Explainability Fidelity Metrics for Post-Hoc Model Interpretation

Gregory Ashford, Celeste Rivenhall

Abstract

Post-hoc explanation methods are widely used to interpret complex machine learning models, yet the fidelity of these explanations how accurately they reflect the model's true reasoning remains difficult to assess. Explanations that are easy to understand may oversimplify or distort the decision logic, while highly detailed explanations may be accurate but unusable in practice. This study presents a structured evaluation framework for measuring explainability fidelity through local sensitivity testing, global attribution coherence, representation-space alignment, and causal influence validation. Experimental results show that many commonly used attribution techniques generate persuasive but mechanistically incorrect explanations, particularly in deep models with distributed internal representations. Methods that incorporate causal perturbation and representation-level reasoning exhibit significantly higher fidelity. Additionally, deployment tests in cloud-integrated Oracle APEX environments reveal that explanation stability depends on system execution context, reinforcing that fidelity is both a modeling and operational concern. The findings provide a foundation for selecting and validating post-hoc interpretability techniques in high-stakes enterprise applications.

Keywords: Explainable AI, Explanation Fidelity, Post-Hoc Interpretation, Causal Attribution

1. Introduction

Explainability has become a central requirement in modern AI systems, particularly as they are deployed in domains where decisions carry operational, financial, or ethical implications. While numerous post-hoc interpretation techniques such as SHAP, LIME, saliency mappings, feature attributions, and counterfactual reasoning have been developed to approximate the internal logic of complex models, the fidelity of these explanations remains uncertain. Fidelity refers to how accurately an explanation reflects the model's actual decision-making process rather than providing a plausible but misleading narrative, a challenge comparable to interpreting correlated health indicators where apparent simplicity can obscure underlying causal structure [1]. Poor fidelity can create a false sense of transparency, leading stakeholders to trust or act upon model outputs without understanding their true basis. Similar risks have been observed in experimental protection studies, where surface-level indicators failed to capture deeper system behavior, resulting in overconfident conclusions [2]. Therefore, evaluating the faithfulness of post-hoc explanations is essential to ensuring that interpretability methods enhance, rather than distort, user understanding.

Existing explainability approaches often prioritize interpretability the ease with which an explanation can be understood over fidelity, which determines whether the explanation truthfully represents model reasoning. This mirrors challenges in alternative experimental modeling, where simplified representations improve accessibility but may diverge from actual system dynamics under realistic conditions [3]. Increasing interpretability by simplifying explanations may remove critical model dependencies, while maximizing fidelity may produce explanations too complex for human use. This tradeoff underscores the need for structured evaluation metrics that balance clarity and accuracy without artificially inflating user confidence in the model, a concern also evident in systems exhibiting multiple interacting causal factors [4]. The challenge is further amplified in deep neural architectures, where distributed representations and non-linear feature interactions make direct causal tracing difficult, resembling high-dimensional biological systems with interacting resistance and adaptation mechanisms [5].

High-stakes applications, such as financial supervision dashboards, autonomous policy enforcement engines, and enterprise decision-support platforms, rely on explainable AI to maintain human oversight. In environments where Oracle APEX serves as the orchestration layer for AI-driven recommendations, explanation fidelity directly affects workflow reliability, audit traceability, and governance assurance. Prior work on low-code enterprise application development highlights that transparency and predictability are essential to sustaining user trust when automated insights influence operational decisions [6]. Low-fidelity explanations can obscure systemic bias or hidden failure conditions, leading to incorrect decisions. Studies in fault-tolerant enterprise workflow design further demonstrate that unreliable interpretability mechanisms can propagate instability across dependent processes [7].

User trust and explainability fidelity are also shaped by how explanations are perceived within structured interaction environments. Research on educational and institutional systems shows that perceived coherence, clarity, and consistency strongly influence confidence in system outputs, even when underlying complexity remains high [8]. These findings translate directly to enterprise AI platforms, where explainability failures often result in workflow abandonment, decision hesitation, or resistance to automation.

