Rethinking XAI Evaluation: A Human-Centered Audit of Shapley Benchmarks in High-Stakes Settings

Inês Oliveira e Silva
University of Porto, Feedzai
Porto, Portugal
ines.silva@feedzai.com

Sérgio Jesus
Feedzai
Porto, Portugal
sergio.jesus@feedzai.com

Iker Perez
Feedzai
London, UK
iker.perez@feedzai.com

Rita P. Ribeiro
University of Porto
Porto, Portugal
rpribeiro@fc.up.pt

Carlos Soares
University of Porto
Porto, Portugal
csoares@fe.up.pt

Hugo Ferreira
Feedzai
Lisbon, Portugal
hugo.ferreira@feedzai.com

Pedro Bizarro
Feedzai
Lisbon, Portugal
pedro.bizarro@feedzai.com

Abstract

Shapley values are a cornerstone of explainable AI, yet their proliferation into competing formulations has created a fragmented landscape with little consensus on practical deployment. While theoretical differences are well-documented, evaluation remains reliant on quantitative proxies whose alignment with human utility is unverified. In this work, we use a unified amortized framework to isolate semantic differences between eight Shapley variants under the low-latency constraints of operational risk workflows. We conduct a large-scale empirical evaluation across four risk datasets and a realistic fraud-detection environment involving professional analysts and 3,735 case reviews. Our results reveal a fundamental misalignment: standard quantitative metrics, such as sparsity and faithfulness, are decoupled from human-perceived clarity and decision utility. Furthermore, while no formulation improved objective analyst performance, explanations consistently increased decision confidence, signaling a critical risk of automation bias and overconfidence in high-stakes settings. These findings demonstrate that current evaluation proxies are insufficient for predicting human impact, motivating our release of interaction measurements to support research into behaviorally grounded XAI evaluation.

1 Introduction

Machine learning (ML) systems are increasingly deployed in high-stakes domains such as fraud detection, credit assessment, and healthcare, where predictions have direct human and financial consequences [Khan et al., 2025, Mienye et al., 2024]. In these settings, model outputs rarely constitute final decisions. Instead, predictions are reviewed by human decision-makers operating under time, attention, and regulatory constraints. As a result, explanations are viewed as indispensable for accountability and oversight, and have become a core requirement in operational ML deployments [Bhatt et al., 2020, Bücker et al., 2022, Rudin, 2019]. Yet, despite widespread adoption, the practical value of explanations in human-in-the-loop workflows remains poorly understood, often assumed rather than empirically established [Amarasinghe et al., 2024, Bhatt et al., 2020, Jesus et al., 2021].

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Track on Evaluations and Datasets.

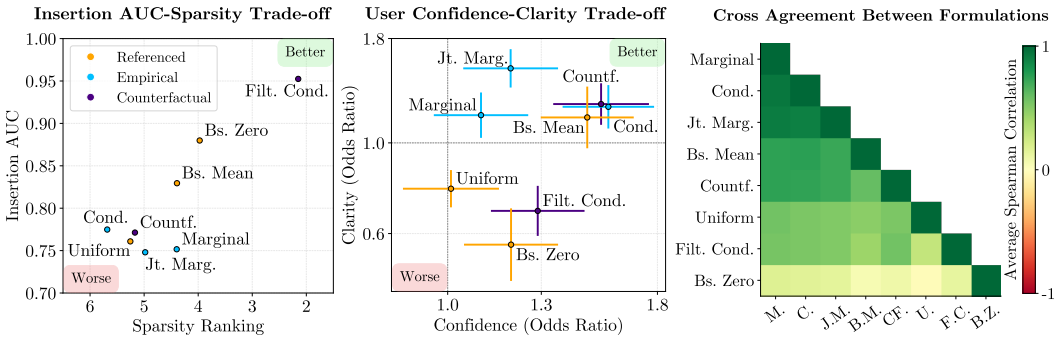

Figure 1: Overview of the empirical XAI audit in this paper. **Left:** Quantitative benchmarks reveal systematic functional variation (e.g., sparsity, faithfulness) across Shapley formulations. **Center:** Controlled analyst studies show differences in perceived clarity and decision confidence. **Right:** Pairwise agreement between Shapley formulations reveals structured alignment.

Among explanation methods, local approaches grounded in cooperative game theory, most notably *Shapley values*, have emerged as the *de facto* standard for feature attribution [Ali et al., 2023], by providing an axiomatic decomposition of model predictions into feature-level contributions [Lundberg and Lee, 2017]. However, the framework has fragmented into competing formulations based on divergent assumptions about the semantics of feature absence, realized in popular implementations such as KernelSHAP, TreeSHAP, and related tools [Albini et al., 2022, Chen et al., 2023, Covert and Lee, 2021, Lundberg et al., 2019, Olsen and Jullum, 2026, Yang, 2021, Zern et al., 2023]. This raises a critical evaluation question for practitioners: *does the choice of formulation matter to the end-user, and do standard evaluation procedures anticipate the impact?*

Modern XAI evaluation relies on theoretical analysis and quantitative proxies [Chen et al., 2023, Covert et al., 2021, Merrick and Taly, 2020], as well as mathematical distinctions between "faithfulness" to the model or "truthfulness" to the data [Chen et al., 2020]. Yet, systematic evidence regarding how explanation methods perform against human-centered benchmarks remains scarce. Existing evaluations often focus on isolated *functional* (model-based) properties and rarely stress-test these metrics against human behavior under realistic operational constraints [Canha et al., 2025, Nauta et al., 2023]. Furthermore, comparisons are often confounded by implementation choices, which mask the true semantic differences of the definitions themselves.

