Evaluation Cards for XAI Metrics

Anonymous ACL submission

Abstract

The evaluation of explainable AI (XAI) methods is affected by a lack of standardization. Metrics are inconsistently defined, incompletely reported, and rarely validated against common baselines. In this paper, we identify transparency of evaluation reporting as a central, under-addressed problem. We propose the XAI Evaluation Card, a documentation template analogous to model cards, designed to accompany any study that introduces an XAI evaluation metric. The card covers explicit declaration of target properties, grounding levels, metric assumptions, validation evidence, gaming risks, and known failure cases. We argue that adopting this template as a community norm would reduce evaluation fragmentation, support meta-analysis, and improve accountability in XAI research.

1 Introduction

As AI systems are deployed in high-stakes settings, explainable AI (XAI) has become essential for accountability and transparency. However, the difficult part is not just explaining model predictions, but rigorously evaluating those explanations. A growing body of literature has attempted to systematize XAI evaluation metrics, but the field remains fragmented. Metrics are introduced without clear mappings to the properties they measure, and reported without contextual information needed for reproducibility.

This paper focuses on standardizing documentation for XAI evaluation through the XAI Evaluation Card. Drawing on a meta-review of surveys published between 2021 and 2025, we argue that the absence of structured reporting is a primary addressable cause of many of the field's evaluation problems, and that a card-based template, analogous to model cards (Mitchell et al., 2019) and datasheets (Gebru et al., 2021), can address them directly.

Our contribution is a concrete template with four structured sections covering identity, scope, validation, and relationships. We discuss how the template maps onto recurring gaps identified across the surveyed literature, and how its adoption as a community norm could improve reproducibility and comparability in XAI evaluation.

2 Background and Motivation

Across eleven surveys covering XAI evaluation metrics published between 2021 and 2025, three main problems appear. First, the field lacks terminological consensus. The same metric names appear under different definitions across papers, and the same underlying properties can be named differently by different research groups. Second, evaluation practice is heavily skewed toward functionally-grounded proxy tasks while human-grounded and application-grounded methods remain underrepresented despite being more directly informative about real-world utility. Third, implementation availability is inconsistently reported, with some proposed metrics remaining at the level of theoretical definitions with no accompanying code.

2.1 Meta-review

Terminological fragmentation is documented most extensively by Pawlicki et al. (2024), who identify close to a hundred XAI metrics and note that several distinct papers use identical metric names for different operationalizations, while papers measuring the same underlying property use different terminology for it. Kadir et al. (2023) observe the resulting difficulty in benchmarking and comparison across XAI evaluation methods. Nauta et al. (2023) also note standardization-related disagreements, and attempt a systematic response, reviewing 29 quantitative evaluation metrics across 12 evaluation properties and arguing that explainability must be treated as a multi-faceted concept eval-

uated along multiple dimensions simultaneously.

The skew toward functionally-grounded evaluation is noted by Lopes et al. (2022), whose taxonomy distinguishes human-centered from computer-centered evaluations, and by Mangold et al. (2025), who argue that current evaluation processes are often too technical and insufficiently focused on human users. Mohseni et al. (2021) find that human-AI task performance dominates the evaluation literature, while mental model alignment and user trust are rarely operationalized. Sisk et al. (2022) similarly highlight a lack of consensus on how explanation reliability and validity should be assessed from a human perspective.

Implementation gaps are highlighted by Coroama and Groza (2022), who check reviewed papers for the presence or absence of an accompanying implementation, finding that many proposed metrics remain theoretical constructs without practical instantiation. Banerjee and Barnwal (2022) similarly note that quantitative metrics depend heavily on the type of machine learning problem and model used, which compounds the difficulty of providing general-purpose implementations. Zhou et al. (2021) find that quantitative metrics for model-based and example-based explanations largely measure simplicity, while metrics for attribution-based explanations target soundness. The most recent survey, Dembinsky et al. (2025), attempts to unify the field through the VXAI framework, consolidating 362 publications into 41 functionally-similar metric groups and proposing a three-dimensional categorization scheme. However, this framework focuses on taxonomy rather than documentation standards, and does not address the reporting gaps we identify.