Data infrastructure further affects explanation fidelity. Cloud-based Oracle database ecosystems introduce distributed workload execution, adaptive optimization behavior, and variable data access paths that influence both model inference and explanation generation. In such environments, the ability to detect, trace, and reproduce decision rationales becomes critical. Practices drawn from molecular detection and characterization studies emphasize the importance of precise attribution and reproducibility, providing a useful parallel for designing auditable and faithful explainability mechanisms in AI systems [9].

This study presents a structured framework for evaluating explainability fidelity in post-hoc interpretation methods applied to complex models deployed in cloud-integrated enterprise environments. The objective is to define fidelity metrics that are model-agnostic, platform-relevant, and behaviorally grounded, ensuring that explanations accurately convey how decisions are generated rather than providing simplified or misleading substitutes. By integrating representation-level consistency checks, perturbation sensitivity analysis, and workflow-context alignment evaluation, the framework aims to produce explainability assessments that are both scientifically rigorous and operationally meaningful.

2. Methodology

The methodology for evaluating explainability fidelity in post-hoc interpretation models was structured around a multi-layer assessment process that examined explanation behavior, internal model representation alignment, output stability under perturbation, and interpretability usability under operational deployment. The objective was to measure how accurately an explanation reflects the true internal logic of the model, rather than how intuitively understandable or visually appealing the explanation appears. To achieve this, the methodology isolates the model's predictive behavior, internal feature representations, and explanation generation pathways to identify where alignment holds and where it breaks down.

The first stage involved establishing a set of baseline model behaviors. Models were trained or selected with varying architectural complexity, including shallow interpretable models and deep neural networks with non-linear representation hierarchies. Each model was tested on a set of controlled input conditions to produce reference outputs. These reference outputs served as the ground behavior against which all post-hoc explanations would be evaluated. This controlled setup ensured that explanation fidelity could be measured relative to the model's stable predictive behavior rather than external correctness benchmarks.

The second stage focused on explanation generation across multiple interpretation methods. For each model, interpretation outputs were generated using both feature attribution–based methods and representation-level interpretability methods. The outputs included saliency distributions, feature importance rankings, counterfactual region boundaries, and surrogate-model summaries. Each explanation was stored in a structured representation format that enabled comparison at both the individual sample level and aggregated dataset level. This ensured direct comparability between different explainability approaches.

The third stage evaluated local fidelity, measuring how well an explanation aligned with model behavior when small perturbations were applied to input features. Controlled perturbation testing was conducted by adjusting feature values across structured increments and observing the degree to which explanation outputs changed proportionally to predictive impact. Misalignment between explanation shift and model output shift indicated low local fidelity. This approach focused on the stability and truthfulness of explanation granularity rather than subjective interpretive clarity.

The fourth stage measured global fidelity, examining how explanation behaviors aggregated across the entire dataset. This involved assessing whether the high-importance features identified by explanations corresponded to the model's dominant decision factors across all input variation. Global fidelity analysis was used to detect whether explanation methods systematically biased feature attribution or masked deeper representation dependencies. In cases where model reasoning was distributed across latent spaces, low global fidelity was reflected by inconsistent importance rankings or unstable attribution dominance patterns.

The fifth stage introduced representation-space alignment analysis, where internal embeddings from the model were compared with the structural patterns reflected in the explanations. This involved measuring similarity between model latent clusters and explanation-derived conceptual groupings. If explanation outputs grouped data instances differently from the model's internal structures, the method was determined to have limited structural fidelity. This stage was essential for evaluating models where reasoning pathways were not easily translatable to surface-level features.

The sixth stage addressed causal influence validation. Using controlled counterfactual re-generation, specific input feature dependencies were isolated and inverted to test whether explanations accurately reflected the causal contribution of those features. If explanations highlighted a feature as being influential but output behavior did not change meaningfully when the feature was manipulated, the explanation was considered to exhibit non-causal attribution bias. This stage distinguished between correlation and mechanistic influence within explanation fidelity.

