully. Instead, the relative alignment varies with both the model, and the explanation functions. Here, we include *GradientSHAP* (SHAP-G) (Lundberg & Lee, 2017), and *Gradient* (GRAD) (Morch et al., 1995; Baehrens et al., 2010)) on top, and bottom rows in Fig. 3, respectively, with more results in Appendix A.6). The overlapping contours (*e.g.,* Avila results in Fig. 3) underscore a simple but nonetheless systematically overlooked aspect of perturbation-based evaluation (Sec. 2.3): that a uniform perturbation of its inputs may affect highly non-linear systems in a non-uniform way. If the effects were uniform, the system would likely be linear.

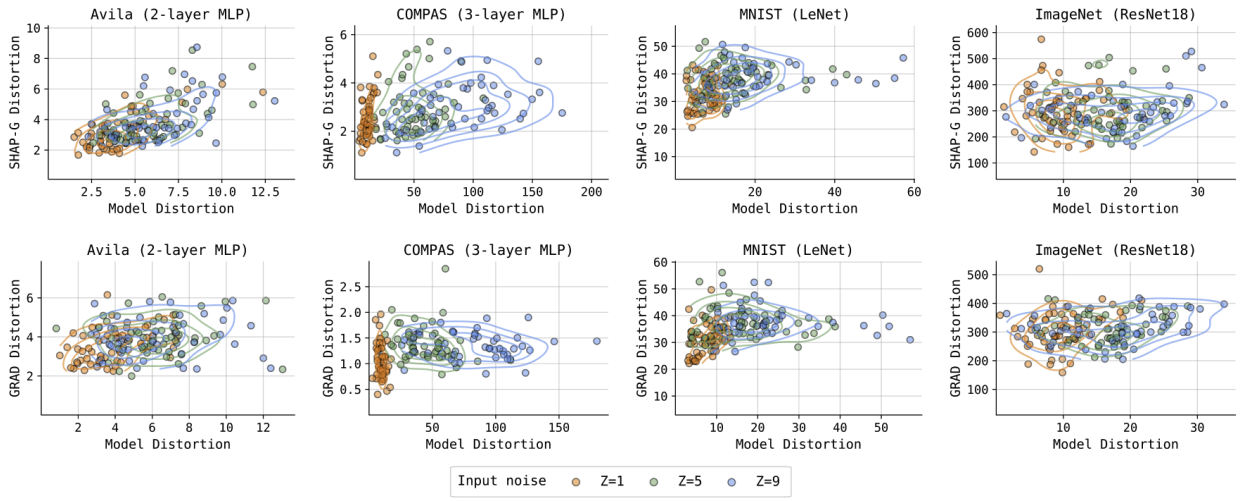



Figure 3: Model (x-axis), and explanation distortions (y-axis) under varying levels of additive Gaussian input noise for vision, and tabular tasks. The scatter points represent individual samples, coloured by perturbation magnitude (z=1, z=5, z=9), with overlapping contours highlighting the relative alignment patterns. The top, and bottom rows represent *GradientSHAP* (SHAP-G), and *Gradient* (GRAD) explanations, respectively.

### 4.2 The Impact of Geometry

By considering the geometric nature of the spaces these functions inhabit, we can understand the observed misalignment better. In differential geometry, each space—whether it is the model output space $\mathcal{Y}$ or the explanation output space $\mathcal{E}$—can be viewed as a manifold with its unique geometric characteristics (Lee, 2012). When a perturbation is applied, a new point on these manifolds may be accessed, and then, when functional distortion (Def. 1) is computed in each space, we are effectively computing a distance between two points on each manifold. For example, with model parameter perturbation, *i.e.,* $\hat{\theta} = \mathcal{P}_\Theta(\theta)$, see Eq. 2, we obtain perturbed model outputs $\hat{\boldsymbol{y}}$ (or, a logit $\hat{y}$) given $\hat{\boldsymbol{y}} = f_{\hat{\theta}}(\boldsymbol{x})$. From this, model distortion (Def. 1) is calculated using, *e.g.,* Euclidean distance between the original, and the perturbed instance. A key observation is that, when distances in two different spaces are functionally compared, we ignore the fact that manifolds have their own separate geometric characteristics which are distinct, with distances in one space not necessarily reflecting equivalent distances in another. In direct comparisons such as correlation (Ancona et al., 2018; Bhatt et al., 2020) or Lipschitz calculations (Alvarez-Melis & Jaakkola, 2018a; Agarwal et al., 2022a), a globally flat metric is assumed. We refer to Fig. 4 (A), and (B) for an illustration of the problem of ignoring the impact of geometry. As a result, the estimation of explanation quality may be misleading.

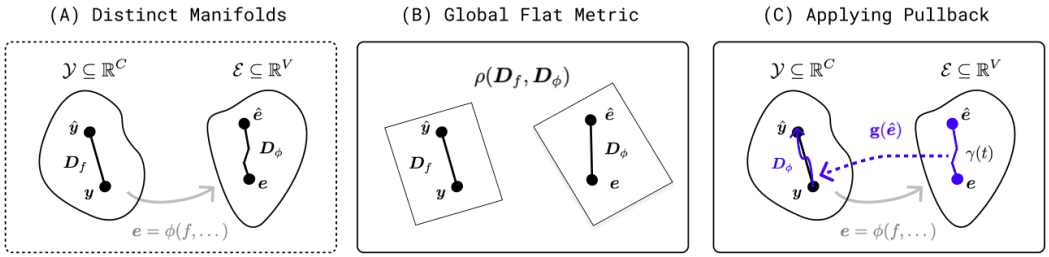



Figure 4: An illustration of the relationship between the manifolds of the model, and explanation. (A) shows how the explanation function maps between the model, and explanation spaces, $\mathcal{Y}$, and $\mathcal{E}$. (B) displays the problem with directly comparing distortions across spaces, assuming a flat metric. (C) illustrates the pullback operation using metric tensor $\mathbf{g}$ to adjust distortions in $\mathcal{E}$ for comparison in $\mathcal{Y}$.

### 4.3 Reconciling Geometric Discrepancies

To enable a geometrically sound comparison between explanation, and model distortion, the aim is to recompute $\boldsymbol{D}_\phi$ to incorporate the non-linear mappings used in generating explanations. This can be achieved by mapping the distortion from the explanation space $\mathcal{E}$ to the model space $\mathcal{Y}$, effectively "pulling back" the measured distance into $\mathcal{Y}$ (see Ch. 11 of Lee (2012) for further details). Guided by differential geometry, we create a metric tensor $\mathbf{g}$ that serves as this pullback onto $\mathcal{Y}$. This process is illustrated as $\mathbf{g}(\hat{\boldsymbol{e}})$ in Fig. 5 (C).

To construct the metric tensor $\mathbf{g}$, we consider an infinitesimal neighbourhood around the parameter perturbation $\theta + du$, for a fixed $\boldsymbol{x}$, and $y$. By applying a first-order Taylor expansion in this neighbourhood, we obtain:

$$\phi(f_{\theta+du}, \ldots) \approx \boldsymbol{e} + J_f du, \tag{9}$$

where $J_f \in \mathbb{R}^{V \times C}$ is the Jacobian for fixed input $\boldsymbol{x}$, with elements $J_{i,j} = \frac{\partial e_i}{\partial f_j}$. We use $f_j$ as shorthand for $f_j(\boldsymbol{x})$. Effectively, $\theta + du$ yields a new perturbed model $f_{\hat{\theta}}$ which is computed with model parameter perturbation (Sec. 5.1). With Eq. 9, we can compute the elements of the pullback tensor $\mathbf{g} \in \mathbb{R}^{V \times V}$ as the sum of the resulting changes in each explanation element $e_v$ *w.r.t.* the changes in each model element $f_j$:

$$g_{i,j}(\boldsymbol{e}) = \sum_{v=1}^{V} \frac{\partial e_v}{\partial f_i} \frac{\partial e_v}{\partial f_j}. \tag{10}$$

Thus, Eq. 10 captures the sensitivity of $\phi$ to model output changes, with $\mathbf{g}$ corresponding to the squared Jacobian $\mathbf{g} = J_f^\top J_f$. Multiplying with the pullback facilitates accurate measurement of distances in the *pseudo-Riemannian manifold* $(\mathcal{Y}, \mathbf{g})$ of space $\mathcal{Y}$ with metric $\mathbf{g}$. With the pullback metric tensor $\mathbf{g}$ in place, we can measure explanation distortion that is equivalent to computing the path length under the induced parameter changes in the "pulled-back" space:

$$\boldsymbol{D}_\phi := L(\gamma) = \int_0^1 \frac{d\gamma(t)}{dt}^\top g_{\gamma(t)} \frac{d\gamma(t)}{dt} dt, \tag{11}$$

where $\gamma(t)$ is a path between endpoints $\boldsymbol{e}, \hat{\boldsymbol{e}} \in \mathcal{E}$ derived from the original, and perturbed models, respectively. Here, $t$ denotes the step size. With Eq. 11, we replace $\boldsymbol{D}_\phi = \delta(\boldsymbol{e}, \hat{\boldsymbol{e}})$ (Def. 1) with the *total accumulated* distortion along the path, with longer paths corresponding to greater distortions. Upon taking this geometric perspective, we can study $\mathcal{Y}$ using extrinsically-defined geometry, contrasting it with the simpler assumption of a flat, intrinsic Euclidean metric. As a result, $\boldsymbol{D}_\phi$, and $\boldsymbol{D}_f$ are more fairly compared in the same space.

## 5 Method: From Theory to Practice

While our unified theory (Sec. 3), and solution to reconcile geometric discrepancies in measurement (Sec. 4), provide first steps towards resolving issues in perturbation-based evaluation, many practical concerns have been raised regarding the choice of perturbation. In this section, we describe how to reliably translate our theory (Def. 5) to practice—we propose a general-purpose, task-agnostic perturbation technique based on model parameter scaling (Sec. 5.1), and introduce the full evaluation algorithm (Sec. 5.2).

### 5.1 Selecting Perturbation Strategy

While all perturbation-based evaluations require parameterisation, input-based perturbation (Defs. 2 and 4), has proven particularly challenging to calibrate (Sturmfels et al., 2020; Haug et al., 2021). Without ground truth labels, selecting parameters such as patch size, pixel, or token replacement strategies is typically based on researchers' judgment. It has been shown that small changes to input parameters can drastically affect evaluation outcomes (Brunke et al., 2020; Brocki & Chung, 2022; Rong et al., 2022; Blücher et al., 2024). Input perturbation is not only impractical from a practitioner's standpoint but also compromises impartiality—if parameters must be adjusted for each model, and dataset, how can task-specific confounds be controlled? In Appendix A.5.1, we provide empirical evidence for the existence of confounds in faithfulness evaluations (Def. 4).

Researchers need a general-purpose, dataset, and architecture-agnostic perturbation strategy that facilitates evaluation across explanation method types (*e.g.,* local, and global methods), and magnitudes (*i.e.,* $\xi$). Following Bykov et al. (2022), we propose the following:

**Model Parameter Scaling.** *Introduce perturbations* $\forall z \in [1, Z]$ *by scaling parameters* $\theta \in \mathbb{R}^W$ *with Gaussian noise* $\eta_i \sim \mathcal{N}(\mathbf{1}, \sigma_z^2 \mathbb{1})$*, and* $\sigma_z^2 \in \mathbb{R}^+$ *such that* $\hat{\theta}_z = \theta \cdot \eta_i$*, yielding a perturbed model* $f_{\hat{\theta}_z}$*.*

By systematically perturbing model parameters instead of the input, from low to high magnitudes with incremental increases of $\sigma_z^2$, ranging from robustness at $z = 1$ to sensitivity at $z = Z$, explanation behaviour is evaluated comprehensively, and agnostically across tasks. With $\xi := \delta(f(\boldsymbol{x}), f_{\hat{\theta}_z}(\boldsymbol{x}))$ (Eq. 3), we can measure the perturbation impact at each $z^{th}$ step so that robustness, *i.e.,* $y = \hat{y}$, and sensitivity, *i.e.,* $y \neq \hat{y}$ criteria are fulfilled (Def. 5). Our approach contrasts with Adebayo et al. (2018), which proposes layer-wise *randomisation* in a top-down order, an approach that faces methodological concerns (Sundararajan & Taly, 2018; Binder et al., 2022; Kokhlikyan et al., 2021; Yona & Greenfeld, 2021). For an illustration of how model parameter scaling affects the classifier's decision boundary, we refer to Fig. 1 of Bykov et al. (2022).

### 5.2 Introducing GEF Evaluator

From an algorithmic perspective, we evaluate GEF in three main steps. First, given a model, and a test set of input-output pairs, we generate perturbed models $f_{\hat{\theta}_1}, \ldots, f_{\hat{\theta}_Z}$ given $Z$ sets of parameters $\hat{\theta}_1, \ldots, \hat{\theta}_Z$ along a perturbation path (see Algo. 1, line 6). Then, for each model $f_{\hat{\theta}_z}$, we compute the model, and explanation distortion quantities, *i.e.,* $\boldsymbol{D}_f^z$, and $\boldsymbol{D}_\phi^z$, using the pullback tensor $\mathbf{g}$ (lines 7, 9, and 10). Finally, distortion vectors are constructed, and correlated with $\rho(\boldsymbol{d}_f, \boldsymbol{d}_\phi)$ (lines 14, and 15). Due to the stochastic nature of model perturbation, we repeat this process $M$ times to average out the effects. We refer to Fig. 5, and Algo. 1, for an overview of the steps involved, presented from a practitioner's perspective.

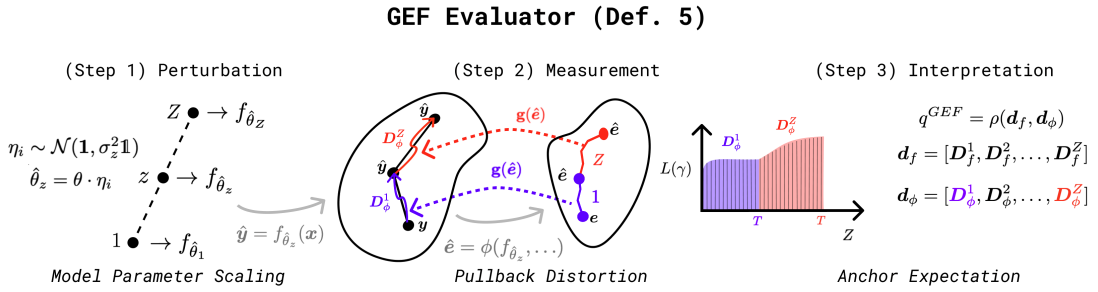



Figure 5: Three-step *GEF* evaluation (Algo. 1) for estimating GEF (Def. 5). First, a perturbed model is obtained by model parameter scaling (Sec. 5.1). Second, distortions are computed via the pullback operation (Sec. 4.3), integrating along the path (*blue*, and *red* areas) to assess the impact on $\phi$. Third, distortion vectors are correlated for the final GEF estimate (Sec. 3.3)

**Practical Benefits.** Our proposed evaluation (Algo. 1) provides several practical benefits. First, *anchoring*, negates the need to rely on arbitrary thresholds in evaluation, *e.g.,* when determining a permissible value for the evaluations themselves (Eqs. 5, 6, and 7) or what perturbation magnitude leads to model alignment for a particular task. Second, perturbing via *model parameter scaling*, at varying intensities combines distinct criteria of explanation quality into a single unified evaluation metric (Sec. 3.3) that is agnostic to the data, model, and explanation approach. Third, constructing the *pullback*, provides a geometrically grounded measurement of GEF, capturing the true functional impacts of the explanation *w.r.t.* its model.

Unless stated otherwise, we use *Euclidean distance* for $\delta$ in the functional distortion calculations (Def. 1), and define $\rho$ using *Spearman Rank Correlation*, assessing the degree of monotonic relationship between the distortion quantities. For the experiments, we set $Z = 5$ (see discussion of the influence of $Z$ in Appendix A.1.4) but can be flexibly updated in the open-source implementation[4]. In Appendix A.2, we provide comprehensive details on the implementation, including how to generate the perturbation path (line

[4]To be released upon publication of the non-anonymous version of the paper.

**Algorithm 1** *GEF Evaluator*

1: **Require:** Model $f$, explanation function $\phi$, input-prediction pairs $\boldsymbol{x}, y \in \boldsymbol{X}, \boldsymbol{Y}$ with $\boldsymbol{X} \subseteq \mathcal{X}, \boldsymbol{Y} \subseteq \mathcal{Y}$
2: **Parameters:** Integers $Z, M, T, K$, correlation measure $\rho$
3: **for** $\boldsymbol{x}, y$ in range$(\boldsymbol{X}, \boldsymbol{Y})$ **do**
4: $\quad \boldsymbol{e} \leftarrow \phi(f, y, \ldots)$
5: $\quad$ **for** $z$ in range$(Z)$ **do**
6: $\quad\quad \hat{y} \leftarrow f_{\hat{\theta}_z}(\boldsymbol{x})$
7: $\quad\quad \boldsymbol{D}_f^z \leftarrow \delta(y, \hat{y})$ // Eq. (4)
8: $\quad\quad$ **if** *Fast-GEF* **then**
9: $\quad\quad\quad \boldsymbol{D}_\phi^z \leftarrow \delta(\boldsymbol{e}, \hat{\boldsymbol{e}})$ with $\hat{\boldsymbol{e}} \leftarrow \phi(f_{\hat{\theta}_z}, \ldots)$ // Eq. (4)
10: $\quad\quad$ **else**
11: $\quad\quad\quad \boldsymbol{D}_\phi^z \leftarrow$ `compute_path_length`$(f_{\hat{\theta}_z}, \boldsymbol{x}, y, T, K)$ // Eq. (11)
12: $\quad\quad$ **end if**
13: $\quad$ **end for**
14: $\quad$ **Construct:** $\boldsymbol{d}_f \leftarrow [\boldsymbol{D}_f^1, \boldsymbol{D}_f^2, \ldots, \boldsymbol{D}_f^Z]$, and $_\phi \leftarrow [\boldsymbol{D}_\phi^1, \boldsymbol{D}_\phi^2, \ldots, \boldsymbol{D}_\phi^Z]$
15: $\quad$ **Calculate:** $q^{GEF} \leftarrow \rho(\boldsymbol{d}_{f,\phi})$
16: $\quad$ **Return:** $q^{GEF}$
17: **end for**

2), and how to tune parameters (line 2). This also includes information on `compute_path_length` (line 11); where we follow an approximation procedure outlined Eq. 20, and 21 in Appendix A.2.1.

### 5.3 Balancing Computational Constraints

While the pullback operation ensures a fair geometric comparison of distortion quantities, its use of high-dimensional Jacobian calculations, and integral steps (Eq. 10) also increases computational demands. To accommodate evaluation contexts involving large model architectures or high-dimensional explanations, we offer an alternative method. For a faster yet *naive* approximation of explanation quality, we omit the pullback operation, and instead define $\boldsymbol{D}_\phi$ according to Eq. 4. This approach, entitled `Fast-GEF`, is less computationally demanding, and complements the *exact* approach with pullback, entitled `GEF`, providing a geometrically grounded quality estimate.