2.2 The Evaluation Problem in XAI

A fundamental distinction in XAI is between properties (conceptual qualities such as fidelity, robustness, or clarity) and metrics, which operationalize these properties into measurable quantities. This distinction is important because the mapping from properties to metrics is neither unique nor exact. Multiple metrics may target the same property while giving different conclusions, and a metric's validity is often context-dependent, varying with model architecture, data modality, and explanation scope.

Across eleven reviewed surveys, five recurring evaluation problems were identified: (1) metrics are introduced without declaring which properties they target; (2) results are reported without specifying the evaluation context in which they are valid; (3) few metrics include sensitivity or stability analysis as part of their original proposal; (4) metric disagreements are rarely acknowledged and interpreted; and (5) implementation availability is inconsistent, with many metrics remaining at the level of theoretical definitions.

These problems are not independent. The absence of declared target properties makes disagreement handling difficult, and the lack of contextual reporting prevents meaningful replication. The XAI Evaluation Card is designed to address all five, as detailed in Section 3.

2.3 Related Documentation Frameworks

The idea of structured documentation for AI artifacts is already established in other areas. Model cards (Mitchell et al., 2019) standardize reporting of model performance. Datasheets for Datasets (Gebru et al., 2021) document dataset origin, collection procedures, and intended uses. To our knowledge, no equivalent documentation standard exists specifically for XAI evaluation metrics, despite the fact that, currently, these metric results are used as evidence for claims about explanation quality.

3 The XAI Evaluation Card

We propose the XAI Evaluation Card as a standardized supplement to any study that proposes a new XAI evaluation metric. The card is organized into four sections, each addressing a different class of documentation gap identified in the meta-review. A filled out example can be found in Appendix A.

3.1 Identity

The identity section requires authors to name the metric, list all explainability properties it operationalizes (with explicit references to property definitions), and declare its grounding level following the framework of Doshi-Velez and Kim (2017): functionally-grounded (proxy tasks), human-grounded (user studies), or application-grounded (domain experts in deployment settings). This directly addresses the widespread conflation of properties and metrics, and the lack of declared evaluation grounding observed across the surveyed literature. Explicit grounding declarations prevent a common failure mode: drawing human-centered conclusions from purely technical metrics.

XAI Evaluation Card	
I. Identity	
Metric Name	Unique, descriptive name for the evaluation metric.
Target Property / Properties	List all explainability properties this metric operationalizes (e.g., *fidelity*, *robustness*, *clarity*), with references to definitions used.
Grounding Level	One or more of: *functionally-grounded* / *human-grounded* / *application-grounded* (Doshi-Velez and Kim, 2017).
II. Scope and Context	
Evaluation Context	Model architecture, data modality, and explanation scope (*local* / *global*) under which results are reported.
Assumptions	All assumptions required by the metric (e.g., feature independence, locality, linearity, calibrated probabilities, meaningful baselines).
III. Implementation and Validation	
Implementation Available?	Yes / No. If yes, provide URL or repository reference.
Validation Evidence	Summary of sensitivity analysis, stability analysis, and correlation with related metrics. Report computational cost where relevant.
Gaming Risk	How a method could achieve a high score on this metric without improving the target explainability property.
Known Failure Cases	Conditions under which the metric is known to fail or produce misleading results.
IV. Relationships and Limitations	
Relationship to Other Metrics	Metrics targeting the same property. Known agreements or disagreements in results.
Disagreement Handling	If this metric conflicts with others reported, state which property is prioritised for the target deployment scenario and why.
Limitations	Main limitations as an operationalization of the target property. Note contexts where the metric should not be used.

Table 1: XAI Evaluation Card template. Fields marked N/A require a brief justification.