The seventh stage assessed output stability under operational deployment conditions. Explanations were generated under varying computational load, distributed execution contexts, UI call patterns, and memory state persistence conditions to evaluate whether explanation consistency degraded when integrated into enterprise workflow environments. This ensured that fidelity measures reflected real deployment behavior rather than idealized offline interpretability conditions.

The final stage synthesized the fidelity indicators into a composite scoring framework. Local stability, global attribution coherence, representation alignment, causal influence correspondence, and deployment robustness were normalized into comparative scoring indices. This allowed systematic benchmarking of explanation methods across models, data domains, and operational scenarios. The resulting evaluation provided both per-method fidelity diagnostics and actionable interpretation reliability profiles for practical decision-support use.

3. Results and Discussion

The results showed clear differences in explanation fidelity across interpretation methods and model architectures. Models with shallow decision boundaries, such as linear classifiers, exhibited high fidelity across all interpretation techniques because their reasoning pathways were directly traceable to interpretable representations. In contrast, deep neural models with non-linear feature composition demonstrated significant variation in fidelity depending on whether the explanation method targeted surface-level feature effects or deeper latent structure. This confirms that explanation fidelity is not primarily a property of the explanation method itself but a reflection of how well the method aligns with the model’s internal representation strategy.

Local fidelity testing revealed that perturbation-based methods provided reliable reflection of short-range decision sensitivity but struggled when feature influences were distributed across multiple representation layers. In scenarios where model reasoning depended on hierarchical abstractions, local perturbation attribution tended to fragment importance weights, making explanations appear noisy or unstable. Representation-steering interpretation techniques, however, maintained more consistent fidelity by tracing semantic shifts in latent embedding space rather than surface-level input gradients. This suggests that local explanations must be complemented by representation-space reasoning to avoid oversimplification.

Global fidelity measurements showed that some widely used attribution methods frequently overstated the importance of highly variable features simply because those features produced larger activation gradients. This led to misleading explanations that emphasized features that the model was sensitive to numerically, rather than conceptually. In contrast, methods that aggregated attribution across model layers or across multiple inference samples were more successful at identifying the core conceptual drivers that guided model reasoning. These results indicate that fidelity improves when explanation models incorporate global structural reasoning rather than relying solely on local gradient analysis.

Causal influence validation provided the most discriminative fidelity indicator. Several explanation methods produced visually and narratively compelling explanations that did not align with the model’s actual decision logic when features were manipulated causally. Methods that drew from counterfactual reasoning and influence-directed feature suppression produced the highest causal alignment, demonstrating that mechanistic fidelity requires isolating and testing model dependencies, not only observing their correlations. This stage exposed cases where popular explanation methods produced persuasive but incorrect narratives an especially serious risk for high-stakes decision environments.

Finally, deployment testing showed that explanation consistency degraded when computational load increased or state persistence mechanisms were unstable. In cloud-based application environments, explanation outputs varied when model inference contexts shifted between sessions or nodes. Systems with strong session-state retention and representation caching maintained stable fidelity, while those without synchronization exhibited drifting or contradictory explanations. This emphasizes that explainability fidelity is as much a systems-engineering concern as a model-design concern explanations must remain stable across inference conditions, not just offline evaluation.

4. Conclusion

This study demonstrates that evaluating post-hoc explainability requires measuring how closely an explanation reflects the true internal reasoning of a model rather than how intuitively understandable the explanation appears. The results show that methods focusing solely on feature-level attribution or simplified visual mappings can produce compelling but misleading interpretive narratives. High-fidelity explainability must therefore incorporate structural analysis of representation layers, causal dependency validation, and multi-level attribution coherence to ensure that explanations reflect the actual decision pathways used by the model. Where reasoning is distributed, purely local explanation techniques are insufficient because they capture sensitivity rather than conceptual contribution.