**Choosing between `GEF` or `Fast-GEF`.** Users can choose between these methods based on their specific computational constraints and demands for accurate quality estimations. We recommend `GEF` wherever possible due to its ability to account for manifold-specific distortions. However, `Fast-GEF` provides a computationally efficient alternative that is suitable for large-scale tasks or resource-constrained environments. Empirical results show that while the `GEF` or `Fast-GEF` may diverge in individual estimates (Appendix A.6), they often shares categorical rankings of explanation methods (Appendix A.8.2).

## 6 Experiments

Our experiments aim to answer the following questions:

**(Q1)** Are unified, GEF evaluations more empirically reliable than competitive singular approaches?

**(Q2)** How does generalised faithfulness of local, and global explanation methods compare across domains?

**(Q3)** How faithful are LLMs as a top-$K$ token post-hoc explainers for NLP classifications?

**(Q4)** Does expanding the capacity of SAEs improve the faithfulness of their explanations?

To answer these questions, we selected a diverse set of datasets, model architectures on tabular, vision, and NLP classification tasks. See Table 2 for an overview. Our experiments evaluate the faithfulness of various explanation approaches, as detailed below. Code for reproducing the experiments is available at `url`.

Table 2: Datasets, and models, with references in Appendix A.4. Multiple models for a dataset are separated by a semicolon.

| Modality | Dataset (n. classes) | Model (size) | Acc. % | Source | Expl. dim | Task |
|---|---|---|---|---|---|---|
| Text | SMS Spam (2) | BERT-TINY FT (4.4M) | 98.0 | HF | 128 | Spam |
| | IMDb (2) | Pythia FT (7.6M); Gemma-2 (2B) | 86.4; 95.6 | HF | 512 | Sentiment |
| | SST-2 (2) | BERT-tiny FT (4.4M) | 98.0 | HF | 59 | Sentiment |
| Vision | ImageNet-1K (1000) | ResNet18 (11.7M) | 89.1 | Torchvision | 50176 | Object |
| | PATH (9) | MedCNN (235.2K) | 84.3 | Local | 784 | Pathology |
| | Derma (7) | MedCNN (234.9K) | 73.2 | Local | 784 | Dermatology |
| | MNIST (10) | LeNet (61.7K) | 97.7 | Local | 784 | Digit |
| | fMNIST (10) | LeNet (61.7K) | 87.7 | Local | 784 | Fashion |
| Tabular | Adult (2) | 3-layer MLP (11.7K); LR (28) | 84.6; 83.3 | OpenXAI | 13 | Income |
| | Compas (2) | 3-layer MLP (11.1K); LR (16) | 85.0; 85.3 | OpenXAI | 7 | Recidivism |
| | Avila (12) | 2-layer MLP (3.5K) | 80.8 | Local | 10 | Letter |

**Global, and Local Methods.** For global methods, we include feature visualisation techniques with different regularisation, and optimization procedures: *Deep-Viz* (DV) (Yosinski et al., 2015), *Magnitude Constrained Optimization* (MACO) (Fel et al., 2024), and *Fourier preconditioning* (FO) (Olah et al., 2017). Optimization steps are set to 50, 100, and 250, otherwise, default values are used as provided in the respective publications ((Fel et al., 2024), and (Nguyen, 2020)). For local methods, two variants of *Layer-wise Relevance Propagation* (LRP), the $\varepsilon$*-rule* (LRP-$\varepsilon$) (Bach et al., 2015) with $\varepsilon = 1e^{-6}$, and the $z^+$*-rule* (LRP-$z^+$) (Montavon et al., 2017) are used. Also, several gradient-based approaches: *Gradient* (GRAD) (Morch et al., 1995; Baehrens et al., 2010), *Saliency* (SAL) (Simonyan et al., 2014), *Input*$\times$*Gradient* (IXG) (Shrikumar et al., 2016), *GradCAM* (G-CAM) (Selvaraju et al., 2020), *Guided Backpropagation* (GBPG) (Springenberg et al., 2015), *SmoothGrad* (SMG) (Smilkov et al., 2017) with 10 noisy samples, and noise level $0.1/(x_{\max} - x_{\min})$, *Integrated Gradients* (INT-G) (Sundararajan et al., 2017) with 10 iterations, and zero baseline. For NLP tasks, we evaluate *LayerIntegratedGradients* (L-INTG) explanations *w.r.t.* the first embedding layer. Two Shapley-based algorithms (Lundberg & Lee, 2017) are included: *GradientSHAP* (SHAP-G) with 10 samples, and *PartitionShap* (SHAP-P) for NLP tasks.

**LLM-x Methods.** An emerging research area in explainability uses separate LLMs to generate post-hoc attributions for important features of a given model (Bills et al., 2023; Kroeger et al., 2023; Krishna et al., 2023; Amara et al., 2024). We create LLM-x explanations by prompting `Gemma-2B-IT` (Mesnard et al., 2024) to rank the top-$K$ most important tokens given a textual input, which is then parsed, decoded, and mapped to input tokens, producing binary attribution vectors. LLM prompts describe the model's classification task, and prediction certainty before, and after model perturbation (Sec. 5.1). The temperature is set to 0 for deterministic outputs. Varying synonyms, the order of tokens, and the number of top-$K$ values $\{5, 10\}$ contribute to the robustness of our findings. The full explanation methodology is described in Appendix A.4.4 with an illustration in Fig. A.2.

**Sparse Autoencoders.** SAEs have lately come forth as an interpretability method for understanding the internal representations of LLMs (Templeton et al., 2024; Huben et al., 2024). In our work, we generate SAE explanations using `Gemma-Scope` (Lieberum et al., 2024), pretrained on the residual block representations of the `Gemma-2-2B` model. Explanations are saved for all 26 layers at both 16K, and 65K widths. Given the sparsity of the explanation vectors, we use cosine distance to compute explanation distortion, defined as $1 - \frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{u}||\mathbf{v}|}$, as it effectively measures similarity regardless of vector magnitude. Appendix A.4.4 provides a detailed description of the generation process for each SAE explanation.

**Control variants** We also evaluate the faithfulness of two control variants: a random explanation (RAN) sampled from a uniform distribution, $\hat{\boldsymbol{e}}_i \sim \mathcal{U}(1, 0)$, and a top-$K$ control variant (RAN-$K$) with $K$ non-zero attributions, each equal to 1. Unless specified, all experiments evaluate 250 explanations for the logit of the predicted class. For comparability, global, and local explanations are normalised by dividing the attribution map by the square root of its average second-moment estimate (Eq. 23) (Binder et al., 2022), with further explanation preprocessing details provided in Appendix A.4.4. For metric implementation, and meta-evaluation, we use the `Quantus` (Hedström et al., 2023b), and `MetaQuantus` (Hedström et al., 2023a)

libraries, respectively. Further experimental details for Q1, Q2, and Q3 are provided in Appendix A.8.1, A.8.4, and A.8.5, respectively.

### 6.1 Measuring Empirical Reliability

To investigate the empirical reliability of GEF evaluations compared to singular approaches, we perform meta-evaluation, which is the practice of evaluating the evaluation method itself. To this end, we adopt the meta-evaluation methodology from Hedström et al. (2023a), which bypasses the lack of ground truth labels by focusing on *metric consistency* ("does this evaluation method produce similar results under consistent conditions?"). For this, two practical meta-evaluative tests are performed: the Input Perturbation Test (IPT), and the Model Perturbation Test (MPT). Each test returns a meta-consistency (MC) score (see Eq. 22), which ranges between $[0, 1]$, with higher values indicating greater reliability. Full meta-evaluation scoring methodology is provided in Appendix A.3. As a sanity check, we also show in Appendix A.8.3 that our proposed evaluators assign low scores to different random control variants, where other metrics fail to do so.

**Setup.** We benchmark three evaluation methods per criterion. In the robustness category, we include *Relative Input Stability* (RIS), *Relative Representation Stability* (RRS), *Relative Ouput Stability* (ROS) (Agarwal et al., 2022a). In the sensitivity category, we include *Model Parameter Randomisation Test* (MPRT) (Adebayo et al., 2018), *Smooth MPRT* (sMPRT), and *Efficient MPRT* (EMPRT) (Hedström et al., 2024). In the faithfulness category, we include *Faithfulness Correlation* (FC) (Bhatt et al., 2020), *Pixel-Flipping* (PF) (Bach et al., 2015), and *Region-Perturbation* (Samek et al., 2017). All metrics are mathematically described in Appendix A.4.5. To ensure comparability with the original publication (Hedström et al., 2023a), we run meta-evaluation on the same set of tasks, which includes ImageNet, MNIST and fMNIST datasets with ResNets and LeNets architectures. Each metric evaluates GRAD, SAL, G-CAM, SHAP-G explanations. Further results, and details are provided in Tabs. A.1, and A.2, and Appendix A.8.5.

**Results.** Fig. 6 (A) shows that our proposed unified methods (`GEF` and `Fast-GEF`) achieve the highest *overall* MC scores, averaged over both MPT, and IPT tests. Our unified methods significantly outperform the most comparable evaluation approach, the faithfulness metrics, which also use $Z$ perturbation steps, with average MC scores of 0.733 compared to 0.601. Although no evaluation method achieves a perfect score (*i.e.,* MC=1), the unified methods still perform comparably to robustness metrics, and surpass sensitivity metrics, with average scores of 0.727, and 0.673, respectively. These results are encouraging as they show that unified methods can achieve high reliability, even when explanation behaviour is evaluated under multiple model conditions, unlike robustness, and sensitivity metrics that focus on a single perspective. While the ROS metric has the highest individual score, this is not statistically significant, and it only offers a limited view of explanation quality. Fig. 6 (B) shows that unified metrics excel in MPT, while robustness metrics perform slightly better in IPT. These score differences correspond to robustness metrics using input perturbations, and unified metrics relying on model perturbations. Further details are provided in Appendix A.8.4.

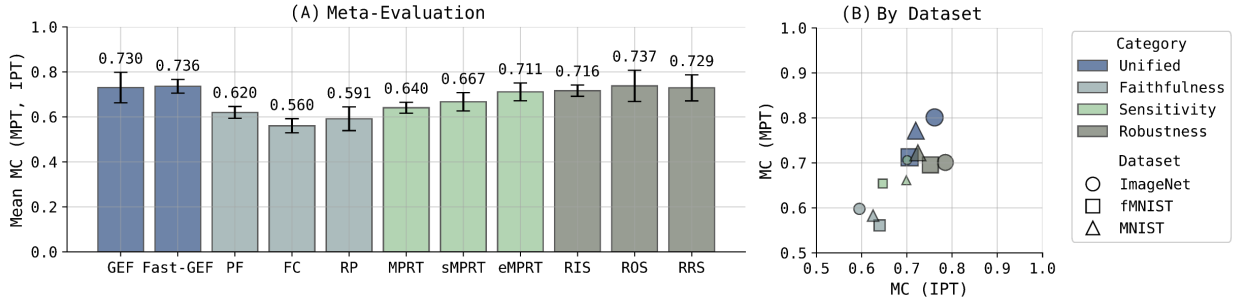



Figure 6: Meta-evaluation, and comparison to established explanation evaluation methods. (A) shows the mean MC scores across MPT, and IPT, aggregated over all datasets, with the error bars showing the standard deviation. (B) displays MC scores aggregated by the test type, and dataset, where the size of the scatter point denotes the standard deviation. `GEF` scores are computed for fMNIST, and MNIST datasets due to computational constraints.

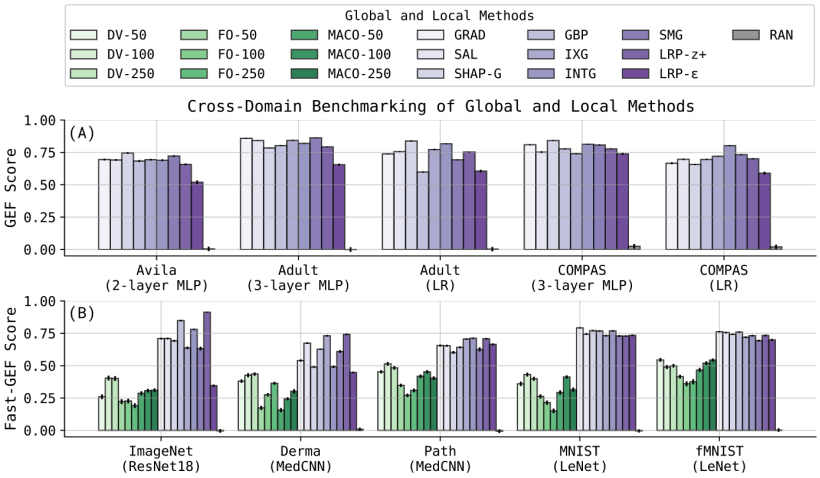



Figure 7: `GEF` and `Fast-GEF` results on (A) local across tabular, and (B) local versus global methods across vision tasks. The error bar shows the standard error, *i.e.,* $\frac{\sigma}{\sqrt{N}}$, where $\sigma$ is the standard deviation, and $N$ is the sample size.

### 6.2 Cross-Evaluating Local, and Global Methods

While local, and global explanations serve distinct purposes, and provide different insights *w.r.t.* their model, it is beneficial to compare them side-by-side in a unified view, as they often rely on similar methodological components, such as network gradients (LeCun et al., 1998; Olah et al., 2017). The absence of general-purpose evaluations has however so far prevented such comparison. `GEF` and `Fast-GEF` effectively fill this gap, facilitating a first, cross-domain comparative faithfulness benchmarking between global, and local methods. Extended results are provided in Appendix A.8.4.

Figs. 7 and 8 provide an overview of cross-domain results for tabular, and vision tasks. For all tabular tasks, `GEF` estimates are computed. For vision tasks, due to the high computational cost of global methods, `Fast-GEF` is used to allow for a fair comparison to local methods. As shown in Fig. 7, no explanation method is perfectly faithful to its model (*i.e.,* no score equals 1) nor consistently outperforms others across tested tasks. This variation aligns with most benchmarking studies of local linear approximation methods, which rarely identify a single winning method (Hedström et al., 2024; Hesse et al., 2024). Among tested global feature visualisation methods, MACO generally outperforms FO variants, consistent with Fel et al. (2024). Comparing the faithfulness scores of DV, MACO, and FO reveals that more optimisation steps do not necessarily result in higher explanation faithfulness. All tested methods significantly outperform the random baseline (RAN), which serves as the theoretical lower bound. As expected, RAN produces faithfulness scores centred around zero. In Fig. 8 (A), and (B), we observe that RAN explanation distortion quantities are flat, *i.e.,* independent of the distortion. Tabs. A.4, and A.5 in Appendix A.8 present the result of Fig. 7.

**Local Methods are Moderately Aligned.** Despite local methods showing imperfect, and highly varying scores across models, and datasets, most `GEF` estimates in tabular tasks, and `Fast-GEF` estimates in vision exceed 0.5, suggesting that the explanation retains some alignment with its model. This is not surprising given that parameter scaling has a direct effect on the model's curvature, to which local gradient-based methods are highly sensitive (Dombrowski et al., 2019), thereby instantaneously influencing their responsiveness to perturbation. Fig. 7 (A) shows that some local methods produce distortion outputs nearly monotonically related to its model, particularly at lower magnitudes (*i.e.,* a $z \leq 3$). This finding nuances studies by Adebayo et al. (2018), which provide single-point sensitivity estimates, conclusively reporting low reactivity to parameter randomisation in local methods. Corroborating recent rebuttal works (Yona & Greenfeld, 2021; Sundararajan & Taly, 2018; Binder et al., 2022) that challenges stark claims of method failure (Adebayo et al., 2018), we find that gradient-based methods are moderately faithful.

**Global Methods are Constrained by Regulariser.** Fig. 8 (C) shows aggregate `Fast-GEF` scores, indicating that global feature visualisation methods typically are less faithful compared to local linear approx-

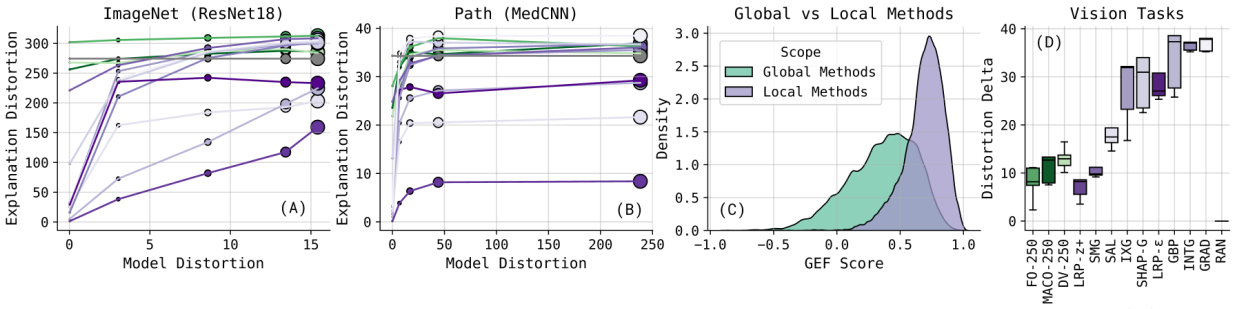



Figure 8: `Fast-GEF` results for vision tasks. (A), and (B) plot the model, and explanation distortion for ImageNet (ResNet18), and Path (MedCNN) along the perturbation path with $Z = 5$ perturbation steps. Here, global methods (DV, MACO, FO) are selected with 250 optimisation steps. (C) displays the distribution of `Fast-GEF` scores for local, and global methods, respectively, aggregated over all vision tasks. (D) reports the difference in explanation distortion between start $z = 1$, and end point $z = 5$, aggregated over all vision tasks.

imation methods. These differences in GEF estimates may be attributed to the global methods' inherent reliance on optimisation procedure (Olah et al., 2017), and NN's ability to retain its learned features despite perturbation via parameter perturbation (Binder et al., 2022). For reference, DV applies multiple regularisation techniques directly to the image, such as Gaussian blur, and cropping regions based on norm, and pixel contribution, while MACO, and FO regularise the frequency domain representation, with MACO adding an extra layer of regularisation via a predefined magnitude template. As observed in Fig. 8 (A), and (B), despite model perturbation, explanation distortions stay relatively flat, with lower distortion deltas compared to most local methods, as displayed in Fig. 7 (D). A strongly regularised optimisation procedure may inherently limit the faithfulness of global methods, in favour of a maximally activated neuron response.