3.2 Scope and Context

The scope section requires a description of the evaluation context (model architecture, data modality, and whether evaluation is local or global) and an explicit enumeration of all assumptions made by the metric, such as feature independence, locality, or the availability of meaningful baselines. Coroama and Groza (2022) highlight that metric validity is highly context-dependent, yet contextual information is routinely omitted from publications. Without it, reported metric scores are not interpretable and cannot be meaningfully compared across studies. This section operationalizes the principle that metric values should never be reported in isolation.

3.3 Implementation and Validation

The validation section asks authors to (a) indicate whether an implementation is available and provide a link, (b) summarize validation evidence including sensitivity analysis, stability analysis, and correlation with related metrics, (c) describe gaming risk (how a method could score highly on this metric without actually improving the target property) and (d) document known failure cases. These requirements respond directly to findings that some proposed metrics lack implementations and that validation beyond the original development context is rare. The gaming risk field is particularly important as it highlights a class of validity threat that is rarely made explicit in evaluation papers.

3.4 Relationships and Limitations

The final section places the metric in its broader evaluation ecosystem. Authors should list metrics targeting the same property and note any known agreements or disagreements. When metrics diverge, they should state which property is prioritized and why, for the target deployment scenario. A limitations field requires listing conditions under which the metric should not be used. Pawlicki et al. (2024) find a large diversity of metrics without consensus on their properties, making inter-metric relationships a critical missing element of most publications. This section makes those relationships explicit and searchable, directly supporting meta-analysis.

4 Discussion

4.1 Relevance to Evaluation Practice

Each field in the XAI Evaluation Card maps onto a documented failure mode in the literature. Cards can be completed as a supplementary table or appendix, imposing minimal overhead while creating a structured, machine-readable record of evaluation decisions. Requiring cards as part of peer review would shift evaluation norms toward greater rigor.

The card is explicitly non-prescriptive. It does not mandate any particular grounding level, metric, or validation procedure. Fields that are not applicable may be marked N/A with justification. This design choice reflects the diversity of XAI evaluation contexts, from post-hoc attribution methods on tabular data to global explanation of deep vision models, and avoids imposing a single evaluation paradigm.

4.2 Connections to Model Cards and Datasheets

The XAI Evaluation Card is intentionally analogous to model cards and datasheets, which have achieved significant community adoption. The card differs from these in its focus: where model cards document what a model does and for whom, evaluation cards document how the quality of an explanation is being measured and under what conditions that measurement is valid. This distinction matters because metrics can be viewed as methodological claims, not empirical systems, and their documentation needs to reflect that difference.

5 Conclusion

XAI evaluation is a critical yet fragmented field. Motivated by a meta-review of eleven recent surveys, we have identified transparency of evaluation reporting as a central, under-addressed challenge. We propose the XAI Evaluation Card, a four-section structured documentation template covering metric identity, evaluation scope and context, implementation and validation, and inter-metric relationships. We argue that adopting this template as a community norm would reduce evaluation fragmentation, enable meaningful comparison across studies, and improve the accountability of empirical claims about explanation quality.

6 Limitations

The card template is a minimal viable standard. It does not resolve deeper disagreements about what properties XAI explanations should satisfy, nor does it provide a formal ontology or taxonomy for aligning differently-named concepts across studies (e.g., whether "faithfulness" in one paper corresponds to "fidelity" in another). Future work could extend the card with formal property references and machine-readable schemas to support automated meta-analysis. Additionally, adoption depends on community and venue buy-in. Mandating cards at the reviewer level could be an effective enforcement mechanism.

References

Puja Banerjee and Rajesh P Barnwal. 2022. Methods and metrics for explaining artificial intelligence models: A review. *Explainable AI: Foundations, methodologies and applications*, pages 61–88.

Loredana Coroama and Adrian Groza. 2022. Evaluation metrics in explainable artificial intelligence (XAI). In *International conference on advanced research in technologies, information, innovation and sustainability*, pages 401–413. Springer.