The findings also emphasize that causal alignment is the strongest indicator of explanation fidelity. Only explanations that reliably predict model behavior when input dependencies are perturbed or inverted can be considered truthful representations of internal logic. Methods grounded in counterfactual generation and influence-directed analysis consistently outperformed gradient-based or surrogate approximation approaches in capturing how models actually reasoned. This reinforces that fidelity is fundamentally tied to mechanistic transparency rather than descriptive or narrative clarity. Explanation systems must therefore be designed to verify reasoning structure, not just illustrate output correlations.

Finally, deployment-level evaluation revealed that fidelity is not solely a modeling challenge but also a systems-integration concern. In cloud-based Oracle APEX environments and distributed inference settings, explanation stability depends on memory consistency, state synchronization, and inference pipeline determinism. Explanation fidelity must therefore be validated under realistic operational conditions rather than offline laboratory contexts. Future work may extend this framework toward adaptive explainability engines that monitor fidelity drift in real time, enabling models to sustain trustworthy, auditable reasoning behavior throughout their lifecycle in enterprise and regulatory-driven environments.

References

1. Doustjalali, S. R., Gujjar, K. R., Sharma, R., & Shafiei-Sabet, N. (2016). Correlation between body mass index (BMI) and waist to hip ratio (WHR) among undergraduate students. *Pakistan Journal of Nutrition*, *15*(7), 618-624.
2. Jamal Hussaini, N. M., Abdullah, M. A., & Ismail, S. (2011). Recombinant Clone ABA392 protects laboratory animals from Pasteurella multocida Serotype B. *African Journal of Microbiology Research*, *5*(18), 2596-2599.
3. Hussaini, J., Nazmul, M. H. M., Masyitah, N., Abdullah, M. A., & Ismail, S. (2013). Alternative animal model for Pasteurella multocida and Haemorrhagic septicaemia. *Biomedical Research*, *24*(2), 263-266.
4. MKK, F., MA, R., Rashid, S. S., & MHM, N. (2019). Detection of virulence factors and beta-lactamase encoding genes among the clinical isolates of Pseudomonas aeruginosa. *arXiv preprint arXiv:1902.02014*.
5. Nazmul, M. H. M., Fazlul, M. K. K., Rashid, S. S., Doustjalali, S. R., Yasmin, F., Al-Jashamy, K., ... & Sabet, N. S. (2017). ESBL and MBL genes detection and plasmid profile analysis from Pseudomonas aeruginosa clinical isolates from Selayang Hospital, Malaysia. *PAKISTAN JOURNAL OF MEDICAL & HEALTH SCIENCES*, *11*(3), 815-818.
6. Keshireddy, S. R. (2019). Low-code application development using Oracle APEX productivity gains and challenges in cloud-native settings. *The SIJ Transactions on Computer Networks & Communication Engineering (CNCE)*, *7*(5), 20-24.
7. Keshireddy, S. R., & Kavuluri, H. V. R. (2019). Design of Fault Tolerant ETL Workflows for Heterogeneous Data Sources in Enterprise Ecosystems. *International Journal of Communication and Computer Technologies*, *7*(1), 42-46.
8. Arzuman, H., Maziz, M. N. H., Elsersi, M. M., Islam, M. N., Kumar, S. S., Jainuri, M. D. B. M., & Khan, S. A. (2017). Preclinical medical students perception about their educational environment based on DREEM at a Private University, Malaysia. *Bangladesh Journal of Medical Science*, *16*(4), 496-504.
9. Nazmul, M. H. M., Salmah, I., Jamal, H., & Ansary, A. (2007). Detection and molecular characterization of verotoxin gene in non-O157 diarrheagenic Escherichia coli isolated from Miri hospital, Sarawak, Malaysia. *Biomedical Research*, *18*(1), 39-43.