### 6.3 Evaluating LLMs as Post-hoc Explainers

While researchers have recently begun exploring the potential of using LLMs as post-hoc explainers, there is still limited theoretical understanding, and empirical evidence on the general faithfulness of such approach. Can an LLM which is inherently decoupled from the model it seeks to explain, provide faithful outcomes? In our evaluation, we prompt `Gemma-2B-IT` for a top-$K$ token explanation for a given input, and prediction pair for datasets characterised by short tokenized lengths, *i.e.,* 59 for SST-2, and 128 for SMS Spam. The post-processed binary explanation vectors are then evaluated with `GEF` and `Fast-GEF`. See Appendix A.8.5 for further details, and extended results.

**LLM-x Explanations Comparable to Random.** Our `GEF` and `Fast-GEF` results in Fig. 9 (A), and (B) show that `Gemma-2B-IT` as an explainer is (i) significantly less faithful than local methods such as SHAP-P, and L-INTG, and (ii) similarly unfaithful as random explainers RAN-5 or RAN-10, on both SST-2, and SMS Spam classification tasks. Fig. 9 (C), and (D) demonstrate that these findings generalise over both top-5,

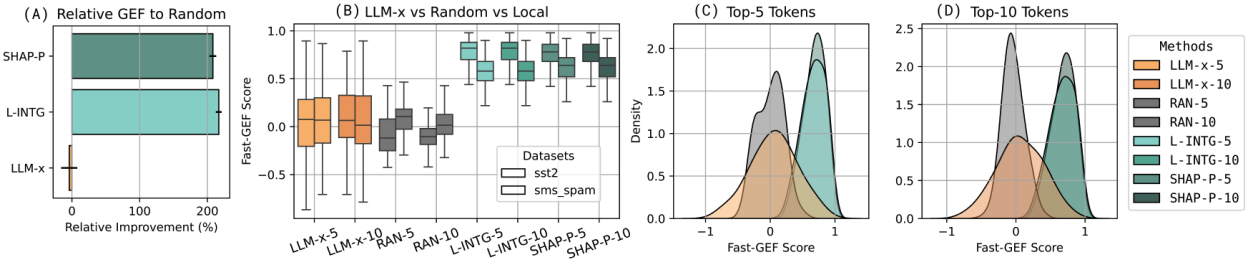



Figure 9: `GEF` (with $M$=3), and `Fast-GEF` results on different top-$K$ explanation NLP tasks. (A) shows the percentage improvement in `GEF` scores relative to RAN, aggregated over all tasks, with error bars showing the standard error. (B) shows the results in the form of box plots for the two datasets with SST-2 (*left*), and SMS Spam (*right*). (C), and (D) show the distribution of `Fast-GEF` scores for top-5, and top-10 explanations respectively, both aggregated over all tasks.

and top-10 tokens tasks, aggregated over both datasets. Our results, showing that LLM-x explanations are not more faithful than random, differ from the encouraging results reported by Kroeger et al. (2023), who found `GPT-4` to be as faithful as local methods in identifying top-$K$ tokens for tabular tasks. This divergence may naturally stem from variations in the experimental setup, including the specific explanation task, LLM used, methodology to evaluate faithfulness, and prompting strategies, however, it also underscores that the faithfulness of LLM-x is still an open research question. To more fully understand the potential of LLM-x as an explainer, further research with additional LLMs would be beneficial.

### 6.4 Measuring Faithfulness of Sparse Autoencoders

SAEs are gaining attention for their claimed ability to construct interpretable "monosemantic" features, but their general faithfulness remains underexplored (Makelov et al., 2024; Mallen & Belrose, 2024). To this end, we evaluate SAE explanations for `Gemma-2-2B` model on $N = 250$ samples on the IMDb dataset (Maas et al., 2011). The `Gemma-2-2B` model is repurposed as a binary classifier by extracting the logits of the "positive" or "negative" classes at the final token position of the prompt. Appendix A.4.4 provides more details.

**High Faithfulness Independent of SAE Width.** Fig. 10 (A) demonstrate that the SAE explanations generally are faithful with respect to the model's intermediate representations. `Fast-GEF` scores are consistently above 0.75 except for fluctuations in layers 1, and $16 - 19$. No significant difference is observed between 16K, and 65K widths, suggesting that the width of the encoding, *i.e.,* the capacity of the SAE, does not correlate with faithfulness. Moreover, Fig. 10 (B) shows that although sparsity of the latent activations decreases in later layers, it does not influence faithfulness (*i.e.,* $\rho = 0.023$, computed with Spearman Rank correlation). This suggests that a higher activation in SAE latents does not necessarily impact faithfulness. This raises the question of whether the faithfulness of SAE explanations is inherently scalable within different statistical or qualitative situations where SAEs are studied. Fig. 10 (C)-(G) further illustrates how explanation distortions vary across layers — with values (*y-axis*) increasing in the middle layers, *i.e.,* 10-20.

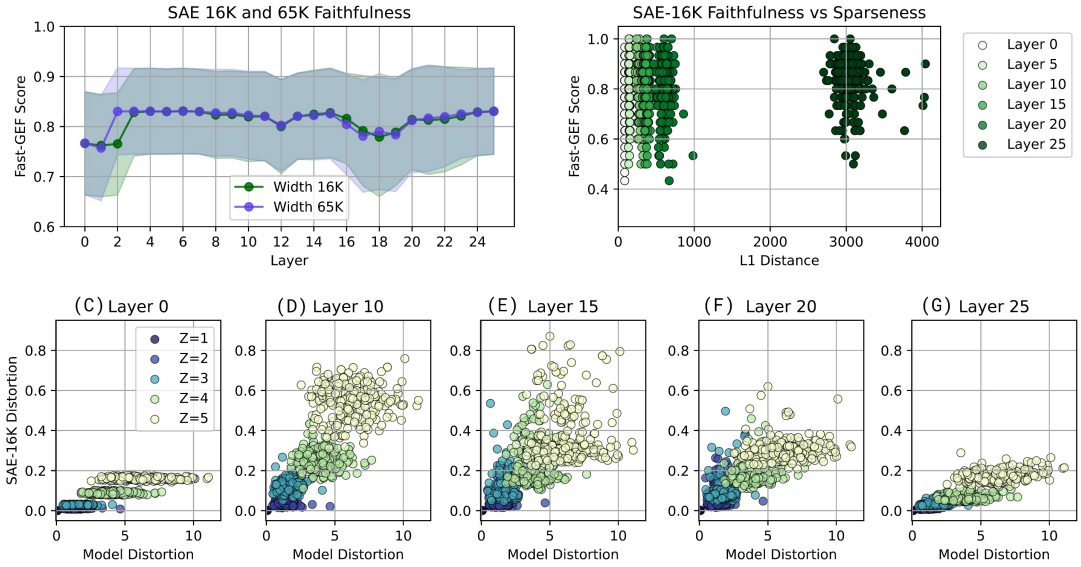



Figure 10: `Fast-GEF` results for SAE explanations on IMDb dataset, using $Z = 5$ perturbation levels, and $M = 3$ models. (A) shows `Fast-GEF` faithfulness scores across layers. (B) shows the sparsity as defined by L1 distance against the `Fast-GEF` indicating no relationship. (C)-(G) illustrates how SAE distortions develop across model layers, coloured by perturbation level.

## 7 Discussion: Where Are We Going?

With the evolving landscape of interpretability, redefining both the conceptual framework, and the geometric foundations of explanation faithfulness is critical. Our work puts forward a long-overdue unification of robustness, sensitivity, and faithfulness evaluations, providing a novel, and urgently needed, revised approach

(Def. 5) to evaluate the direct alignment between explanation, and model functions (Sec. 3). By using differential geometry as the basis of our proposed evaluation method, `GEF` combined with a task-agnostic perturbation strategy, we address the fundamental flaws of many existing evaluations: systematic overlook of the intrinsic geometry of non-linear spaces (Sec. 4). Our solution offers a threshold-free, fair comparison of functional distortions, making our approach not just another evaluation method but a necessary foundation for future interpretability research (Sec. 5).

**Novel, Empirical Insights.** In a first-ever cross-domain faithfulness benchmarking of global, and local explanations on vision, tabular, and NLP tasks (Sec. 6), we learn that tested local explanation methods generally are moderately faithful. We find that global feature visualisation methods are comparatively less faithful, which is an important understanding considering the recent evidence pointing to their general susceptibility to adversarial manipulation (Geirhos et al., 2023; Bareeva et al., 2024a). While it would be valuable to compare our findings with existing studies, to our knowledge, there is no direct study on the faithfulness of feature visualisations. Existing evaluations focus on alignment with human preferences or improvement on downstream task (Borowski et al., 2021; Zimmermann et al., 2021; Krishna et al., 2023; Bareeva et al., 2024b) or similarity to natural samples of the explained class (Fel et al., 2024). Our findings on generalised faithfulness thus provide complementary insights into the quality of feature visualisation as *model* explainers. Additionally, due to the recent interest LLMs as potential post-hoc explainers (Krishna et al., 2023; Kroeger et al., 2023), we study their faithfulness. We find no improved faithfulness compared to random explanations, and urge more investigation on this question. Finally, we observe that residual stream SAEs on `Gemma-2-2B` exhibit generally high faithfulness, with limited influence by the width (16K or 65K) and sparsity of the explanation. Further investigation is required to fully understand the potential of SAEs, and LLM-x as *generally faithful* explainers.

### 7.1 Limitations

While the results in our paper allow us to claim that our proposed method is sounder geometrically (Sec. 4), more reliable empirically (Sec. 6.1), and easier to use practically (Sec. 5), our evaluation alone does not guarantee explanation quality. Without ground truth labels, we cannot assess the statistical validity of an explanation function. An explanation may be estimated to be *generally faithful* but still lack intrinsic value (Bhattacharjee & von Luxburg, 2024) or interpretable qualities (Bordt & von Luxburg, 2024). The need for a thorough, application-grounded assessment of explanation quality that asserts value on a downstream task (Krishna et al., 2023; Lanham et al., 2023) is not eliminated with `GEF`. Evaluation using synthetic models with known ground truth (Carmichael & Scheirer, 2023) could complement our proposal.

### 7.2 Future Work

There are several exciting geometric, and empirical questions worth exploring. The geometric considerations in `GEF` suggest a deeper examination of the computational trade-offs of computing accurate pullbacks on individual explanation functions, specifically in comparing global versus local methods. In future work, there is opportunity to build on the growing body of research in ML that draws from geometry, and related topics in higher mathematics to deepen our understanding of NNs, and problems to which they are applied (Stephenson et al., 2021; Burns & Tang, 2023; Papamarkou et al., 2024). Recent theoretical studies on LLMs, and transformer models (Hoogland et al., 2024; Burns, 2024) have illustrated how neural activations may arrive at, and utilise "superpositional" encoding strategies (Elhage et al., 2022), which prominently feature considerations or findings of a geometric or topological nature. Continued development of general frameworks, and theories that conceptualise NNs in terms of geometry, and topology (Bianchini & Scarselli, 2014; Hauser & Ray, 2017; Naitzat et al., 2020; Benfenati & Marta, 2023a;b; Burns & Fukai, 2023) will likely facilitate a deeper understanding of both explanations, and evaluations, particularly in relation to the underlying mathematical characteristics of data, optimisation processes, and learned functions.

Recent advances in manifold geometry have introduced tools to analyse how input data modulates internal processing through perturbations (Kvinge et al., 2023). Exploring how explanation faithfulness varies with training data, and how it intersects with the geometric characteristics of the model presents an exciting direction. We also expect models optimized with non-Euclidean methods (Fei et al., 2023) to reveal stronger

differences between `GEF` and `Fast-GEF`, providing new opportunities to study the interplay between geometry, and faithfulness in explainability.

Lastly, we plan to expand our benchmarking scope to include natural activation-maximisation explanations (Borowski et al., 2021), concept-based explanations like INVERT (Bykov et al., 2023), and non-classification tasks. Given that pullback calculations can be computationally prohibitive for high-dimensional explanations, and highly parameterised models, exploring ways to speed up the Jacobian calculation (Eq. 21), and employ adaptive noise schedules would be valuable.

### Broader Impact Statement

Interpretability, or XAI, is widely acknowledged as essential for responsible ML. This paper critically examines current evaluation methods from unifying, and geometric perspectives, and proposes improvements. While negative societal impacts are improbable, overreliance on any single evaluation method is not advised.

## References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 9525–9536, 2018.

Chirag Agarwal, Nari Johnson, Martin Pawelczyk, Satyapriya Krishna, Eshika Saxena, Marinka Zitnik, and Himabindu Lakkaraju. Rethinking stability for attribution-based explanations. *CoRR*, abs/2203.06877, 2022a.

Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. OpenXAI: Towards a transparent evaluation of model explanations. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022b.

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models. *CoRR*, abs/2402.04614, 2024.

AlignmentResearch. robust-llm-pythia-imdb-14m-mz-ada-v3 (revision da044f3), 2024.

Tiago A. Almeida, Jose Maria Gomez Hidalgo, and Akebo Yamakami. Contributions to the study of sms spam filtering: New collection and results. In *Proceedings of the 2011 ACM Symposium on Document Engineering (DOCENG'11)*, 2011.

David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 7786–7795, 2018a.

David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 7786–7795, 2018b.

Kenza Amara, Rita Sevastjanova, and Mennatallah El-Assady. Challenges and opportunities in text generation explainability. In Luca Longo, Sebastian Lapuschkin, and Christin Seifert (eds.), *Explainable Artificial Intelligence*, pp. 244–264, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-63787-2.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Christopher J. Anders, David Neumann, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin. Software for dataset-wide xai: From local explanations to global insights with Zennit, CoRelAy, and ViRelAy. *CoRR*, abs/2106.13200, 2021.

Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilovic, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John T. Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques, 2019.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 283–294. Association for Computational Linguistics, 2023.

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), 2015.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *J. Mach. Learn. Res.*, 11:1803–1831, 2010.

Dilyara Bareeva, Marina MC Höhne, Alexander Warnecke, Lukas Pirch, Klaus Robert Muller, Konrad Rieck, and Kirill Bykov. Manipulating feature visualizations with gradient slingshots. In *ICML 2024 Next Generation of AI Safety Workshop*, 2024a.

Dilyara Bareeva, Galip Ümit Yolcu, Anna Hedström, Niklas Schmolenski, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. Quanda: An interpretability toolkit for training data attribution evaluation and beyond, 2024b. URL `https://arxiv.org/abs/2410.07158`.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

I. Bellido and E. Fiesler. Do backpropagation trained neural networks have normal weight distributions? In Stan Gielen and Bert Kappen (eds.), *ICANN '93*, pp. 772–775, London, 1993. Springer London. ISBN 978-1-4471-2063-6.

Alessandro Benfenati and Alessio Marta. A singular riemannian geometry approach to deep neural networks i. theoretical foundations. *Neural Networks*, 158:331–343, 2023a. ISSN 0893-6080. doi: 10.1016/j.neunet.2022.11.022.

Alessandro Benfenati and Alessio Marta. A singular riemannian geometry approach to deep neural networks ii. reconstruction of 1-d equivalence classes. *Neural Networks*, 158:344–358, 2023b. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2022.11.026.

Jose Manuel Benitez, Juan Luis Castro, and Ignacio Requena. Are artificial neural networks black boxes? *IEEE Trans. Neural Networks*, 8(5):1156–1164, 1997.

Pietro Berkes and Laurenz Wiskott. On the analysis and interpretation of inhomogeneous quadratic forms as receptive fields. *Neural Comput.*, 18(8):1868–1895, 2006.

Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations. In Christian Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pp. 3016–3022. ijcai.org, 2020.

Robi Bhattacharjee and Ulrike von Luxburg. Auditing local explanations is hard, 2024.

Monica Bianchini and Franco Scarselli. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *IEEE Transactions on Neural Networks and Learning Systems*, 25(8):1553–1565, 2014. doi: 10.1109/TNNLS.2013.2293637.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. *URL https://openaipublic. blob. core. windows. net/neuron-explainer/paper/index. html.(Date accessed: 14.05. 2023)*, 2, 2023.

Alexander Binder, Leander Weber, Sebastian Lapuschkin, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Shortcomings of top-down randomization-based sanity checks for evaluations of deep neural network explanations. *CoRR*, abs/2211.12486, 2022.

Stefan Blücher, Johanna Vielhaben, and Nils Strodthoff. Decoupling pixel flipping and occlusion strategy for consistent xai benchmarks. *arXiv preprint arXiv:2401.06654*, 2024.

Douglas G Bonett and Thomas A Wright. Sample size requirements for estimating pearson, kendall and spearman correlations. *Psychometrika*, 65:23–28, 2000.

Sebastian Bordt and Ulrike von Luxburg. Statistics without interpretation: A sober look at explainable machine learning. *CoRR*, abs/2402.02870, 2024.

Judy Borowski, Roland Simon Zimmermann, Judith Schepers, Robert Geirhos, Thomas S. A. Wallis, Matthias Bethge, and Wieland Brendel. Exemplary natural images explain CNN activations better than state-of-the-art feature visualization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Lennart Brocki and Neo Christopher Chung. Evaluation of interpretability methods and perturbation artifacts in deep neural networks. *CoRR*, abs/2203.02928, 2022.

Lukas Brunke, Prateek Agrawal, and Nikhil George. Evaluating input perturbation methods for interpreting CNNs and saliency map comparison. In *Computer Vision – ECCV 2020 Workshops*, pp. 120–134. Springer International Publishing, 2020.

Thomas F Burns. Semantically-correlated memories in a dense associative model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4936–4970. PMLR, 21–27 Jul 2024.

Thomas F Burns and Tomoki Fukai. Simplicial hopfield networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=_QLsH8gatwx`.

Thomas F Burns and Robert Tang. Detecting danger in gridworlds using Gromov's Link Condition. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Kirill Bykov, Anna Hedström, Shinichi Nakajima, and Marina M.-C. Höhne. Noisegrad - enhancing explanations by introducing stochasticity to model weights. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, February 22 - March 1, 2022*, pp. 6132–6140. AAAI Press, 2022.

Kirill Bykov, Laura Kopf, Shinichi Nakajima, Marius Kloft, and Marina MC Höhne. Labeling neural representations with inverse recognition. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Zachariah Carmichael and Walter Scheirer. How well do feature-additive explainers explain feature-additive predictors? In *XAI in Action: Past, Present, and Future Applications*, 2023.

Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Xi Wu, and Somesh Jha. Concise explanations of neural networks using adversarial training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1383–1391. PMLR, 2020.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 782–791. Computer Vision Foundation / IEEE, 2021.

Yu-Neng Chuang, Guanchu Wang, Chia-Yuan Chang, Ruixiang Tang, Fan Yang, Mengnan Du, Xuanting Cai, and Xia Hu. Large language models as faithful explainers. *CoRR*, abs/2402.04678, 2024.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Ian Covert, Chanwoo Kim, and Su-In Lee. Learning to estimate shapley values with vision transformers. *CoRR*, abs/2206.05282, 2022.

Sanjoy Dasgupta, Nave Frost, and Michal Moshkovitz. Framework for evaluating faithfulness of local explanations. In *International Conference on Machine Learning*, pp. 4794–4815. PMLR, 2022.

Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 13567–13578, 2019.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Dumitru Erhan, Y. Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *Technical Report, Univeristé de Montréal*, 01 2009.

Yanhong Fei, Xian Wei, Yingjie Liu, Zhengyu Li, and Mingsong Chen. A survey of geometric optimization for deep learning: From euclidean space to riemannian manifold. *arXiv preprint arXiv:2302.08210*, 2023.

Thomas Fel, Thibaut Boissin, Victor Boutin, Agustin Picard, Paul Novello, Julien Colin, Drew Linsley, Tom Rousseau, Rémi Cadène, Laurent Gardes, and Thomas Serre. Unlocking feature visualization for deeper networks with magnitude constrained optimization. *Advances in Neural Information Processing Systems*, 36, 2024.

Robert Geirhos, Roland S. Zimmermann, Blair L. Bilodeau, Wieland Brendel, and Been Kim. Don't trust your eyes: on the (un)reliability of feature visualizations. *CoRR*, abs/2306.04719, 2023.

Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, 2021. ISSN 2589-7500. doi: 10.1016/S2589-7500(21)00208-9.

Riccardo Guidotti. Evaluating local explanation methods on ground truth. *Artif. Intell.*, 291:103428, 2021. doi: 10.1016/J.ARTINT.2020.103428.

Peter Hase, Harry Xie, and Mohit Bansal. The out-of-distribution problem in explainability and search methods for feature importance explanations. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 3650–3666, 2021.

Johannes Haug, Stefan Zürn, Peter El-Jiz, and Gjergji Kasneci. On baselines for local feature attributions. *CoRR*, abs/2101.00905, 2021.

Michael Hauser and Asok Ray. Principles of riemannian geometry in neural networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016.

Anna Hedström, Philine Lou Bommer, Kristoffer Knutsen Wickstrøm, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. The meta-evaluation problem in explainable AI: identifying reliable estimators with metaquantus. *Trans. Mach. Learn. Res.*, 2023, 2023a.

Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34): 1–11, 2023b.

Anna Hedström, Leander Weber, Sebastian Lapuschkin, and Marina Höhne. A fresh look at sanity checks for saliency maps. In *Explainable Artificial Intelligence*, pp. 403–420, Cham, 2024. Springer Nature Switzerland.

Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. Benchmarking the attribution quality of vision models. *CoRR*, abs/2407.11910, 2024. doi: 10.48550/arXiv.2407.11910.

Jesse Hoogland, George Wang, Matthew Farrugia-Roberts, Liam Carroll, Susan Wei, and Daniel Murfet. The developmental landscape of in-context learning. *arXiv preprint arXiv:2402.02364*, 2024.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 9734–9745, 2019.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=F76bwRSLeK`.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 4198–4205. Association for Computational Linguistics, 2020.

Neil Jethani, Adriel Saporta, and Rajesh Ranganath. Don't be fooled: label leakage in explanation methods and the importance of their quantitative evaluation. In Francisco J. R. Ruiz, Jennifer G. Dy, and Jan-Willem van de Meent (eds.), *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*, volume 206 of *Proceedings of Machine Learning Research*, pp. 8925–8953. PMLR, 2023.

Niklas Koenen and Marvin N. Wright. Toward understanding the disagreement problem in neural network feature attribution. *CoRR*, abs/2404.11330, 2024.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch. *CoRR*, abs/2009.07896, 2020.

Narine Kokhlikyan, Vivek Miglani, Bilal Alsallakh, Miguel Martin, and Orion Reblitz-Richardson. Investigating sanity checks for saliency maps with image and text classification. *CoRR*, abs/2106.07475, 2021.

Laura Kopf, Philine Lou Bommer, Anna Hedström, Sebastian Lapuschkin, Marina MC Höhne, and Kirill Bykov. Cosy: Evaluating textual explanations of neurons. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024.

Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *CoRR*, abs/2202.01602, 2022.

Satyapriya Krishna, Jiaqi Ma, Dylan Z Slack, Asma Ghandeharioun, Sameer Singh, and Himabindu Lakkaraju. Post hoc explanations of language models can improve language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Nicholas Kroeger, Dan Ley, Satyapriya Krishna, Chirag Agarwal, and Himabindu Lakkaraju. Are large language models post hoc explainers? *CoRR*, abs/2310.05797, 2023.

Henry Kvinge, Grayson Jorgenson, Davis Brown, Charles Godfrey, and Tegan Emerson. Internal representations of vision models through the lens of frames on data manifolds. In *NeurIPS 2023 Workshop on Symmetry and Geometry in Neural Representations*, 2023.

Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. Rethinking explainability as a dialogue: A practitioner's perspective. *CoRR*, abs/2202.01875, 2022.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *CoRR*, abs/2307.13702, 2023.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

John M Lee. *Smooth manifolds*. Springer, 2012.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, 2024. URL `https://arxiv.org/abs/2408.05147`.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4765–4774, 2017.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*, 2024. URL https://openreview.net/forum?id=MHIX9H8aYF.

Alex Mallen and Nora Belrose. Balancing label quantity and quality for scalable elicitation, 2024.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Cristian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, and et al. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295, 2024.

Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognit.*, 65:211–222, 2017. doi: 10.1016/j.patcog.2016.11.008.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.*, 73:1–15, 2018.

Niels J. S. Morch, Ulrik Kjems, Lars Kai Hansen, Claus Svarer, Ian Law, Benny Lautrup, Stephen C. Strother, and Kelly Rehm. Visualization of neural networks using saliency maps. In *Proceedings of International Conference on Neural Networks (ICNN'95), Perth, WA, Australia, November 27 - December 1, 1995*, pp. 2085–2090. IEEE, 1995.

Gregory Naitzat, Andrey Zhitnikov, and Lek-Heng Lim. Topology of deep neural networks. *Journal of Machine Learning Research*, 21(184):1–40, 2020.

Michael Neely, Stefan F. Schouten, Maurits J. R. Bleeker, and Ana Lucic. Order in the court: Explainable AI methods prone to disagreement. *CoRR*, abs/2105.03287, 2021.

Anphi Nguyen and Maria Rodriguez Martinez. On quantitative aspects of model interpretability. *CoRR*, abs/2007.07584, 2020.

Hoa Nguyen. Activation maximization. https://github.com/Nguyen-Hoa/Activation-Maximization, 2020.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

Theodore Papamarkou, Tolga Birdal, Michael M. Bronstein, Gunnar E. Carlsson, Justin Curry, Yue Gao, Mustafa Hajij, Roland Kwitt, Pietro Lio, Paolo Di Lorenzo, Vasileios Maroulas, Nina Miolane, Farzana Nasrin, Karthikeyan Natesan Ramamurthy, Bastian Rieck, Simone Scardapane, Michael T Schaub, Petar Veličković, Bei Wang, Yusu Wang, Guowei Wei, and Ghada Zamzmi. Position: Topological deep learning is the new frontier for relational learning. In *Forty-first International Conference on Machine Learning*, 2024.

ProPublica. Compas recidivism risk score data and analysis, 2016. URL https://github.com/propublica/compas-analysis.

Luyu Qiu, Yi Yang, Caleb Chen Cao, Jing Liu, Yueyuan Zheng, Hilary Hei Ting Ngai, Janet H. Hsiao, and Lei Chen. Resisting out-of-distribution data problem in perturbation of XAI. *CoRR*, abs/2107.14000, 2021.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi (eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144. ACM, 2016a.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi (eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144. ACM, 2016b.

Laura Rieger and Lars Kai Hansen. IROF: a low resource evaluation metric for explanation methods. *CoRR*, abs/2003.08747, 2020.

Manuel Romero. bert-tiny-finetuned-sms-spam-detection (revision 012e268), 2024.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18770–18795. PMLR, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Trans. Neural Networks Learn. Syst.*, 28(11):2660–2673, 2017.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.*, 128(2):336–359, 2020.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *CoRR*, abs/1605.01713, 2016.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3145–3153. PMLR, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Irwin Sobel, Gary Feldman, et al. A 3x3 isotropic gradient operator for image processing. *a talk at the Stanford Artificial Project in*, 1968:271–272, 1968.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

Claudio Stefano, Francesco Fontanella, Marilena Maniaci, and Alessandra Freca. Avila. UCI Machine Learning Repository, 2018. DOI: 10.24432/C5K02X.

Cory Stephenson, suchismita padhy, Abhinav Ganesh, Yue Hui, Hanlin Tang, and SueYeon Chung. On the geometry of generalization and memorization in deep neural networks. In *International Conference on Learning Representations*, 2021.

Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 2020. doi: 10.23915/distill.00022. https://distill.pub/2020/attribution-baselines.

Mukund Sundararajan and Ankur Taly. A note about: Local explanation methods for deep neural networks lack sensitivity to parameter values. *CoRR*, abs/1806.04205, 2018.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017.

Zeren Tan and Yang Tian. Robust explanation for free or at the cost of faithfulness. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33534–33562. PMLR, 2023.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

VityaVitalich. bert-tiny-sst2 (revision 2e14b76), 2023.

Matthew Watson, Bashar Awwad Shiekh Hasan, and Noura Al Moubayed. Agree to disagree: When deep learning models with identical architectures produce distinct explanations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 875–884, January 2022.

Kristoffer K Wickstrøm, Marina M. C. Höhne, and Anna Hedström. From flexibility to manipulation: The slippery slope of xai evaluation. *CoRR*, abs/INSERT, 2024.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. ISSN 00994987.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I. Inouye, and Pradeep Ravikumar. On the (in)fidelity and sensitivity of explanations. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 10965–10976, 2019.

Gal Yona and Daniel Greenfeld. Revisiting sanity checks for saliency maps. *CoRR*, abs/2110.14297, 2021.

Jason Yosinski, Jeff Clune, Anh Mai Nguyen, Thomas J. Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *CoRR*, abs/1506.06579, 2015.

Jerrold H Zar. Spearman rank correlation. *Encyclopedia of biostatistics*, 7, 2005.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, volume 8689 of *Lecture Notes in Computer Science*, pp. 818–833. Springer, 2014a.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, volume 8689 of *Lecture Notes in Computer Science*, pp. 818–833. Springer, 2014b.

Roland S. Zimmermann, Judy Borowski, Robert Geirhos, Matthias Bethge, Thomas S. A. Wallis, and Wieland Brendel. How well do feature visualizations support causal understanding of CNN activations? In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 11730–11744, 2021.

## Appendix

The Appendix is organised as follows: theoretical considerations (Sec. A.1), implementation notes for `GEF` and `Fast-GEF` (Sec. A.2), details on the `MetaQuantus` framework (Sec. A.3), the general experimental setup (Sec. A.4), analysis of model assumptions (Sec. A.5), alignment patterns extended results (Sec. A.6), ablation experiment results (Sec. A.7), extended results from individual experiments (Sec. A.8), and notation tables (Sec. A.9).

### A.1 Theoretical Considerations

The following subsections provide detailed proofs, extensions, and discussions surrounding the GEF criterion.

#### A.1.1 GEF: Penalising Random Explanations

Following the discussion in Sec. 3, a quality estimator should be able to recongise model-independent unfaithful explanations. In the following, we show that our proposed GEF criterion (Def. 5) identifies two specific types of unfaithful explanations by design: constant, and random explanations.

**Corollary 1 (Penalising Unfaithful Explanations)** *Let $\Psi^{GEF}$ be a quality estimator that yields estimates $q^{GEF} \in \mathbb{R}$, with $q^{GEF} = 0$ indicating a lack of generalised faithfulness. To be a valid measure of explanation quality, $\Psi^{GEF}$ should assign low scores to both (I) constant, and (II) random explanations:*

$$\textit{Constant (I):} \ \forall \hat{\boldsymbol{e}} : \boldsymbol{e} = \hat{\boldsymbol{e}} \Rightarrow q^{GEF} = 0$$
$$\textit{Random (II):} \ \forall \hat{\boldsymbol{e}} : \hat{\boldsymbol{e}} \sim \mathcal{U}(0,1) \Rightarrow q^{GEF} = 0$$

*where $\hat{\boldsymbol{e}}$, and $\boldsymbol{e}$ are perturbed, and unperturbed explanations, and $\mathcal{U}(0,1)$ denotes a uniform distribution. The GEF estimate $q^{GEF} = \rho(\boldsymbol{d}_f, \boldsymbol{d}_\phi)$ (Def. 5) assigns low scores in the first, and the second case.*

**Proof.** In case (I), the explanation does not change across perturbations, leading to an explanation distortion vector $\boldsymbol{d}_\phi$ that contains only zeros:

$$\forall \hat{\boldsymbol{e}}, z \in [1, Z] : \boldsymbol{e} = \hat{\boldsymbol{e}} \Rightarrow \ \boldsymbol{D}_\phi^z = 0,$$

whereas the model's distortion vector $\boldsymbol{d}_f$ will contain non-zero values due to perturbations:

$$\forall z \in [1, Z] : \boldsymbol{D}_f^z \neq 0.$$

Consequently, the correlation coefficient $\rho(\boldsymbol{d}_f, \boldsymbol{d}_\phi)$ will be zero with $q^{GEF} = 0$.

In case (II), the explanation distortion $\boldsymbol{D}_\phi^z$ will be approximately uniform across all perturbation steps since each perturbation is independently drawn from the same distribution:

$$\forall \hat{\boldsymbol{e}}, z, j \in [1, Z], z \neq j : \hat{\boldsymbol{e}} \sim \mathcal{U}(0,1) \Rightarrow \boldsymbol{D}_\phi^z \approx \boldsymbol{D}_\phi^j,$$

whereas the model distortion $\boldsymbol{D}_f^z$ will vary according to the degree of perturbation:

$$\forall \hat{\boldsymbol{e}}, z, j \in [1, Z], z \geq j : \hat{\boldsymbol{e}} \sim \mathcal{U}(0,1) \Rightarrow \boldsymbol{D}_f^j \geq \boldsymbol{D}_\phi^z.$$

The lack of correlation between $\boldsymbol{d}_f$, and $\boldsymbol{d}_\phi$ results in a quality measure$q^{GEF}$ of 0. This completes the proof.

#### A.1.2 GEF: Extension

To extend the applicability of GEF (Def. 5) to global methods that explain *any* neuron within a model, we adopt Kopf et al. (2024), and view the model $f$ as a composition of two functions, $F : \mathcal{X} \to \mathcal{G}$, and $L : \mathcal{G} \to \mathcal{Y}$, such that $f = L \circ F$. Here $\mathcal{G} \subset \mathbb{R}^{c \times w^* \times h^*}$, where $c \in \mathbb{N}$ is the number of neurons in the layer, and $w^*, h^* \in \mathbb{N}$ represent the width, and height of the feature map, respectively. The function $F$, is referred to as the *feature extractor*. We redefine the model function as a chosen feature extractor, and replace $y$ in Def. 5 with the activation of the $c^{th}$ neuron such that *i.e.,* $y = F_c(\boldsymbol{x}, \theta) : \mathcal{X} \to \mathbb{R}^{w^* \times h^*}$. While the model's output space $\mathcal{Y}$ is replaced by $\mathcal{G}$, we similarly define the perturbed instance $\hat{y}$.

#### A.1.3 GEF: Derivation of Linear Case

Our definition of GEF is based on the observation that any distortion present in the model output space $\mathcal{Y}$, should be mirrored in the explanation space $\mathcal{E}$. Since neural networks are non-linear functions, a fair distortion in $\mathcal{Y}$, and $\mathcal{E}$, requires the introduction of the pullback (Sec. 4). In the case of a linear model, however, the relationship between the distortion quantities $\boldsymbol{D}_f$, and $\boldsymbol{D}_\phi$ can be derived analytically. Here, the explanation is based on the first-order Taylor term, which is a linear approximation of the model's behaviour, forming the foundation of many established explanation methods (*e.g.,* (Montavon et al., 2017)). We proceed to derive this relationship explicitly below.

**Proof.** Consider $f$ to be a linear model of the form $f(\boldsymbol{x};\theta) = \theta\boldsymbol{x} + c$. The explanation is the parameter vector $\theta$. We can derive the expected distortion $\boldsymbol{D}_f := \mathbb{E}_{\hat{\theta}_m}[(f(\boldsymbol{x};\theta) - f(\boldsymbol{x};\hat{\theta}_m))^2]$ (see Eq. 4) where $m \in [1, M]$ denotes the number of perturbed models for a fixed perturbation magnitude $\xi$, *i.e.,* a step $z$.

$$\boldsymbol{D}_f^z = (\theta\boldsymbol{x} + c)^2 - 2(\theta\boldsymbol{x} + c)\mathbb{E}_{\hat{\theta}_m}\left[(\hat{\theta}_m\boldsymbol{x} + c)\right] + \mathbb{E}_{\hat{\theta}_m}\left[(\hat{\theta}_m\boldsymbol{x} + c)^2\right] \tag{12}$$

$$\boldsymbol{D}_f^z = \theta^2\boldsymbol{x}^2 + 2c\theta\boldsymbol{x} + 2c^2 - 2\boldsymbol{x}(\theta\boldsymbol{x} + c)\mathbb{E}_{\hat{\theta}_m}\left[\hat{\theta}_m\right] - 2c(\theta\boldsymbol{x} + c) + \mathbb{E}_{\hat{\theta}_m}\left[\hat{\theta}_m^2\boldsymbol{x}^2 + 2c\hat{\theta}_m\boldsymbol{x}\right] \tag{13}$$

$$\boldsymbol{D}_f^z = \theta^2\boldsymbol{x}^2 - 2\boldsymbol{x}(\theta\boldsymbol{x} + c)\mathbb{E}_{\hat{\theta}_m}\left[\hat{\theta}_m\right] + 2c\boldsymbol{x}\mathbb{E}_{\hat{\theta}_m}\left[\hat{\theta}_m\right] + \boldsymbol{x}^2\mathbb{E}_{\hat{\theta}_m}\left[\hat{\theta}_m^2\right] \tag{14}$$

$$\boldsymbol{D}_f^z = \theta^2\boldsymbol{x}^2 - 2\theta\boldsymbol{x}^2\mathbb{E}_{\hat{\theta}_m}\left[\hat{\theta}_m\right] + \boldsymbol{x}^2\mathbb{E}_{\hat{\theta}_m}\left[\hat{\theta}_m^2\right] \tag{15}$$

For the explanation distortion, a similar decomposition can be performed:

$$\boldsymbol{D}_\phi^z = \theta^2 - 2\theta\mathbb{E}_{\hat{\theta}_m}\left[\hat{\theta}_m\right] + \mathbb{E}_{\hat{\theta}_m}\left[\hat{\theta}_m^2\right] \tag{16}$$

By combining Eq. 16 and 15, we arrive at:

$$\boldsymbol{D}_\phi^z = \frac{1}{\boldsymbol{x}^2}\boldsymbol{D}_f^z \tag{17}$$

We can construct the distortion vectors $\mathbf{d}_\phi$, and $\mathbf{d}_f$, and for each entry Eq. 17 holds. When $\rho$ is defined as the Pearson correlation coefficient, we find the distortion of the model $\mathbf{d}_f$, and the distortion of the explanation function $\mathbf{d}_\phi$ to be perfectly correlated:

$$\rho(\mathbf{d}_\phi, \mathbf{d}_f) = \frac{\mathrm{cov}_\xi\left(\mathbf{d}_f, \mathbf{d}_\phi\right)}{\sqrt{\mathrm{Var}_\xi(\mathbf{d}_f)}\sqrt{\mathrm{Var}_\xi(\mathbf{d}_\phi)}} \tag{18}$$

$$\rho(\mathbf{d}_\phi, \mathbf{d}_f) = \frac{1/\boldsymbol{x}^2\mathrm{Var}_\xi(\mathbf{d}_f)}{1/\boldsymbol{x}^2\mathrm{Var}_\xi(\mathbf{d}_f)} = 1. \tag{19}$$

This proves that in a simplified scenario, the key assumption of correlated distortion quantities holds, *i.e.,* the model parameters $\theta$ provide a *faithful* explanation. Since monotonicity is a weaker condition than linearity, Eq. 19 also holds when $\rho$ is defined as the Spearman Rank correlation coefficient (Bonett & Wright, 2000).

