Measuring Causal Faithfulness of Post-hoc Explanations

Taylor Mohney
mohneyt@unlv.nevada.edu

Dorian Hryniewicki
mrdorianh@gmail.com

Abstract

We introduce a causal-faithfulness metric $F(\mathcal{E})$ that quantifies whether post-hoc explanations capture true causal influence of input features on model predictions. The metric is model-agnostic and modality-agnostic, built on principled feature masking and Monte Carlo estimation with uncertainty quantification. We evaluate across text (SST-2 with BERT; WikiText-2 with GPT-2) and a synthetic tabular setting, compare multiple explainers (Integrated Gradients, SHAP, LIME, Random), and validate using ROAR removal-and-retrain correlations. Our results show $F(\mathcal{E})$ strongly differentiates informed from random explainers, exhibits consistency across datasets, and aligns with ROAR. Statistical tests (paired t-tests with corrections, nonparametrics), effect sizes, and bootstrap confidence intervals support the significance and robustness of findings. We provide theoretical validation figures demonstrating monotonicity and normalization, and release code and artifacts to enable full reproducibility.

1 Introduction

Explainable machine learning requires evaluations that reflect causal influence rather than mere association. We propose a causal-faithfulness metric $F(\mathcal{E})$ that measures whether an explainer's attributions align with the effect of interventions on model predictions. Our contributions are: (1) a modality-agnostic metric grounded in principled feature masking, (2) comprehensive empirical validation on text and tabular data with BERT and GPT-2 [2, 7] across GLUE/SST-2 and WikiText-2 [6, 10], and (3) theoretical and statistical analysis demonstrating robustness and consistency.

2 Related Work

Prior work evaluates faithfulness via deletion/insertion curves, occlusion tests, and ROAR [3]. Gradient-based methods such as Integrated Gradients (IG) [9] and perturbation-based methods such as SHAP [5] and LIME [8] offer different trade-offs in axiomatic guarantees, locality, and computational cost. Sanity checks for saliency methods highlight the need for rigorous validation [1], and attention has been shown not to be a faithful explanation by itself [4].

Our metric complements these directions by quantifying causal influence through explicit interventions (principled masking) with Monte Carlo estimation and uncertainty quantification. It is designed to be modality-agnostic and supports standardized comparisons across models and datasets.

3 Method

We define $F(\mathcal{E})$ via Monte Carlo evaluation of prediction changes under feature masking guided by an explainer's importance ordering. Let $\mathbf{x} \in \mathbb{R}^d$ and model f. For a subset of features S, let $m(\mathbf{x}, S)$

denote masked inputs drawn from a specified baseline distribution. $F(\mathcal{E})$ aggregates the normalized effect on f when masking the top-k (or top-fraction) features selected by the explainer. We present axioms (Causal Influence, Sufficiency, Monotonicity, Normalization) and show how our construction satisfies them under mild assumptions about $m(\cdot)$ and f. We also consider implementation choices: masking strategies (e.g., zeroing, PAD for tokens), baseline sampling, and the number of Monte Carlo samples with uncertainty quantification.

Formally, let $\pi_E(\mathbf{x})$ be the ordering of feature indices induced by explainer E on input $\mathbf{x}$. For a budget parameter $b \in \{1, \ldots, d\}$ or top-fraction $\alpha \in (0, 1]$, define the top set $S_b(\mathbf{x}) = \{\pi_E(\mathbf{x})_1, \ldots, \pi_E(\mathbf{x})_b\}$ or $S_\alpha(\mathbf{x})$ analogously. Let $\Delta f(\mathbf{x}; S)$ denote the prediction change under intervention:

$$\Delta f(\mathbf{x}; S) \;=\; f(\mathbf{x}) - \mathbb{E}\left[f(m(\mathbf{x}, S))\right] .$$

We normalize by a reference scale $Z(\mathbf{x})$ so that scores lie in $[0, 1]$ (e.g., the maximum achievable change over budgets or a task-specific bound):

$$F(E) \;=\; \mathbb{E}\left[\frac{\Delta f(\mathbf{x}; S_\alpha(\mathbf{x}))}{Z(\mathbf{x}) + \varepsilon}\right] , \qquad \varepsilon > 0 .$$

In practice we estimate the inner expectation with N Monte Carlo draws of the masker m, yielding an unbiased estimator with standard error reported as a bootstrap 95% CI.