David Dembinsky, Adriano Lucieri, Stanislav Frolov, Hiba Najjar, Ko Watanabe, and Andreas Dengel. 2025. Unifying VXAI: A Systematic Review and Framework for the Evaluation of Explainable AI. *arXiv preprint arXiv:2506.15408*.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Md Abdul Kadir, Amir Mosavi, and Daniel Sonntag. 2023. Evaluation metrics for XAI: A review, taxonomy, and practical applications. In *2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES)*, pages 000111–000124. IEEE.

Pedro Lopes, Eduardo Silva, Cristiana Braga, Tiago Oliveira, and Luís Rosado. 2022. XAI systems evaluation: a review of human and computer-centred methods. *Applied Sciences*, 12(19):9423.

Aline Mangold, Juliane Zietz, Susanne Weinhold, and Sebastian Pannasch. 2025. On the Design and Evaluation of Human-centered Explainable AI Systems: A Systematic Review and Taxonomy. *arXiv preprint arXiv:2510.12201*.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In

Proceedings of the conference on fairness, accountability, and transparency, pages 220–229.

Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–45.

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

Marek Pawlicki, Aleksandra Pawlicka, Federica Uccello, Sebastian Szelest, Salvatore D'Antonio, Rafał Kozik, and Michał Choraś. 2024. Evaluating the necessity of the multiple metrics for assessing explainable AI: A critical examination. *Neurocomputing*, 602:128282.

Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*.

Marion Sisk, Makeen Majlis, Cameron Page, and Abbas Yazdinejad. 2022. Analyzing XAI metrics: Summary of the literature review. *Authorea Preprints*.

Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593.

A Example of a Filled Out XAI Evaluation Card

XAI Evaluation Card	
I. Identity	
Metric Name	Deletion Area Under the Curve (Deletion AUC / DAUC) (Petsiuk et al., 2018)
Target Property / Properties	Faithfulness / Fidelity. It operationalizes this by measuring if the features identified as "important" by the explanation are necessary for the model to maintain its predictive confidence.
Grounding Level	Functionally-grounded (proxy task).
II. Scope and Context	
Evaluation Context	Applied to local feature attributions (e.g., saliency maps) across both vision and NLP modalities. Requires a model with probability or logit outputs.
Assumptions	Assumes feature independence to some degree. Assumes that iteratively masking top-rated features will degrade model performance if the explanation is faithful. Assumes the chosen baseline/imputation method (e.g., replacing pixels with zeros, mean values, or blurring) is meaningful and does not artificially break the model.
III. Implementation and Validation	
Implementation Available?	Yes. https://github.com/eclique/RISE/blob/master/evaluation.py.
Validation Evidence	Empirical studies show DAUC is highly sensitive to the choice of the baseline/imputation value (e.g., zero-masking vs. generative inpainting). It carries a moderate-to-high computational cost, requiring multiple forward passes per instance as features are incrementally removed.
Gaming Risk	An explanation method could achieve a high DAUC score by intentionally selecting features that, when masked, create severe out-of-distribution (OOD) artifacts. The model's confidence drops because the input looks unnatural (like adversarial noise), not because the true explanatory features were removed.
Known Failure Cases	Can produce misleading results when features are highly correlated. The model might rely on a redundant, unmasked feature, making the DAUC score artificially low despite a good explanation.
IV. Relationships and Limitations	
Relationship to Other Metrics	Conceptually similar to comprehensiveness (used in NLP). Often paired with the Insertion AUC metric. May disagree with Faithfulness Correlation if the model's response to feature removal is highly non-linear.
Disagreement Handling	If DAUC conflicts with Insertion AUC, DAUC should be prioritized if the deployment scenario strictly requires identifying the features that are necessary for the model to work (e.g., safety auditing for failure modes).
Limitations	The main limitation is the OOD problem. The metric might evaluate the model's robustness to missing data rather than the explanation's true fidelity. Should not be used in isolation without an insertion or OOD-compensated baseline.

Table 2: Example of a filled out XAI Evaluation Card for Deletion Area Under the Curve (Deletion AUC / DAUC), a popular metric used to evaluate feature attribution methods.