#### A.1.4 GEF: Influence of Z

The parameter $Z$ represents the number of steps in the perturbation path, and consequently dictates how finely the model's response will be captured by the GEF criterion (Def. 5). As such, selecting an appropriate value for $Z$ is critical because it affects the interpretation of the results. A higher $Z$ allows for a finer evaluation of how well an explanation aligns with the model's behaviour under varying conditions. When using Spearman's rank correlation coefficient as our measure of $\rho$, a larger $Z$ generally stabilises the faithfulness score due to the reduction in confidence intervals with more samples (*i.e.,* $CI \sim \frac{1}{Z}$) (Bonett & Wright, 2000). Nonetheless, this assumes a monotonic response from both the model, and the explanation, which may not be realistic (Sec. 4). If the model itself is not monotonic across perturbations, expecting the explanation to behave monotonically is also unrealistic.

Fig. A.1 (A) ($Z = 2$), and (B) ($Z = 20$) demonstrate the violation of the monotonicity assumption, as we observe large error bars, and divergent behaviour for $Z = 20$, indicating non-monotonic responses. Accordingly, a moderate value of $Z$ (Zar, 2005) is advised for meaningful measurement.

### A.2 Notes on `GEF` and `Fast-GEF` Implementation

In the following, we provide details on the GEF algorithm.

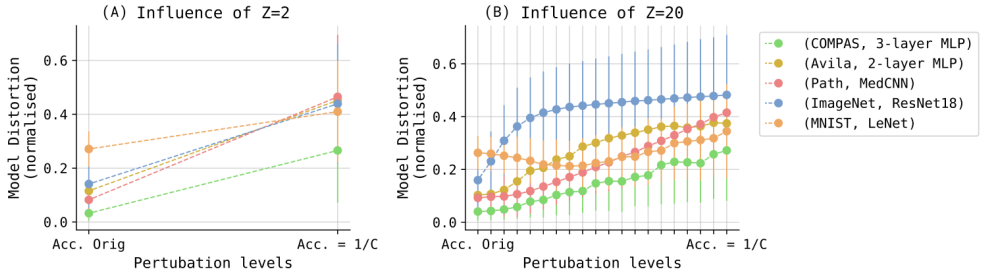



Figure A.1: Model distortion (normalised by its maximum value) with (A) showing $Z = 2$ perturbation steps, and (B) showing $Z = 20$ perturbation steps.

### A.2.1 Generate Perturbation Path

To generate the perturbation path of length $Z$, satisfying $y \neq \hat{y}$, $\forall y, \hat{y} \in \mathcal{Y}$, we computationally find the minimum noise level $\sigma_z^2$ at $z = Z$, such that the perturbed model's accuracy (ACC) approximates $\frac{1}{C}$, where $C$ is the number of classes, within a threshold, *i.e.,* $\epsilon << 1$. Here, $\text{ACC} = \frac{1}{N}\sum_{i=1}^{N}(f(\boldsymbol{X}_i;\theta) = \boldsymbol{Y}_i)$ where $N$ is the number of samples in the test set, denoted $\boldsymbol{X}$. This is achieved by progressively increasing $\sigma^2$, and applying it to the model according to Sol. 5.1, which process concludes when the model's accuracy satisfies the condition $|\text{ACC} - \frac{1}{C}| < \epsilon$, thereby determining perturbation level for subsequent evaluation.

**Compute Path Length.** For a more faithful estimate of explanation distortion, for each step $z \in [0, Z]$, we compute the path length for $\boldsymbol{D}_\phi$. We replace the integral in Eq. 11 with a sum over $T$ steps:

$$L(\gamma) = \sum_{t=1}^{T} de_t^T (J_f(\hat{\boldsymbol{e}}_t)^T J_f(\hat{\boldsymbol{e}}_t)) de_t, \tag{20}$$

where $de_t \in \mathbb{R}^V$ denotes the feature-wise difference in explanations *i.e.,* $(\boldsymbol{e} - \hat{\boldsymbol{e}}_t)$ with $\phi(f_{\hat{\theta}_t}, \ldots) = \hat{\boldsymbol{e}}_t$, and $J_f(\hat{\boldsymbol{e}}_t) \in \mathbb{R}^{V \times C}$ is the Jacobian for fixed $\boldsymbol{x}$, and $f_{\hat{\theta}_t}$. To numerically approximate this Jacobian, for each step $t \in [0, T]$, we perturb the neural activations (*i.e.,* logits $\hat{\boldsymbol{y}}$) by adding infinitesimal noise. In practice, we sample from a Gaussian distribution $\upsilon_k \sim \mathcal{N}(0, 0.001)$ such that $\hat{y}_k = \hat{\boldsymbol{y}} + \upsilon_k$, $k \in [1, K]$ times. After each perturbation, we recalculate the corresponding explanation $\phi(\hat{y}_k, \ldots) = \hat{\boldsymbol{e}}_k$. Elements of the Jacobian $J_f(\hat{\boldsymbol{e}}_t)$ are then computed as feature-wise difference between $\boldsymbol{e}$, and $\hat{\boldsymbol{e}}_k$:

$$\frac{\partial e_i}{\partial f_j} \approx \lim_{K \to \infty} \frac{1}{K} \sum_{k=1}^{K} (e_j - \hat{e}_{j,k}) \nu_k^{-1}. \tag{21}$$

where $i, j$ refers to the indices of the Jacobian $J_f(\hat{\boldsymbol{e}}_t)$.

Note that, unless specified otherwise, we set $M$, $Z$, $T$, and $K$ to 5, for all experiments. Please find Appendix A.7 for an ablation study on these hyperparameters.

## A.3 Notes on `MetaQuantus` framework

For meta-evaluation, a two-step process is employed. First, two types of controlled perturbations are introduced: minor, and disruptive. These are designed to evaluate the metric's resilience to noise ($NR$), and its sensitivity to adversarial conditions ($AR$), respectively. Specifically, these perturbations are applied in both the input, and model spaces, resulting in two distinct tests: the Input Perturbation Test (IPT), and the Model Perturbation Test (MPT)[5]. Second, the effects of the perturbations are measured in two

[5]For the IPT, independent, and identically distributed (i.i.d.) additive uniform noise is applied, defined as $\hat{\boldsymbol{x}}_i = \boldsymbol{x} + \nu_i$, where $\nu i \sim \mathcal{U}(\alpha, \beta)$. For the MPT, multiplicative Gaussian noise is applied to all network weights, represented as $\hat{\theta} i = \theta \cdot \nu i$ with $\nu i \sim \mathcal{N}(\mu, \sigma^2)$. The hyperparameters $\alpha, \beta, \mu, \sigma^2 a$ follow the specifications of the original study (Hedström et al., 2023a).

meta-evaluative criteria: intra-consistency (**IAC**), and inter-consistency (**IEC**). Here, **IAC** refers to measuring the similarity in score distributions post-perturbation, and **IEC** refers to the occurrence of categorical ranking changes within a set of distinct explanation methods[6]. Each metric is then assigned a summarised meta-consistency score, denoted as MC $\in [0, 1]$:

$$\text{MC} = \left(\frac{1}{|\mathbf{m}^*|}\right) \mathbf{m}^{*T}\mathbf{m} \quad \text{where} \quad \mathbf{m} = \begin{bmatrix} \mathbf{IAC}_{NR} \\ \mathbf{IAC}_{AR} \\ \mathbf{IEC}_{NR} \\ \mathbf{IEC}_{AR} \end{bmatrix}, \tag{22}$$

with $\boldsymbol{m}^* \in \mathbb{R}^4$ representing an ideal quality estimator, essentially a vector of ones. A higher MC score, approaching 1, indicates superior reliability according to the defined evaluation criteria. Metrics that demonstrate both resilience to minor perturbations, and reactivity to disruptive changes achieve higher MC scores. We refer to the original publication (Hedström et al., 2023a) for further details on the elements in the meta-evaluation vector $\mathbf{m}$ (Eq. 22), and the framework in general.

### A.4 General Experimental Setup

Here, we describe the models, datasets, tooling, hardware, explanation, and evaluation methods in this work.

#### A.4.1 Models, and Datasets

We employ various models for vision, text, and tabular tasks in our experiments. See Tab. 2.

For vision classification, we use ImageNet-1K for object recognition (Russakovsky et al., 2015) with ResNet18 (He et al., 2016), Pathology, and Derma for medical image analysis with proposed MedCNN architecture (Yang et al., 2023), MNIST, and fMNIST (LeCun et al., 2010; Xiao et al., 2017) for digit, and fashion recognition with LeNet (LeCun et al., 1998). For text classification, we use SMS Spam (Almeida et al., 2011) with a tiny, fine-tuned BERT model (Romero, 2024), IMDb (Maas et al., 2011) with Pythia (AlignmentResearch, 2024), and SST-2 (Socher et al., 2013) with a tiny, fine-tuned BERT model (VityaVitalich, 2023). For tabular classification, we use Adult (Becker & Kohavi, 1996) and, COMPAS (ProPublica, 2016), with 3-layer MLP, and and Avila (Stefano et al., 2018) with 2-layer MLP.

All models that are not publicly accessible are released at GitHub repository `url`.

#### A.4.2 Tooling

Several libraries, and open-source implementations enabled this work, including `transformers` (Wolf et al., 2020), `OpenXAI` (Agarwal et al., 2022b), `Captum` (Kokhlikyan et al., 2020), `Zennit` (Anders et al., 2021), `Shap` (Lundberg & Lee, 2017), `Activation-Maximization` (Nguyen, 2020), and `Horama` (Fel et al., 2024). For metric implementation, and meta-evaluation, we use the `Quantus` (Hedström et al., 2023b), and `MetaQuantus` (Hedström et al., 2023a) libraries, respectively.

#### A.4.3 Hardware

The experiments were conducted using two hardware configurations: a cluster with four Tesla V100S-PCIE-32GB GPUs, each offering 32 GB of memory, and a DGX-2 system featuring eight NVIDIA A100-SXM4-40GB GPUs, each with 40 GB of memory. Both setups support the NVIDIA driver version 535.161.07, and CUDA 12.2.

#### A.4.4 Explanation Methods

All the hyperparameters of the individual explanations methods, are listed in the main manuscript. Concerning the preprocessing, the signs of the attributions are maintained, unless the method algorithmically relies

[6]**IAC** provides a normalised p-value derived from the non-parametric *Wilcoxon signed-rank test* (Wilcoxon, 1945), comparing the original, and perturbed score distributions. For $NR$, similar distributions are expected, whereas for $AR$, the distributions are anticipated to differ. **IEC** counts ranking changes within explanation methods post-perturbation, with an ideal metric showing consistent rankings under minor noise ($NR$), and altered rankings under disruptive noise ($AR$).

on it such as SAL. Note, that not every explanation method is suitable or intended to be used for all data modalities, and/ or model architectures. For example, GradCAM explanations are primarily designed for convolutional neural networks (CNN) models, and global feature visualisation methods are generally applied to vision tasks. As such, we only report `GEF` and `Fast-GEF` results where appropriate.

**Normalisation.** We perform normalisation using the square root of the mean of the squared values (as detailed in the Appendix of (Binder et al., 2022)). This approach introduces less variance compared to normalisation techniques like scaling by the maximum value. It is defined as follows:

$$\text{norm}(\boldsymbol{e}) = \frac{\boldsymbol{e}_{h,w}}{\left(\frac{1}{HW}\sum_{h',w'} \boldsymbol{e}^2_{h',w'}\right)^{1/2}}\,, \tag{23}$$

where $H$, and $W$ represent the height, and width, respectively, and $\hat{\boldsymbol{e}}_{h,w}$ denotes the explanation value at the pixel location $(h, w)$[7].

**LLM-x Methodology.** To generate LLM-x explanations, we use `Gemma-2B-IT` (Mesnard et al., 2024) as the explainer. For each instance, we create a prompt describing the task, softmax confidence before, and after perturbation, and the class labels. The prompt template introduces the task (*e.g.,* classifying `sms messages` or `sentiment analysis`), and uses synonyms for model descriptions (*e.g.,* `"AI"`, `"machine learning"`), and perturbation types (*e.g.,* `"adversarially manipulated"`, `perturbed with noise"`) to vary language. The softmax change is calculated, and added to the template and is described in the context of the model getting `"more"` or `"less"` certain of a class label. The LLM is asked to return the top-$K$ important tokens in a structured JSON format, ranking tokens from `1` to `K`. The temperature is set to 0 for deterministic outputs.

After prompting, invalid or non-JSON outputs are removed. The LLM-ranked tokens are normalised by lowercasing, removing punctuation, and optionally stemming. Then, these tokens (or words) are encoded with the original model's tokeniser. Binary explanation vectors are created by matching the LLM-ranked tokens to the original input tokens, with a value of 1 for matching tokens, and 0 otherwise. For full details, including the code, and prompt template, we refer to our GitHub repository at `url`.

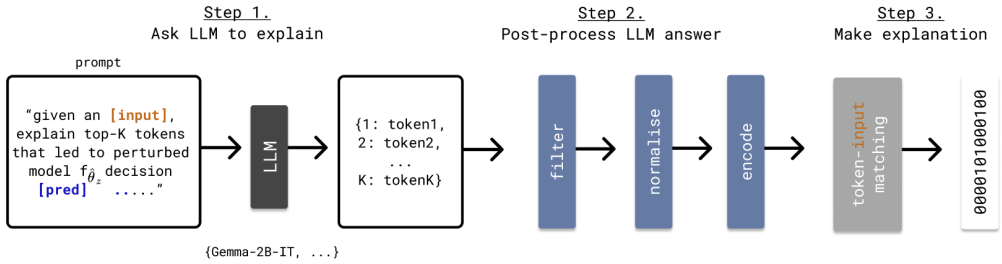



Figure A.2: A high-level overview of the three-step LLM-x methodology.

**SAE Methodology.** SAEs are designed to create sparse, interpretable representations of the internal activations of a LLM, while preserving their reconstruction. Internal representations of a given layer $l$, *i.e.,* $f_\theta(\boldsymbol{x}) \in \mathbb{R}^L$ are encoded into a sparse latent vector $z(\boldsymbol{x}) \in \mathbb{R}^O$, with latent dimensions larger than the internal representation, *i.e.,* $O \gg L$. Then, these decoded representations are reconstructed such that $\boldsymbol{g}(z) \approx f_\theta(\boldsymbol{x})$. This process is defined by the encoder and decoder functions:

$$\boldsymbol{z}(f_\theta(\boldsymbol{x})) := \sigma\left(\boldsymbol{W}_{enc} f_\theta(\boldsymbol{x}) + \boldsymbol{b}_{enc}\right), \tag{24}$$

$$\boldsymbol{g}(z) := \boldsymbol{W}_{dec}\boldsymbol{z} + \boldsymbol{b}_{dec}, \tag{25}$$

where $\sigma$ enforces sparsity through activation functions like ReLU or JumpReLU (Lieberum et al., 2024), using $L1$ or $L0$ regularisation during training. The SAE explanations are generated by performing a forward pass through the SAE encodings, and storing the activated values of $\boldsymbol{z}$.

[7]This normalisation method ensures that the mean squared distance from zero of each explanation score equals one. Unlike other normalisation techniques that constrain attribution values to a predefined range—making them suitable for visualisation—this method retains a metric useful for comparing the distances across different explanation methods.

In the following, we describe the methodology used to produce LLM-x explanations. An illustration is provided in Fig. A.2.

#### A.4.5 Evaluation Methods

Next, we mathematically define the evaluation methods (or "metrics") used in this work (Sec. 6.1).

**Faithfulness.** Within the faithfulness category, we evaluate three metrics, including, *Faithfulness Correlation* (FC) (Bhatt et al., 2020), *Pixel-Flipping* (PF) (Bach et al., 2015), and *Region-Perturbation* (RP) (Samek et al., 2017). FC is defined as follows:

$$\Psi_{\text{FC}} = \underset{S \in |S| \subseteq d}{\text{corr}} \left( \sum_{i \in S} \phi(\boldsymbol{x}, f, \hat{y}; \lambda)_i, f(\boldsymbol{x}) - f\left(\boldsymbol{x}_{[\boldsymbol{x}_s = \overline{\boldsymbol{x}}_s]}\right) \right), \tag{26}$$

where $|S| \subseteq D$ is a subset of indices of a sample $\boldsymbol{x}$, $\overline{\boldsymbol{x}}$ is the chosen baseline value, and $\boldsymbol{x}_{[\boldsymbol{x}_s = \overline{\boldsymbol{x}}_s]}$ are the masked input, with randomly chosen indices.

PF returns a vector of prediction scores $p_i$ corresponding to pixel replacements $i \in n$, which are sorted in descending order by the highest relevant pixel in the explanation $\phi(\boldsymbol{x}, f, \hat{y}; \lambda)$. To return one evaluation score per input sample, we calculate the area under the curve (AUC) as follows:

$$\Psi_{\text{PF}} = \sum_{i=1}^{n} (\hat{y}_i + \hat{y}_{i+1}) \cdot \frac{p_{i+1} - p_i}{2} \tag{27}$$

where $p_i$, and $p_{i+1}$ are the prediction values of the $i^{th}$, and $(i+1)^{th}$ perturbation step, and $\hat{y}_i$, and $\hat{y}_{i+1}$ the corresponding network prediction.

RP is established similarly to the PF metric, but follows the most-relevant-first perturbation strategy, creating consecutive perturbed samples $\hat{y}_i, \hat{y}_{i+1}$ such that for $\hat{y}_i$ perturbed pixels correspond to larger respective explanation values than the pixel perturbed in $\hat{y}_{i+1}$. Across each perturbation curve, the area over the curve is calculated, and averaged across multiple masked inputs $\hat{\boldsymbol{x}}$ as follows:

$$\Psi_{\text{RP}} = \frac{1}{L+1} \mathbb{E}_{\hat{(\boldsymbol{x})}} \left( \sum_{k=1}^{L} (\hat{y}_0 + \hat{y}_k) \right), \tag{28}$$

where $L$ is the number of perturbed features in the input.