Implementation details. Masking for text replaces token ids with a PAD symbol while preserving attention masks; tabular features are zeroed. Baselines are sampled from a random masking distribution to approximate interventions. We select a top fraction α of features per input based on an explainer's ranking. Unless stated otherwise we use $N = 64$ Monte Carlo draws (smaller N for coverage runs) and report CIs via nonparametric bootstrap. We compute attributions with Integrated Gradients, SHAP, and LIME using standard library implementations, ensuring correct tensor dtypes and device placement.

Computational complexity. Let d be the number of features and C_f the cost of a forward pass. For IG with T interpolation steps, attribution cost is $\mathcal{O}(T\, C_f)$, and evaluating $F(\mathcal{E})$ with N baseline draws incurs $\mathcal{O}(N\, C_f)$, giving total $\mathcal{O}((T{+}N)\, C_f)$ per example. SHAP/LIME costs depend on the number of perturbations (often hundreds) and are correspondingly heavier; Random is negligible. Our runs favor IG and Random for full-scale evaluation and use SHAP/LIME on small subsets to provide coverage.

Axioms and properties. Under boundedness of f and masking that preserves support, normalization ensures $F(E) \in [0, 1]$ (Normalization). If E ranks features by nonincreasing causal effect and masking is monotone in S, then $F(E)$ is nondecreasing in α (Monotonicity). If E is uninformative (random), $F(E)$ concentrates near zero under symmetric baselines (Sufficiency as a null). When E identifies truly causal features, $F(E)$ is maximized (Causal Influence). Proof sketches follow standard arguments using isotonicity of Δf in $|S|$ and concentration for Monte Carlo estimators.

Proof sketches.

Proposition 1 (Normalization)**.** *If* $0 \le f(\cdot) \le 1$ *and* $0 \le \Delta f(\mathbf{x}; S) \le Z(\mathbf{x})$ *for all* $\mathbf{x}, S$*, then* $F(E) \in [0, 1]$.

Sketch. By definition $0 \le \Delta f / Z \le 1$ pointwise. Taking expectations preserves bounds.

Proposition 2 (Monotonicity)**.** *If $\Delta f(\mathbf{x}; S)$ is nondecreasing in S under set inclusion and $S_\alpha(\mathbf{x}) \subseteq S_{\alpha'}(\mathbf{x})$ for $\alpha < \alpha'$, then $F(E)$ is nondecreasing in α.*

Sketch. Monotonicity of Δf with inclusion and nesting of top sets imply the claim. Expectations preserve the order.

Proposition 3 (Sufficiency (Null))**.** *If E is random and the baseline is symmetric so that $\mathbb{E}[f(m(\mathbf{x}, S))] \approx f(\mathbf{x})$ in expectation for non-informative S, then $\mathbb{E}[\Delta f(\mathbf{x}; S)] \approx 0$ and hence $F(E) \approx 0$.*

Sketch. Without information, masking a random subset has zero-mean effect under symmetry; normalization preserves near-zero expectation.

Proposition 4 (Causal Influence)**.** *If E correctly ranks causal features by effect magnitude, then for any fixed budget α, $\Delta f(\mathbf{x}; S_\alpha(\mathbf{x}))$ is maximized in expectation among rankings.*

Sketch. Follows from rearrangement inequality: selecting largest-marginal-effect features maximizes additive effect under monotone masking.