**Robustness.** Within the robustness category, we evaluate three metrics, including, *Relative Input Stability* (RIS), *Relative Representation Stability* (RRS), *Relative Ouput Stability* (ROS) (Agarwal et al., 2022a). RIS extends (Alvarez-Melis & Jaakkola, 2018b), which is a measure of how much the explanation changes *w.r.t.* the input under slight perturbation $\hat{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{u}_i$. The change is measured as the $l_p$ norm, and the RIS metric only considers perturbations that result in the same model prediction, *i.e.,* $f(\boldsymbol{x}) = f(\hat{\boldsymbol{x}})$. It is defined as follows:

$$\Psi_{\text{RIS}} = \max_{\hat{\boldsymbol{x}}} \frac{\left\| \frac{\phi(\boldsymbol{x}, f, \hat{y}; \lambda) - \phi(\hat{\boldsymbol{x}}, f, \hat{y}; \lambda)}{\phi(\boldsymbol{x}, f, \hat{y}; \lambda)} \right\|_p}{\max\left( \left\| \frac{\boldsymbol{x} - \hat{\boldsymbol{x}}}{\boldsymbol{x}} \right\|_p, \epsilon_{min} \right)}, \ \forall \hat{\boldsymbol{x}} \in \mathcal{N}_\epsilon; f(\boldsymbol{x}) = f(\hat{\boldsymbol{x}}) \tag{29}$$

where $\epsilon_{min} > 0$ ensures a non-zero denominator.

In contrast to the RIS metric, RRS considers the internal representation of the model $\mathcal{L}(\cdot)$ (*e.g.,* an output embedding), while maintaining similar perturbation conditions:

$$\Psi_{\text{RRS}} = \max_{\hat{\boldsymbol{x}}} \frac{\left\| \frac{\phi(\boldsymbol{x}, f, \hat{y}; \lambda) - \phi(\hat{\boldsymbol{x}}, f, \hat{y}; \lambda)}{\phi(\boldsymbol{x}, f, \hat{y}; \lambda)} \right\|_p}{\max\left( \left\| \frac{\mathcal{L}_{\boldsymbol{x}} - \mathcal{L}_{\hat{\boldsymbol{x}}}}{\mathcal{L}_{\boldsymbol{x}}} \right\|_p, \epsilon_{min} \right)}, \ \forall \hat{\boldsymbol{x}} \in \mathcal{N}_\epsilon; f(\boldsymbol{x}) = f(\hat{\boldsymbol{x}}) \tag{30}$$

where $\epsilon_{min} > 0$ ensures a non-zero denominator.

ROS makes similar adaptations as the RRS metric, assumes however that the model's internal representations are not accessible. Instead the output logits $h(\boldsymbol{x})$, and $h(\hat{\boldsymbol{x}})$ are assessed:

$$\Psi_{\text{ROS}} = \max_{\hat{\boldsymbol{x}}} \frac{\left\| \frac{\phi(\boldsymbol{x}, f, \hat{y}; \lambda) - \phi(\hat{\boldsymbol{x}}, f, \hat{y}; \lambda)}{\phi(\boldsymbol{x}, f, \hat{y}; \lambda)} \right\|_p}{\max\left( \left\| h(\boldsymbol{x}) - h(\hat{\boldsymbol{x}}) \right\|_p, \epsilon_{min} \right)}, \ \forall \hat{\boldsymbol{x}} \in \mathcal{N}_\epsilon; f(\boldsymbol{x}) = f(\hat{\boldsymbol{x}}) \tag{31}$$

where $\epsilon_{min} > 0$ ensures a non-zero denominator.

**Sensitivity.** Within the sensitivity category, we evaluate three metrics, including *Model Parameter Randomisation Test* (MPRT) (Adebayo et al., 2018), *Smooth Model Parameter Randomisation Test* (sMPRT), *Efficient Model Parameter Randomisation* (eMPRT) (Hedström et al., 2024). MPRT computes a quality estimate $\hat{q} \in \mathbb{R}$ measuring the similarity between the original explanation $\boldsymbol{e}_l$, and the explanation $\hat{\boldsymbol{e}} := \phi(\boldsymbol{x}, \hat{f}_l^t, y)$ of the perturbed model $\hat{f}_l^t$ randomised in a top-down fashion up to layer $l \in [L, L-1, \dots, 1]$:

$$\hat{q}^{\text{MPRT}} = \rho(\boldsymbol{e}, \hat{\boldsymbol{e}}_l), \tag{32}$$

with similarity function $\rho : \mathbb{R}^D \times \mathbb{R}^D \mapsto \mathbb{R}$.

sMPRT computes a quality estimate $\hat{q} \in \mathbb{R}$ between explanations $\boldsymbol{e}_i := \phi(\hat{\boldsymbol{x}}_i, f, y; \lambda)$, and $\hat{\boldsymbol{e}}_{l,i} := \phi(\hat{\boldsymbol{x}}_i, \hat{f}_l^b, y; \lambda)$ averaged over $i \in [1, N]$ where $\hat{\boldsymbol{e}}_{l,i}$ corresponds to the perturbed model $\hat{f}_l^b$ randomised in a bottom-down fashion up to layer $l \in [1, 2, \dots, L]$:

$$\hat{q}^{\text{sMPRT}} = \rho\left( \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{e}_i, \frac{1}{N} \sum_{i=1}^{N} \hat{\boldsymbol{e}}_{l,i} \right), \tag{33}$$

with $\hat{\boldsymbol{x}}_i = \boldsymbol{x} + \eta_i$, and $\eta_i \sim \mathcal{N}(0, \sigma)$ with $||\eta_i||_p \leq \epsilon$ holding with high probability, for $\sigma, \epsilon \in \mathbb{R}$.

eMPRT computes a quality estimate $\hat{q} \in \mathbb{R}$ measuring the relative rise in the complexity of the explanation from a fully randomised model $\hat{f}$ such that $\hat{\boldsymbol{e}} := \phi(\boldsymbol{x}, \hat{f}, y; \lambda)$:

$$\hat{q}^{\text{eMPRT}} = \frac{c(\hat{\boldsymbol{e}}) - c(\boldsymbol{e})}{c(\boldsymbol{e})} \tag{34}$$

where $c : \mathbb{R}^D \mapsto \mathbb{R}$ is a complexity function, *e.g.,* discrete entropy.

### A.5 Analysing Violations of Model Assumptions

To understand whether perturbation techniques commonly employed for robustness, sensitivity, and faithfulness evaluations generally fulfill the critical assumptions of model distortion (Ass. 1-3), we performed several experiments. To investigate how often model robustness, sensitivity, and faithfulness (Ass. 1-3) hold versus fail in practice, we set up a simple experiment that tracks model, and explanation distortions, *i.e.,* $\boldsymbol{D}_f$, and $\boldsymbol{D}_\phi$, while applying perturbation commonly used in evaluation such as additive Gaussian noise for robustness evaluation, top-down, and bottom-up layer-by-layer parameter randomisation for sensitivity evaluation, and cumulative masking for faithfulness evaluation.

**Model Robustness Under Additive Noise.** To understand the extent to which model robustness (Ass. 1) is generally satisfied for robustness evaluation (Def. 2), we examine Fig. A.3 (A), (B), and (C). Here, the distribution of $\boldsymbol{D}_f$ is visualised over $Z = 10$ input perturbation steps, showing how model distortion varies with increasing input perturbation magnitude, using additive Gaussian noise, *i.e.,* , $\nu_i \sim \mathcal{N}(0, \sigma)$ to generate perturbed inputs $\hat{\boldsymbol{x}}_i = \boldsymbol{x} + \nu_i$, with $\sigma$ increasing until the model behaves randomly (*i.e.,* accuracy $= 1/C$). While the average trend (*blue* line) indicates that larger perturbation causes higher model distortion, sample-wise exceptions frequently appear. In Fig. **??**, random sample trajectories reveal both correlated and uncorrelated patterns between perturbation levels and model distortions. This is a key observation, as it implies that model robustness cannot be assured by a general threshold without inspecting each evaluation sample individually.

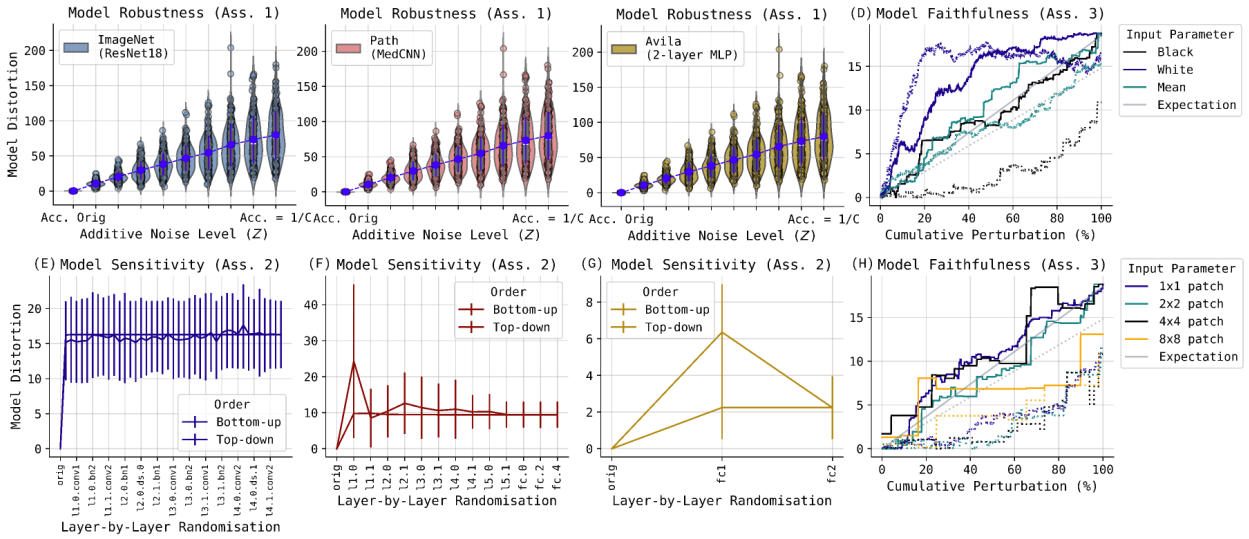



Figure A.3: Impact of model distortion (y-axis) over common perturbation types in robustness, sensitivity, and faithfulness evaluations, across different datasets, and NN architectures. (A), (B), and (C) depict the distribution of model distortions across different perturbation magnitudes of additive Gaussian noise for ImageNet (ResNet18), Path (MedCNN), and Avila (2-layer MLP), respectively. (D), (E), and (F) show the average, and standard deviation of model distortions over layer-wise top-down, and bottom-up randomisation for the same datasets (as indicated by colour). (G), and (H) display model distortions for randomly chosen MNIST (*solid* line), and fMNIST (*dashed* line) samples (LeNet) under cumulative perturbations using different patch sizes ($1 \times 1$, $2 \times 2$, $4 \times 4$, $8 \times 8$), and baseline replacement strategies (*black, white, mean*).

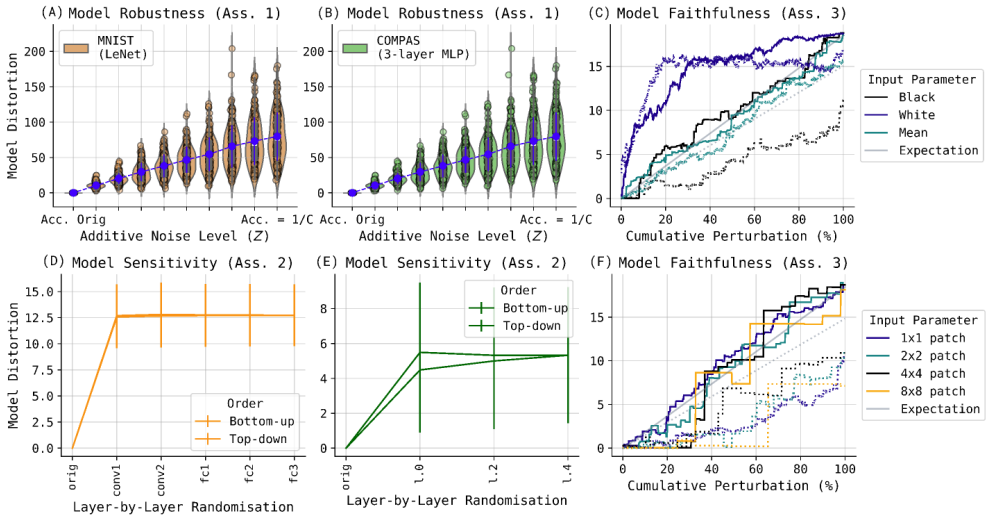



Figure A.4: Impact of model distortion (y-axis) over common perturbation types in robustness, sensitivity, and faithfulness evaluations, across different datasets, and NN architectures. (A), and (B) depict the distribution of model distortions across different perturbation magnitudes of additive Gaussian noise for MNIST (LeNet), and COMPAS (2-layer MLP), respectively. (D), and (E) show the average, and standard deviation of model distortions over layer-wise top-down, and bottom-up randomisation for the same datasets (as indicated by colour). (C), and (F) display model distortions for randomly chosen MNIST (*solid* line), and fMNIST (*dashed* line) samples (LeNet) under cumulative perturbations using different patch sizes (*1x1, 2x2, 4x4, 8x8*), and baseline replacement strategies (*black, white, mean*).

**Model Sensitivity Under Layer-by-Layer Randomisation.** For sensitivity evaluations (Def. 3) to be meaningful, the model distortion caused by perturbation must be significant (Ass. 2). To test this practice, we perform consecutive layer-wise model parameter randomisation; in both a top-down (Adebayo et al., 2018), and bottom-up (Hedström et al., 2024) manner. From Fig. A.3 (E), (F), and (G), we observe that, although model distortion generally increases with layer-wise randomisation, there are exceptions of non-

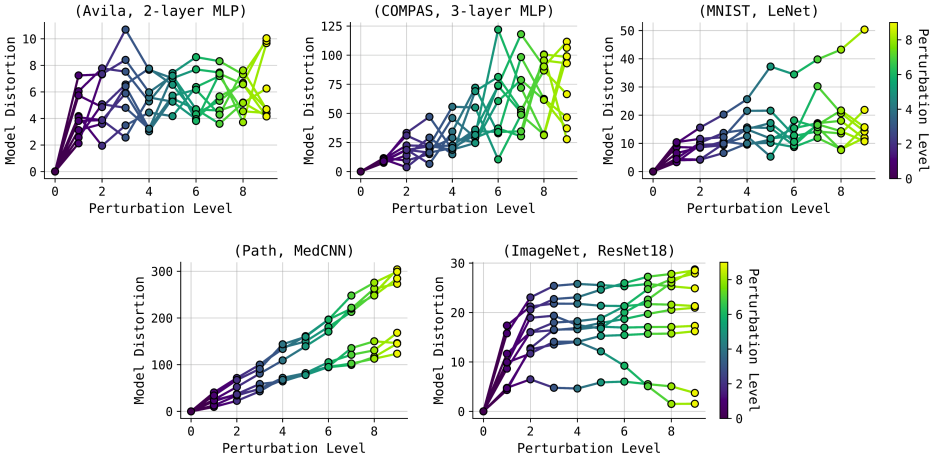



Figure A.5: Sample-wise trajectories of model distortion (*y-axis*) across different perturbation magnitudes of additive Gaussian noise. Each panel includes line plots across $N = 10$ samples, showing both correlated and uncorrelated outcomes.

monotonicity (see, *e.g.,* Path, and Avila results in Fig. A.3 (E), and (F), respectively). The high standard deviation (see the error bars) suggests that layer-wise randomisation fails to predictably dictate the degree of model distortion, undermining the assumption that significant model distortions will always occur in sensitivity evaluations.

**Model Faithfulness Under Cumulative Input Perturbation.** To investigate whether model distortion increases monotonically under cumulative input perturbation (Ass. 3), we measure $\boldsymbol{D}_f$ using a standard "pixel-flipping" faithfulness procedure (Bach et al., 2015). By randomising the perturbation order, the resulting faithfulness curve should reflect *only* the model's response; any deviation from a linear trend suggests that Ass. 3 is failed. Observing Fig. A.3 (D), and (H), we see that neither patch size (top) nor replacement strategy (bottom) induces monotonic non-decreasing model behaviour. While these results are expected due to the model's inherent nonlinearity, and OOD effects (Hase et al., 2021; Hesse et al., 2024), it is not accounted for in the faithfulness evaluation itself (Def. 4). When genuine signals (*i.e.,* explanation quality) are not decoupled from noise (*i.e.,* non-monotonic model behaviour), interpretations may become biased (Hooker et al., 2019; Brocki & Chung, 2022; Brunke et al., 2020).

Together, these results reveal how easily, and systematically Ass. 1-3 are violated by perturbation strategies commonly applied in practice (Sec. 2.1.1). Our findings are consequential as they demonstrate that the validity of existing robustness, sensitivity, and faithfulness evaluations (Def. 2-4) are frequently undermined. As displayed in Fig. A.3, there are many sample-wise exceptions where the perturbation magnitude, and the model distortion quantity are not strictly monotonically related, challenging the assumption that increased perturbations lead to proportionally greater distortions, and vice-versa.

### A.5.1 Issues with Cumulative Input Perturbation

If small changes in input parameters, cause large variations in evaluation outcomes, evaluation reliability is compromised. Corroborating previous studies (Brunke et al., 2020; Brocki & Chung, 2022; Rong et al., 2022), the varied faithfulness curves in Fig. A.3 (G), and (H) demonstrate how input parameter choices, such as patch size or pixel value, can drastically influence the evaluation outcomes across tasks, *i.e.,* act as evaluation confounds (*cf.* the same parameter for MNIST *solid* line vs. fMNIST *dotted* line). These variations between tasks expose a simple, yet systematically overlooked issue in faithfulness evaluations: that parameter choices to perturb the input inherently introduce task-specific biases to the evaluation. Attempts to mitigate these biases—using inverse curves (Blücher et al., 2024) or assessing the OOD impact of perturbations (Qiu et al., 2021; Haug et al., 2021)—fail to address the core problem: that evaluation methods (Sec. 2.1.2) that require input parameters to be tuned according to its task, are inherently biased, impeding impartial comparisons across tasks, and explanation approaches.