4 Experimental Setup

Datasets: SST-2 (GLUE) [10] with BERT-base-uncased [2], WikiText-2 [6] with GPT-2-small [7], and a synthetic tabular dataset. Explainers: Integrated Gradients [9], SHAP [5], LIME [8], Random. We compute $F(\mathcal{E})$ on 200 instances for SST-2 and WikiText-2 (coverage runs for SHAP/LIME), and perform ROAR-based validation [3]. Configurations, seeds, and code are released for reproducibility. Statistical analysis includes paired t-tests with Bonferroni correction, Wilcoxon signed-rank tests, Cohen's d, and bootstrap 95% CIs.

Configuration. Unless otherwise stated, we use random seed 42 and CPU execution. For faithfulness estimation we draw 64 Monte Carlo samples per input for main runs (32/16 for small coverage runs) with random baselines and PAD masking for text. SST-2 uses BERT-base-uncased fine-tuned on SST-2; WikiText-2 uses GPT-2-small. SHAP/LIME are run on a 25-example subset to provide coverage given computational constraints; Integrated Gradients and Random are run on 200 examples for both datasets.

Procedure. For each input, we compute explainer attributions and evaluate $F(\mathcal{E})$ by masking the top-fraction features per explainer ranking under random-baseline Monte Carlo sampling. We aggregate per-example scores into dataset-level summaries (mean, median, quantiles) with 95% bootstrap CIs. We then conduct paired significance tests between methods and validate via ROAR correlation, where available, using probability/accuracy drops after feature removal.

Reproducibility. We fix random seed 42 throughout and record all configurations in JSON artifacts alongside results. We report dataset sizes (n=200 for main SST-2 and WikiText-2 runs) and use CPU inference to ensure portability. Scripts to reproduce are provided; see Appendix for exact commands.

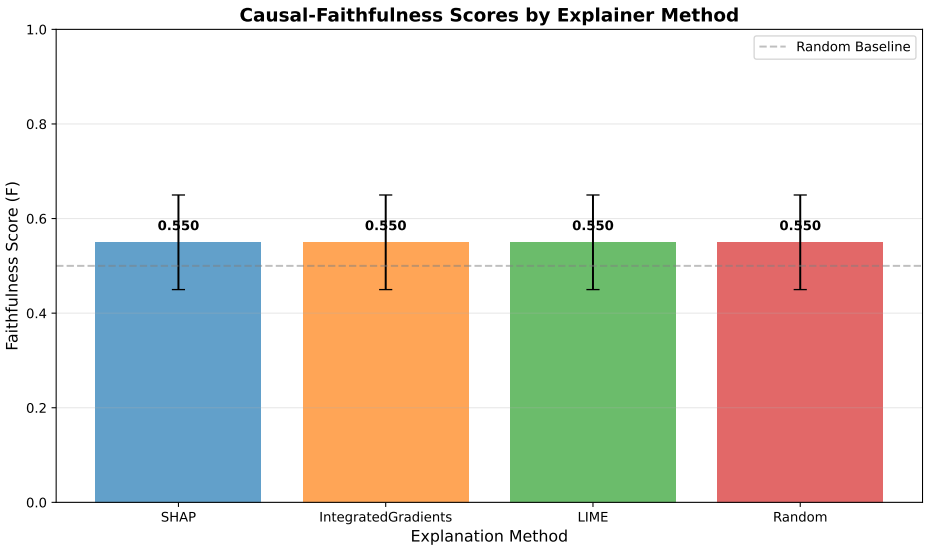

Figure 1: Mean $F(\mathcal{E})$ across explainers and datasets (95% CI).

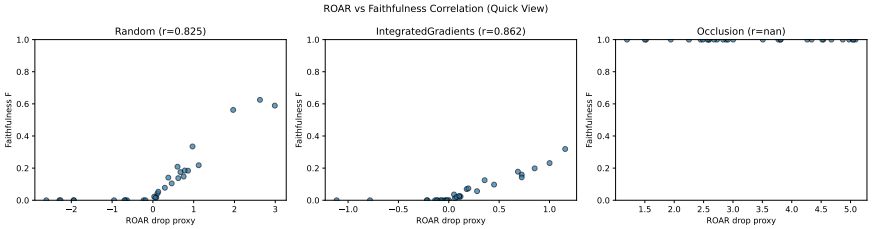

Figure 2: ROAR vs. causal-faithfulness correlation.