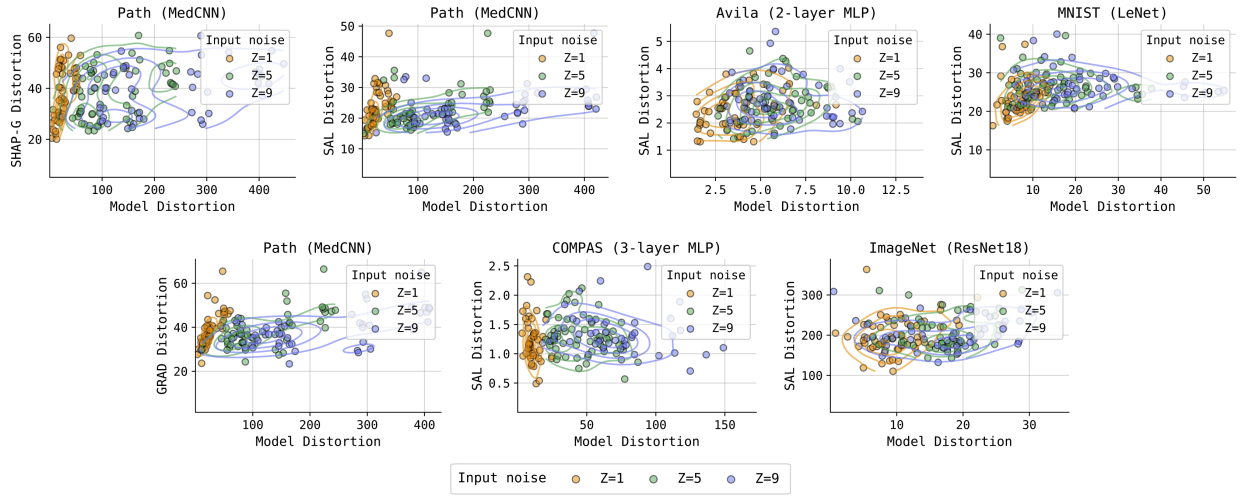



Figure A.6: Model (x-axis), and explanation distortions (y-axis) under varying levels of additive Gaussian input noise for vision and tabular tasks, as indicated in the titles. Scatter points represent individual samples, coloured by perturbation magnitude ($Z = 10$), with overlapping contours highlighting the relative alignment patterns. The individual plots contain SHAP-G, GRAD, and SAL explanations, as indicated by the y-axis labels.

## A.6 Alignment Patterns, and Extended Results

Fig. A.6 provides complementary results to Fig. 3 in the main manuscript. The scatter points are coloured by perturbation magnitude up to $Z = 10$, using additive Gaussian noise. The varying but consistent overlaps of points of high and low perturbation magnitudes alongside the almost uniform distribution along the y-axis, illustrate that perturbation effects are not guaranteed to have a proportional effect on the model and explanation functions. Thus, using the perturbation magnitude as an indicator of the magnitude of which the explanation should change (as done in existing evaluations, see Defs. 2-3) is not reliable.

Next, we show the change in the relationship between model distortion and explanation distortion when comparing the commonly used input perturbations with model perturbation (Fast-GEF) and when measuring explanation distortion and model distortion on the same geometry (GEF). Fig. A.7 illustrates the relationship of model and explanation distortion across these three cases from left to right. Each contour plot includes $N = 10$ samples per perturbation magnitudes from $Z = 1$ to $Z = 5$. The scatter points are colored to one of three increasing perturbation magnitudes ($Z = 1$ to $Z = 3$). We can observe the distortion quantities across tasks with increasing model complexity, from a simpler tabular task (first row) to a highly parameterised model for a vision task (last row).

As expected, the first column (input perturbation) coincides with Fig. A.6 and yields the same findings. However, in the following two columns, we can observe how Fast-GEF and GEF behave in practice. Most notably we observe that the alignment between model and explanation distortion changes from less complex to more complex task and also compared to more complex tasks. Furthermore, we find that Fast-GEF tend to generate higher coherence compared to input perturbation, with the exception of ImageNet. Lastly, GEF yields more coherent distortions of the explanation and the model. While these findings appear to support both our approaches, it is important to note that without access to ground truth, it is unclear whether the contour plots should show stronger coherence. In our approaches, we focus on coherence as a measure of the quality of an explanation, and worsened coherence can also indicate an unfaithful explanation.

## A.7 Ablation Study

To better understand the influence of the hyperparameters, on the proposed `GEF` evaluation method, we conducted an ablation study. We employed two tasks, *i.e.,* a tabular dataset (Avila) using a 2-layer MLP model with SAL explanations and a vision dataset (MNIST) using a LeNet model with 250 random explanations, sampled from a uniform distribution, *i.e.,* $\hat{\boldsymbol{e}}_i = \mathcal{U}(0, 1)$. For each hyperparameter, *i.e.,* the number

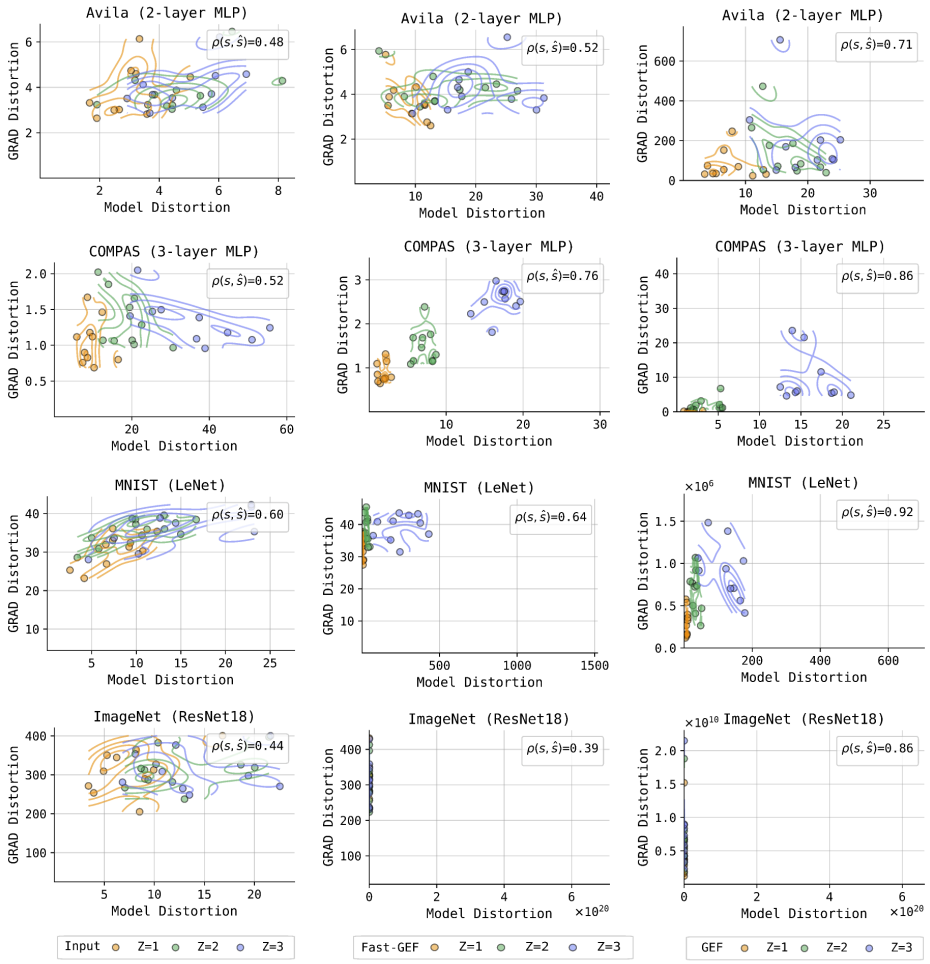



Figure A.7: Each plot shows the model (x-axis), and explanation distortions (y-axis) under different types of noise for *Gradient* (GRAD) explanation. The first column shows distortion outcomes after applying additive Gaussian input noise. The second and third columns show distortion outcomes after applying model parameter scaling (Sec. 5.1). The second column computes explanation distortion using `Fast-GEF` and the third column computes distortion using `GEF` (*i.e.,* with pullback mechanism). Scatter points represent individual samples, coloured by perturbation magnitude (z=1, z=2, z=3), with $Z = 5$ number of steps.

of perturbed models $M$, the length of the perturbation path $Z$, the number of summation steps $T$, and the number of samples $K$, we enumerated over values from 0 to 20, while fixing the others at a default value of 10. For each configuration, we recorded mean (solid line) and standard deviation (shaded area) of the model distortion, explanation distortion, Jacobian quantity, and the mean computation time.

Fig. A.8 demonstrate the results for both the tabular (*left*) and the vision task (*right*). As can be observed by the converging values of the standard deviation and means, the hyperparameters are resilient to key parameter changes once parameter values reach 5 or higher. The Jacobian variance reflects the curvature captured in its estimate. While the variance increases for parameter values above 5, the mean stabilises, indicating diminishing returns in capturing additional curvature. At parameter values of 5, the majority of the curvature is already captured, providing a practical trade-off between computational efficiency and quality of approximation. Considering computational time, all parameters lead to a linear, non-negligible increase. Among them, $Z$ and $M$ are identified as the primary drivers of time. Based on these experimental findings, setting $K = T = Z = M = 5$ balances computational efficiency and stability.

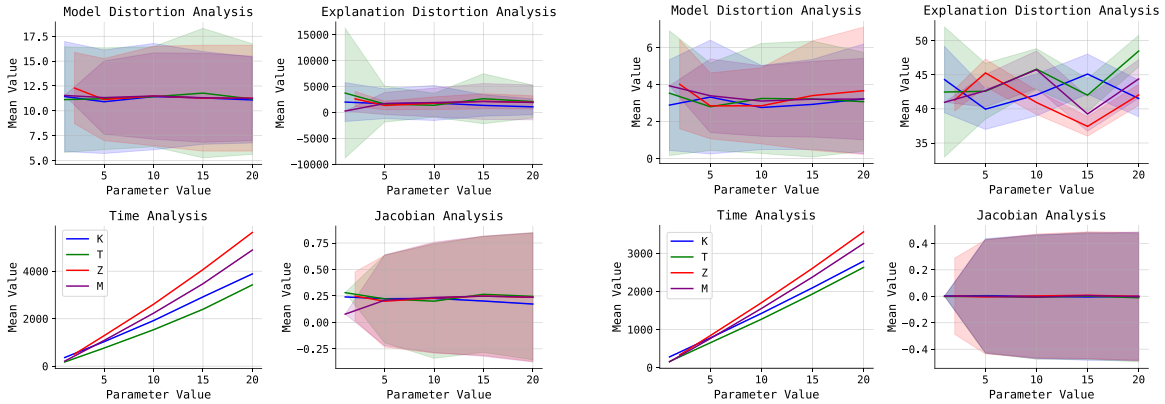



Figure A.8: Ablation study results across hyperparameters $M$, $Z$, $T$, and $K$ for two tasks: (*left*) Saliency explanation on Avila (2-layer MLP), and (*right*) Random explanation on MNIST (LeNet). The mean value (*solid* line) and variance (*shaded* area) are reported. The time analysis is measured in seconds.

## A.8 Experiments, and Extended Results

This section provides descriptions of experimental setups, and extended results, including meta-evaluations, and agreement between different scoring methods. Additionally, we present further results for random control variant sanity checks, cross-domain benchmarking, and LLM-x methodology, and extended results.

### A.8.1 Meta-Evaluation

To employ the scoring methodology (Sec. A.3), we used the pre-existing test suite available in the `MetaQuantus` library[8] with their pre-defined hyperparameters.

**`MetaQuantus` Hyperparameters.** We applied these metrics over $K = 5$ perturbations, conducting 3 iterations with the test configurations specified in the library for two different sets of explanation methods, namely {GRAD, G-CAM}, and {SAL, SHAP-G}, which were evaluated by each metric. The explanation method groups were created by randomly selecting methods from the complete set of available methods, ensuring consistency across various experimental setups, such as dataset, and model combinations. In terms of choosing $K$, and the number of iterations, we followed the recommendations from the original study to keep the standard deviation between different sets relatively low. To ensure a fair comparison across metrics, all shared hyperparameters were assigned the same values.

**Metrics Hyperparameters.** All metrics have been implemented in `Quantus` (Hedström et al., 2023b). Different hyperparameters were chosen for the individual metrics based on the dataset. For the robustness metrics, we use 5 noisy samples, and employ additive Gaussian noise such that $\nu \sim \mathcal{N}(0, 0.001)$. For the faithfulness metrics, we use 28 features per perturbation step, and a patch size of 7 for the MNIST, and fMNIST datasets. For ImageNet, we set the number of features to 896, and the patch size to 28. For FC, similar to the robustness metrics, we let it run 5 times. For the sensitivity metrics, namely MPRT, and sMPRT, we use a noise magnitude of 0.01 for each sample, and sMPRT uses 5 samples in its calculation. For all sensitivity metrics, we use the Spearman rank correlation coefficient.

**Extended Results.** In Tabs. A.1, and A.2, we provide the corresponding results for Fig. 6.

### A.8.2 Agreement between `GEF`, and `Fast-GEF`

To determine whether the simpler, computationally efficient `Fast-GEF` method can serve as an alternative to the more exact but computationally intensive `GEF` method, we compare the agreement between their

[8]Code is provided at `https://github.com/annahedstroem/MetaQuantus/`.

Table A.1: MC scores and standard deviation for unified, and faithfulness methods listed in A.4.5 for ImageNet, MNIST, and fMNIST datasets. The final row shows the mean score for each metric across the datasets. Values range between $[0, 1]$, with higher values indicating better outcomes. Due to computational constraints, `GEF` scores are only computed for fMNIST, and MNIST datasets.

| | Unified | | Faithfulness | | |
|---|---|---|---|---|---|
| | `GEF` | `Fast-GEF` | `PF` | `FC` | `RP` |
| ImageNet | NAN ± NAN | 0.78 ± 0.02 | 0.63 ± 0.01 | 0.51 ± 0.02 | 0.63 ± 0.06 |
| MNIST | 0.75 ± 0.07 | 0.74 ± 0.03 | 0.61 ± 0.04 | 0.63 ± 0.03 | 0.59 ± 0.03 |
| fMNIST | 0.71 ± 0.07 | 0.71 ± 0.03 | 0.63 ± 0.01 | 0.50 ± 0.04 | 0.58 ± 0.09 |
| **Mean** | **0.73 ± 0.07** | **0.74 ± 0.03** | **0.62 ± 0.02** | **0.56 ± 0.03** | **0.59 ± 0.06** |

Table A.2: MC scores and standard deviation for sensitivity, and robustness methods listed in A.4.5 for ImageNet, MNIST, and fMNIST datasets. The final row shows the mean score for each metric across the datasets. Values range between $[0, 1]$, with higher values indicating better outcomes.

| | Sensitivity | | | Robustness | | |
|---|---|---|---|---|---|---|
| | `MPRT` | `sMPRT` | `eMPRT` | `RIS` | `ROS` | `RRS` |
| ImageNet | 0.71 ± 0.02 | 0.69 ± 0.04 | 0.71 ± 0.02 | 0.72 ± 0.06 | 0.76 ± 0.07 | 0.75 ± 0.04 |
| MNIST | 0.63 ± 0.02 | 0.66 ± 0.04 | 0.76 ± 0.03 | 0.73 ± 0.02 | 0.70 ± 0.09 | 0.74 ± 0.09 |
| fMNIST | 0.63 ± 0.01 | 0.67 ± 0.05 | 0.67 ± 0.05 | 0.70 ± 0.02 | 0.77 ± 0.06 | 0.70 ± 0.03 |
| **Mean** | **0.64 ± 0.02** | **0.67 ± 0.04** | **0.71 ± 0.03** | **0.72 ± 0.03** | **0.74 ± 0.07** | **0.73 ± 0.05** |

respective faithfulness estimates. For a subset of explanation methods, and tasks (see Tab. 2), we thus compute scores, and rank explanation methods from R1 to RN. While it is expected that estimates from the two methods differ, a high agreement in a categorical ranking would make `Fast-GEF` a practical alternative in resource-constrained environments. Experimental details are found in Appendix A.8.4.

**Results.** Fig. A.9 (A) visually compares how `GEF` and `Fast-GEF` ranks (x-axis) each explanation method in terms of increases (y-axis), highlighting the relative agreement between them. The explanations in the tabular, and text tasks show perfect ranking agreement. In the MNIST vision task, with minimal nominal differences, GRAD, and SHAP-G methods disagree in their ranking (R1, and R2), but such disagreement can be expected acknowledging the algorithmic similarity between these explanation methods. In the Derma vision task, the same pattern is observed, yet with a slightly larger difference for the global method FO-50. Interestingly, we observe that nominal differences are pronounced for global methods (DV-50, and FO-50), and that `Fast-GEF` tends to generate slightly lower faithfulness estimates *cf.* `GEF`.

### A.8.3 Scoring Control Variants

To ensure the reliability of out method, we validate that both `GEF` and `Fast-GEF` assign low faithfulness scores to different control variant explanations. In our sanity checks, we evaluate explanations generated by uniform sampling, *i.e.,* $\hat{\boldsymbol{e}}_i \sim \mathcal{U}(0, 1)$, a constant value, *i.e.,* $\hat{\boldsymbol{e}}_i = \mathbf{0}$, and with a model-independent Sobel

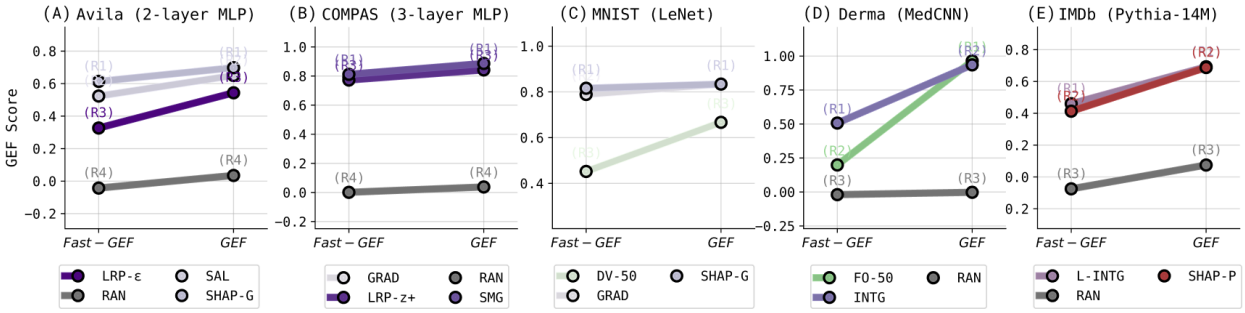



Figure A.9: (A) to (E) illustrates the GEF scores of `GEF` and `Fast-GEF` (with $M = 1$) for various explanation methods, and tasks. Explanation methods are ranked between R1 to RN, in descending order.

filter. For non-random reference, we evaluate GRAD explanations for the predicted class of the Derma task (see Tab. 2) (Sobel et al., 1968). For comparability, we extend this sanity check exercise to one metric per evaluative criteria, *i.e.,* `FC` (faithfulness), `MPRT` (sensitivity), and `RIS` (robustness). Hyperparameters are provided in Appendix A.8.1.