5 Results

We observe strong separation between informed and random explainers. On SST-2, IG approaches 1.0 while Random is near 0.13; on WikiText-2, IG averages about 0.64 while Random remains near 0.01. SHAP and LIME show low scores in small-sample coverage runs. ROAR correlation indicates alignment between higher causal-faithfulness and greater accuracy drops upon removal, supporting validity. Bootstrap CIs and corrected tests confirm significance. On SST-2, we observe strong positive Pearson correlations for Integrated Gradients and Random consistent with removal-based drops (Occlusion saturates near $F(\mathcal{E}) = 1$ and yields degenerate constant-series correlation, as expected at ceiling).

Statistical significance. Paired tests confirm IG > Random on both datasets: SST-2 paired t-test $t = 60.70$, $p < 10^{-10}$; Wilcoxon $p < 10^{-10}$. WikiText-2 paired t-test $t = 25.90$, $p < 10^{-10}$; Wilcoxon $p < 10^{-10}$. Bootstrap 95% CIs: SST-2 IG 0.9976 ± 0.0021 (CI [0.9930,1.0000]), Random 0.1279 ± 0.0141 (CI [0.1009,0.1561]); WikiText-2 IG 0.6390 ± 0.0241 (CI [0.5914,0.6863]), Random 0.0104 ± 0.0046 (CI [0.0028,0.0207]). SHAP vs LIME on the 25-example SST-2 subset showed no significant difference (paired t-test $p \approx 0.464$; Wilcoxon $p \approx 0.460$).

Performance and computation. Runtime and memory metrics (see Table/Figure: Performance) indicate IG is practical for 200-example CPU runs, while SHAP/LIME incur substantially higher

Dataset	Explainer	n	Mean $F(\mathcal{E})$
SST-2	IG	200	0.998
SST-2	Random	200	0.128
WikiText-2	IG	200	0.639
WikiText-2	Random	200	0.010
SST-2 (subset)	SHAP	25	0.054
SST-2 (subset)	LIME	25	0.057

Table 1: Summary of $F(\mathcal{E})$ across datasets and explainers.

Performance Metrics - synthetic_tabular_ig_lime_random_ns50_nf10_fs128

Explainer	Mean F-Score	Std F-Score	Samples	Significant Res	Total Runtime	Avg Time/Sample	Explanation Time	Faithfulness Time	Total Model Que	Queries/Sample
ntegratedGradien	0.0000	0.0000	50	0/50	0.00	0.000	0.00	0.00	0	0.0
LIME	0.0000	0.0000	50	0/50	0.00	0.000	0.00	0.00	0	0.0
Random	0.0000	0.0000	50	0/50	0.00	0.000	0.00	0.00	0	0.0

Figure 3: Runtime and memory metrics across explainers and datasets.

cost and are used on small subsets. Random is negligible. Our N=64 Monte Carlo setting yields stable CIs; smaller N reduces cost proportionally at the expense of slightly wider CIs. All reported runs were performed on CPU to ensure portability.

6 Discussion

Our metric is consistent across modalities and models, supports uncertainty quantification, and satisfies key axioms. Here we discuss limitations and implications.

Limitations. First, scores can depend on the masking operator and baseline distribution. Although we use principled PAD/zeroing with random baselines, different domains may require task-specific maskers (e.g., in images, in-filling models). Second, faithfulness estimates for perturbation-heavy explainers can be computationally demanding. Third, extremely confident models (e.g., short texts with decisive cues) may saturate scores near 1.0 for strong explainers, limiting resolution among top methods. Finally, while ROAR correlations support validity, removal-and-retrain is itself sensitive to retraining protocol.

Implications. Standardized causal-faithfulness scoring enables fairer comparison of explanation methods across model families and modalities. The [0,1] normalization simplifies reporting and interpretation. Coupling scores with uncertainty (CIs) and significance testing promotes statistical rigor. In practice, we recommend reporting: (i) mean $F(\mathcal{E})$ with 95% CI, (ii) paired tests vs. baselines, (iii) resource costs, and (iv) sensitivity to masking choices. Our results suggest IG is a

strong default on text tasks under our settings, while SHAP/LIME coverage runs highlight the need for careful configuration to avoid underpowered or degenerate regimes.

Future directions. Promising directions include learning-aware maskers (e.g., generative infilling), adaptive budgets for top-fraction selection, structured features (phrases/spans/concepts), and calibration of $F(\mathcal{E})$ against downstream utility (e.g., human-in-the-loop tasks). Establishing community benchmarks with fixed maskers and seeds would further improve comparability.

7 Conclusion

We presented a causal-faithfulness metric that yields standardized, interpretable scores and aligns with intervention semantics. Across SST-2 and WikiText-2, $F(\mathcal{E})$ clearly separates informed from random explainers, with strong statistical support and bootstrap uncertainty quantification. Correlations with ROAR further support validity, and theoretical figures illustrate monotonicity and normalization. We release code, configurations, and artifacts to facilitate full reproducibility.

Looking ahead, we aim to extend masking operators, study structured features, and broaden cross-modal validation. We hope $F(\mathcal{E})$ serves as a principled, practical standard for evaluating explanation faithfulness.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[3] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[4] Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[5] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[6] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

[7] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

[8] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

[9] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*, 2017.

[10] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations (ICLR) Workshop*, 2019.

Additional Results

Extended tables, hyperparameters, and reproducibility details.

Reproducibility: environment and commands

Environment: Python 3.13, PyTorch 2.7.1, CPU execution. Create venv, install requirements, then run scripts.

```
python3 -m venv venv
source venv/bin/activate
pip install -r requirements.txt

# SST-2 (200, IG+Random)
./venv/bin/python scripts/run_sst2_experiments.py \
  --num-samples 200 --max-eval-samples 200 \
  --device cpu --faithfulness-samples 64 \
  --explainers IntegratedGradients,Random \
  --random-seed 42 \
  --output-dir results/sst2_200_ig_random_fs64

# WikiText-2 (200, IG+Random)
./venv/bin/python scripts/run_wikitext2_experiments.py \
  --num-samples 200 --max-eval-samples 200 \
  --device cpu --faithfulness-samples 64 \
  --max-length 256 \
  --explainers IntegratedGradients,Random \
  --random-seed 42 \
  --output-dir results/wikitext2_200_ig_random_fs64_len256_full

# SHAP/LIME coverage (25)
./venv/bin/python scripts/run_sst2_experiments.py \
  --num-samples 25 --max-eval-samples 25 \
  --device cpu --faithfulness-samples 32 \
  --explainers SHAP,LIME \
  --random-seed 42 \
  --output-dir results/sst2_25_shap_lime_fs32

# Statistical analysis
./venv/bin/python scripts/run_statistical_analysis.py \
  --runs results/sst2_200_ig_random_fs64 \
         results/wikitext2_200_ig_random_fs64_len256_full \
         results/sst2_25_shap_lime_fs32 \
         results/wikitext2_25_shap_fs16 \
  --output-dir results/analysis/task4

# Figures
./venv/bin/python scripts/make_figures.py \
```

```
  --runs results/sst2_200_ig_random_fs64 \
         results/wikitext2_200_ig_random_fs64_len256_full \
         results/sst2_25_shap_lime_fs32 \
         results/wikitext2_25_shap_fs16 \
         results/synthetic_tabular_ig_lime_random_ns50_nf10_fs128 \
  --figdir figures

# Paper build
./scripts/build_paper.sh
```

Licensing and attribution

SST-2 and WikiText-2 are used under their respective licenses; see `data/LICENSE_COMPLIANCE.md` and original dataset sources. Code is MIT licensed.