Table A.3: Evaluation scores of Derma (MedCNN) explanations for three random, and one regular (GRAD) explanation. The arrow (↑,↓) indicates whether higher or lower values are better. NaN indicates no score is produced.

| Explanation | `GEF` (↑) | `Fast-GEF` (↑) | `FC` (↑) | `MPRT` (↓) | `RIS` (↓) |
|---|---|---|---|---|---|
| Control Var. Constant | nan ± nan | nan ± nan | nan ± nan | nan ± nan | 0.11 ± 0.29 |
| Control Var. Random Uniform | -0.01 ± 0.30 | -0.01 ± 0.22 | -0.00 ± 0.51 | -0.00 ± 0.04 | 3.21 ± 2.89 |
| Control Var. Sobel Filter | nan ± nan | nan ± nan | -0.01 ± 0.50 | 1.00 ± 0.00 | 82197.21 ± 132718.26 |
| GRAD | 0.47 ± 0.23 | 0.48 ± 0.15 | -0.05 ± 0.49 | 0.01 ± 0.04 | 1764.60 ± 10007.26 |

**Results.** Tab. A.3 presents the results. Some metrics produce no values (nan), *e.g.,* when correlating identical vectors, and by that identify the unfaithful explanation. `Fast-GEF`, and `GEF` consistently assign low scores to random explanations, and high scores to non-random GRAD explanations, indicating their ability to identify the control explanations. Conversely, other metrics fail at least in one random test, either showing little discrepancy between regular, and control variants or even giving higher scores to the control. For instance, `MPRT`, and `RIS` score random uniform explanations as good or better than regular ones.

### A.8.4 Cross-Domain Benchmarking

**Extended Results.** We benchmark various local, and global explanation methods with `GEF` and `Fast-GEF`. In Fig. A.10, we extend the results in Fig. 8.

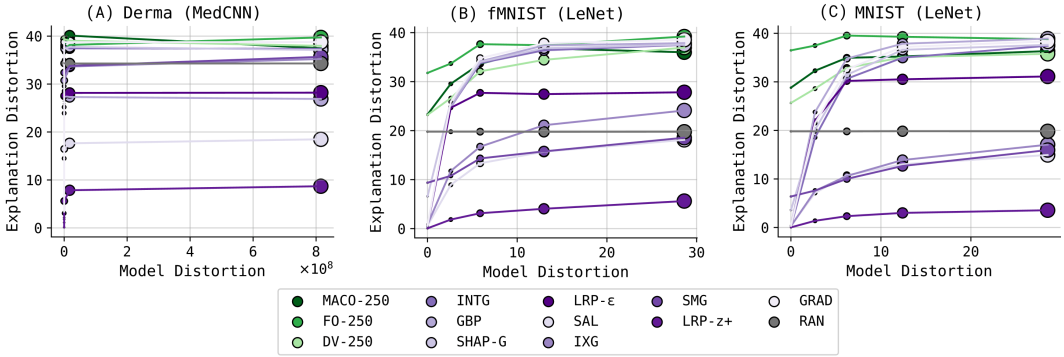



Figure A.10: `Fast-GEF` results for vision tasks. (A), (B), and (C) plot the model and explanation distortion for Derma (MedCNN), and fMNIST (LeNet), and MNIST (LeNet) along the perturbation path with $Z = 5$ perturbation steps. The size of the scatter point represents each perturbation steps, from 1 to 5.

The results presented in Fig. 7 are provided in Tabs. A.4, and A.5.

### A.8.5 LLM-x

In the following, we provided extended results of the LLM-x experiments.

**Extended Results.** The results presented in Fig. 9 are provided in Tabs. **??** and **??**.

Table A.4: GEF results on local methods for tabular tasks. Mean faithfulness scores, and standard errors are reported, with higher values indicating better quality.

| | Task | Adult (3-layer MLP) | Adult LR | Avila (2-layer MLP) | Compas (3-layer MLP) | Compas LR |
|---|---|---|---|---|---|---|
| Local Methods | SMG | 0.86 ± 0.00 | 0.69 ± 0.00 | 0.72 ± 0.01 | 0.81 ± 0.01 | 0.73 ± 0.01 |
| | SHAP-G | 0.78 ± 0.00 | 0.84 ± 0.01 | 0.75 ± 0.01 | 0.84 ± 0.00 | 0.66 ± 0.01 |
| | SAL | 0.84 ± 0.00 | 0.76 ± 0.00 | 0.69 ± 0.01 | 0.75 ± 0.01 | 0.70 ± 0.01 |
| | RAN | -0.00 ± 0.02 | 0.00 ± 0.02 | 0.00 ± 0.02 | 0.02 ± 0.02 | 0.02 ± 0.02 |
| | LRP-$\varepsilon$ | 0.66 ± 0.01 | 0.61 ± 0.01 | 0.52 ± 0.01 | 0.74 ± 0.01 | 0.59 ± 0.01 |
| | LRP-$z^+$ | 0.79 ± 0.00 | 0.75 ± 0.00 | 0.66 ± 0.01 | 0.78 ± 0.00 | 0.70 ± 0.01 |
| | IXG | 0.84 ± 0.00 | 0.77 ± 0.01 | 0.69 ± 0.01 | 0.74 ± 0.00 | 0.72 ± 0.00 |
| | INTG | 0.82 ± 0.00 | 0.82 ± 0.00 | 0.69 ± 0.01 | 0.81 ± 0.00 | 0.80 ± 0.00 |
| | GRAD | 0.86 ± 0.00 | 0.74 ± 0.00 | 0.69 ± 0.01 | 0.81 ± 0.01 | 0.67 ± 0.01 |
| | GBP | 0.80 ± 0.00 | 0.60 ± 0.01 | 0.68 ± 0.01 | 0.78 ± 0.01 | 0.70 ± 0.01 |

Table A.5: Fast-GEF result on local methods for vision tasks. Mean faithfulness scores, and standard errors are reported, with higher values indicating better quality.

| | Task | Derma MedCNN | fMNIST LENET | Imagenet-1k Resnet18 | MNIST LENET | Path MedCNN |
|---|---|---|---|---|---|---|
| Local Methods | SMG | 0.61 ± 0.01 | 0.69 ± 0.01 | 0.63 ± 0.01 | 0.73 ± 0.01 | 0.63 ± 0.02 |
| | SHAP-G | 0.49 ± 0.01 | 0.74 ± 0.01 | 0.69 ± 0.01 | 0.77 ± 0.01 | 0.60 ± 0.01 |
| | SAL | 0.67 ± 0.01 | 0.76 ± 0.01 | 0.71 ± 0.01 | 0.74 ± 0.01 | 0.65 ± 0.01 |
| | RAN | 0.01 ± 0.01 | 0.00 ± 0.01 | -0.00 ± 0.01 | -0.00 ± 0.01 | -0.01 ± 0.01 |
| | LRP-$\varepsilon$ | 0.45 ± 0.01 | 0.70 ± 0.01 | 0.35 ± 0.01 | 0.73 ± 0.01 | 0.66 ± 0.01 |
| | LRP-$z^+$ | 0.74 ± 0.01 | 0.73 ± 0.01 | 0.91 ± 0.00 | 0.73 ± 0.01 | 0.71 ± 0.01 |
| | IXG | 0.73 ± 0.01 | 0.72 ± 0.01 | 0.64 ± 0.01 | 0.73 ± 0.01 | 0.71 ± 0.01 |
| | INTG | 0.49 ± 0.01 | 0.73 ± 0.01 | 0.78 ± 0.01 | 0.77 ± 0.01 | 0.71 ± 0.01 |
| | GRAD | 0.54 ± 0.01 | 0.76 ± 0.01 | 0.71 ± 0.01 | 0.79 ± 0.01 | 0.66 ± 0.01 |
| | GBP | 0.63 ± 0.01 | 0.76 ± 0.01 | 0.85 ± 0.01 | 0.77 ± 0.01 | 0.64 ± 0.01 |
| Global Methods | MACO-50 | 0.16 ± 0.02 | 0.47 ± 0.02 | 0.29 ± 0.02 | 0.29 ± 0.02 | 0.42 ± 0.01 |
| | MACO-250 | 0.30 ± 0.02 | 0.54 ± 0.01 | 0.31 ± 0.01 | 0.31 ± 0.02 | 0.40 ± 0.01 |
| | MACO-100 | 0.24 ± 0.01 | 0.52 ± 0.01 | 0.31 ± 0.01 | 0.41 ± 0.01 | 0.45 ± 0.01 |
| | FO-50 | 0.17 ± 0.02 | 0.42 ± 0.01 | 0.22 ± 0.02 | 0.26 ± 0.02 | 0.35 ± 0.01 |
| | FO-250 | 0.36 ± 0.01 | 0.38 ± 0.02 | 0.19 ± 0.02 | 0.15 ± 0.02 | 0.31 ± 0.01 |
| | FO-100 | 0.28 ± 0.01 | 0.36 ± 0.02 | 0.23 ± 0.02 | 0.21 ± 0.02 | 0.27 ± 0.01 |
| | DV-50 | 0.38 ± 0.01 | 0.54 ± 0.02 | 0.26 ± 0.02 | 0.36 ± 0.02 | 0.45 ± 0.01 |
| | DV-250 | 0.44 ± 0.01 | 0.50 ± 0.01 | 0.40 ± 0.02 | 0.40 ± 0.02 | 0.48 ± 0.01 |
| | DV-100 | 0.43 ± 0.02 | 0.49 ± 0.02 | 0.40 ± 0.02 | 0.43 ± 0.02 | 0.51 ± 0.01 |

Table A.6: Fast-GEF results on LLM-x, and local methods for top-$K$ tasks. Mean faithfulness scores, and standard errors are reported, with higher values indicating better quality.

| | Task | SMS Spam BERT-TINY FT | SST2 BERT-TINY FT |
|---|---|---|---|
| Local Methods | SHAP-P-5 | 0.62 ± 0.01 | 0.75 ± 0.01 |
| | SHAP-P-10 | 0.62 ± 0.01 | 0.75 ± 0.01 |
| | RAN-5 | 0.08 ± 0.01 | -0.08 ± 0.01 |
| | RAN-10 | 0.03 ± 0.01 | -0.10 ± 0.01 |
| | LLM-X-5 | 0.06 ± 0.02 | 0.05 ± 0.02 |
| | LLM-X-10 | 0.05 ± 0.02 | 0.08 ± 0.02 |
| | L-INTG-5 | 0.58 ± 0.01 | 0.77 ± 0.01 |
| | L-INTG-10 | 0.58 ± 0.01 | 0.77 ± 0.01 |

Table A.7: Fast-GEF results on LLM-x, and local methods for top-$K$ tasks. Mean faithfulness scores, and standard errors are reported, with higher values indicating better quality.

| Task | SMS Spam BERT-TINY FT | SST2 BERT-TINY FT |
|---|---|---|
| LLM-X | -4.25 ± 8.42 | -3.73 ± 7.72 |
| L-INTG | 195.49 ± 2.91 | 238.15 ± 2.98 |
| SHAP-P | 185.95 ± 3.06 | 230.27 ± 3.75 |

## A.9 Notation Tables

A comprehensive list of all notations used in this paper is provided in the notation tables below.

**Spaces, and Elements**

| | |
|---|---|
| $\mathcal{X},\ \boldsymbol{x}$ | The input space $\mathcal{X} \subseteq \mathbb{R}^D$ with a sample $\boldsymbol{x} \in \mathcal{X}$ |
| $\mathcal{F},\ \theta$ | The model space $\mathcal{F} \subseteq \mathbb{R}^U$ with parameters $\theta \in \mathcal{F}$ |
| $\mathcal{Y},\ \boldsymbol{y}$ | The function output space $\mathcal{Y} \subseteq \mathbb{R}^C$ with logits $\boldsymbol{y} \in \mathcal{Y}; \boldsymbol{y} = [y_1, \ldots, y_C]^T$ for $C$ classes $y_c \in \boldsymbol{y} \forall c \in [1, C]$ |
| $\mathcal{E},\ \boldsymbol{e}$ | The explanation space $\mathcal{E} \subseteq \mathbb{R}^V$ with an explanation $\boldsymbol{e} \in \mathcal{F}$ |
| $\mathcal{Q},\ q$ | The evaluation space $\mathcal{Q} \subseteq \mathbb{R}^M$ with a quality estimate $q \in \mathcal{Q}$ |
| $\mathcal{S},\ \boldsymbol{s}$ | A set of spaces $S \subset \{\mathcal{X}, \mathcal{F}, \mathcal{Y}, \mathcal{E}, \mathcal{Q}\}$ where $\mathcal{S} \subseteq \mathbb{R}^S, \mathcal{S} \in \mathbb{N}$ with $\boldsymbol{s} \in \mathcal{S}$ |
| $\mathcal{H},\ \boldsymbol{h}$ | A subset of spaces $\mathcal{H} \subseteq \{\mathcal{F}, \mathcal{E}\}$ with $\boldsymbol{h} \in \mathcal{H}$ |
| $\hat{\boldsymbol{s}}, \hat{\boldsymbol{x}}, \hat{\theta}, \hat{y}, \hat{\boldsymbol{e}}$ | A sample, input, parameters, logit, explanation, post-perturbation. |

**Functions**

| | |
|---|---|
| $f$ | A classifier function $f : \mathcal{X} \to \mathcal{Y}$ with $f(\boldsymbol{x}; \theta) = \boldsymbol{y}$ (we refer $f_\theta$ as $f$), parameterised by $\theta$ |
| $\phi_L$ | A local function $\phi_L : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^V$ with $\phi_L(f, \boldsymbol{x}, y; \lambda) = \boldsymbol{e}$, parameterised by $\lambda$ |
| $\phi_G$ | A global explanation function $\phi_G : \mathcal{F} \times \mathcal{Y} \to \mathbb{R}^V$ with $\phi_G(f, y; \lambda) = \boldsymbol{e}$, parameterised by $\lambda$ |
| $\phi$ | Collectively, denoting $\phi_L$, and $\phi_G$ although they formally reside in different spaces. |
| $\Psi$ | An evaluation function $\Psi : \mathcal{E} \times \mathcal{X} \times \mathcal{F} \times \mathcal{Y} \to \mathbb{R}$ with $\Psi(\boldsymbol{e}, \boldsymbol{x}, f, y; \tau) = \boldsymbol{q}$, parameterised by $\tau$ |
| $\mathcal{P}_\mathcal{S}$ | A perturbation function $\mathcal{P} : \mathcal{S} \to \mathcal{S}$ where $\mathcal{P}(\boldsymbol{s}; \omega)$ on space $\mathcal{S}$ |
| $\delta$ | A general discrepancy function $\delta : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ with $\delta(\boldsymbol{s}, \hat{\boldsymbol{s}}) = \xi$, parameterised by $\omega \in \mathbb{R}$ |
| $k$ | A separate mapping function $k : \mathcal{S} \to \mathcal{H}$ mapping $\boldsymbol{s}, \hat{\boldsymbol{s}}$ to a distinct space $\mathcal{H}$ |
| $\boldsymbol{D}_k$ | A functional distortion $\boldsymbol{D}_k : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ with $\boldsymbol{D}_k(\boldsymbol{s}, \hat{\boldsymbol{s}}) = \delta(k(\boldsymbol{s}), k(\hat{\boldsymbol{s}}))$ |
| $\rho$ | A correlation function with $\rho : \mathbb{R}^Z \times \mathbb{R}^Z \to \mathbb{R}$ |

**Constants**

| | |
|---|---|
| $C$ | The number of classes |
| $D$ | The dimension of the input |
| $W$ | The dimension of the parameter vector |
| $V$ | The dimension of the explanation outputs |
| $Z$ | The number of perturbation steps |
| $K$ | The number of samples to approximate the Jacobian |
| $T$ | The number of integral steps between two points, $\boldsymbol{e}$, and $\hat{\boldsymbol{e}}$ |
| $M$ | The number of models to average over in `GEF`, and `Fast-GEF` |

**Variables**

| | |
|---|---|
| $\xi$ | The perturbation magnitude defined as the discrepancy $\delta(\boldsymbol{s}, \hat{\boldsymbol{s}}) = \xi$ between $\hat{\boldsymbol{s}}$, and $\boldsymbol{s}$ |
| $\boldsymbol{D}_f$ | The model distortion $\boldsymbol{D}_f$ across parameter- $\boldsymbol{D}_f(\theta, \hat{\theta})$, and input perturbation $\boldsymbol{D}_f(\boldsymbol{x}, \hat{\boldsymbol{x}})$ |
| $\boldsymbol{D}_\phi$ | The explanation distortion $\boldsymbol{D}_\phi$ across parameter- $\boldsymbol{D}_\phi(\theta, \hat{\theta})$, and input perturbation $\boldsymbol{D}_\phi(\boldsymbol{x}, \hat{\boldsymbol{x}})$ |
| $\varepsilon^{RO}_{\boldsymbol{D}_k}$ | The implicit upper boundary value with $\varepsilon^{RO} \in \mathbb{R}^+$, and $k \in \{\phi, f\}$ used in robustness |
| $\varepsilon^{SE}_{\boldsymbol{D}_k}$ | The implicit lower boundary value with $\varepsilon^{SE} \in \mathbb{R}^+$, and $k \in \{\phi, f\}$ used in sensitivity |
| $\alpha$ | A boundary value for the perturbation magnitude, with $\alpha \in \mathbb{R}^+$ |
| $\eta_i$ | The Gaussian noise matrix with $\eta_i \sim \mathcal{N}(\boldsymbol{0}, \sigma_i^2 \mathbb{1})$ |
| $\sigma_z^2$ | The covariance scale of a Gaussian distribution with $\sigma_z^2 \in \mathbb{R}^+$ at $\mathrm{z}^{th}$ perturbation |
| $J_f$ | The network Jacobian for fixed input $\boldsymbol{x}$, with $J_f \in \mathbb{R}^{V \times C}$, and elements $J_{i,j} = \frac{\partial e_i}{\partial f_j}$ |
| $\mathbf{g}$ | Pullback metric tensor based on the elementwise Jacobian with $\mathbf{g} \in \mathbb{R}^{V \times V}$ |
| $z$ | Index of perturbation steps with $z \in [1, Z]$ |
| $\boldsymbol{D}_f^z$ | The model distortion at perturbation step $z$ with $\boldsymbol{D}_f^z := \boldsymbol{D}_f^z(\theta, \hat{\theta}_z)$ |
| $\boldsymbol{D}_\phi^z$ | The explanation distortion at perturbation step $z$ with $\boldsymbol{D}_\phi^z := \boldsymbol{D}_f^z(\theta, \hat{\theta}_z)$ |
| $\boldsymbol{d}_f$ | The vector of model distortion with $Z$ steps, $\boldsymbol{d}_f = [\boldsymbol{D}_f^1, \boldsymbol{D}_f^2, \ldots, \boldsymbol{D}_f^Z]$ |
| $\boldsymbol{d}_\phi$ | The vector of explanation distortion with $Z$ steps, $\boldsymbol{d}_\phi = [\boldsymbol{D}_\phi^1, \boldsymbol{D}_\phi^2, \ldots, \boldsymbol{D}_\phi^Z]$ |