In this work, we audit the validity of XAI evaluation proxies by treating them as empirical hypotheses, and we contribute a high-fidelity behavioral corpus. We use Shapley values to instantiate a controlled study of the alignment between quantitative benchmarks and human utility through 3,735 granular human-AI interactions. We use a unified amortized framework to eliminate implementation confounders, and instantiate eight Shapley definitions under identical computational conditions. Our scope reflects high-throughput, high-stakes tabular data environments in low-latency financial, healthcare, e-commerce or logistics deployments [Grinsztajn et al., 2022, Sahakyan et al., 2021, Van Breugel and Van Der Schaar, 2024]. We conduct this audit across 5 datasets, including a real production-grade fraud-detection environment involving 37 analysts with professional experts.

Our analysis reveals a fundamental **decoupling between metrics and utility**: standard quantitative proxies are poor predictors of human-perceived clarity or decision confidence. Crucially, while explanations fail to improve objective analyst performance, they consistently boost confidence. This exposes a major deployment risk: current evaluation practices may favor explanations that encourage **automation bias** without improving decision quality. Figure 1 provides an overview of this landscape. Our contributions are:

- In Section 2, we describe a unified framework to isolate semantic explanation effects of Shapley formulations, eliminating algorithmic confounders for fair comparison.
- In Sections 3-5, we demonstrate through expert-driven reviews that common XAI evaluation metrics fail to align with human decision-making outcomes.
- In Section 6, we document a critical failure mode: the promotion of decision confidence without accuracy gains.

Table 1: Shapley value formulations categorized by method used to represent feature-absence.

Category	Definition	Sampling Distribution	Semantics / Assumptions
Referenced	Fixed Baseline	$\delta(\boldsymbol{X}_{\mathcal{F}\setminus\mathcal{S}} = \boldsymbol{x}'_{\mathcal{F}\setminus\mathcal{S}})$	Absent features fixed to a reference; efficient but highly sensitive to baseline selection.
	Uniform	$u(\boldsymbol{X}_{\mathcal{F}\setminus\mathcal{S}})$	Samples uniformly from a hyperbox; ignores data distribution and feature correlations.
Empirical	Marginal	$p(\boldsymbol{X}_{\mathcal{F}\setminus\mathcal{S}})$	Samples absent features from joint empirical distribution; breaks dependencies with $\boldsymbol{x}_{\mathcal{S}}$.
	Joint-Marginal	$\prod_{j\in\mathcal{F}\setminus\mathcal{S}} p(X_j)$	Samples from univariate marginals independently; removes all feature dependencies.
	Conditional	$p(\boldsymbol{X}_{\mathcal{F}\setminus\mathcal{S}} \mid \boldsymbol{X}_{\mathcal{S}} = \boldsymbol{x}_{\mathcal{S}})$	Preserves empirical dependencies via conditioning; hard to estimate in high-dimensions.
Counterfactual	Search Counterfactual	$p(\boldsymbol{X}^{(c)}_{\mathcal{F}\setminus\mathcal{S}} \mid f(\boldsymbol{X}^{(c)}_{\mathcal{F}}) \approx y^*)$	Uses counterfactual perturbations as baselines; high variance and optimization-intensive.
	Filtered Conditional	$p(\boldsymbol{X}_{\mathcal{F}\setminus\mathcal{S}} \mid f(\boldsymbol{X}_{\mathcal{F}}) \in \mathcal{Y})$	Conditions on background with specific model outputs; blends attribution with contrastive logic.

We further release non-proprietary code, interface designs, and 3,735 granular human-AI interaction measurements to support the reproducibility of our analysis and the development of behaviorally grounded XAI benchmarks.[1]

2 Shapley Value Foundations

Shapley values [Shapley, 1953] provide an additive decomposition of a model's output $f : \mathcal{X} \mapsto \mathbb{R}$ (e.g., risk scores or regression estimates) into feature-level contributions. Their axiomatic foundation and status as *de facto* standard for tabular applications [Gupte and Paparrizos, 2025, Sahakyan et al., 2021] is ideal for auditing alignment between XAI evaluation proxies and human utility.

Let $\mathcal{F} = \{1, \ldots, d\}$ denote the full feature set and $\mathcal{S} \subseteq \mathcal{F}$ a coalition of *present* features. For an input instance $\boldsymbol{x} \in \mathcal{X}$, a *value function* $v_{\boldsymbol{x}}(\mathcal{S})$ specifies the expected model output when only features in $\mathcal{S}$ are observed:

$$v_{\boldsymbol{x}}(\mathcal{S}) = \mathbb{E}_{\boldsymbol{X}_{\mathcal{F}\setminus\mathcal{S}} \sim p(\cdot)}\big[f(\boldsymbol{x}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{F}\setminus\mathcal{S}})\big], \tag{1}$$

where the *background* distribution $p(\cdot)$ defines the semantics of feature absence. The Shapley contribution of feature i to a model output is defined as

$$\phi_i = \sum_{\mathcal{S}\subseteq\mathcal{F}\setminus\{i\}} \frac{|\mathcal{S}|!(d - |\mathcal{S}| - 1)!}{d!} \big[v_{\boldsymbol{x}}(\mathcal{S} \cup \{i\}) - v_{\boldsymbol{x}}(\mathcal{S})\big], \tag{2}$$

which satisfies efficiency, symmetry, and dummy axioms [Lundberg and Lee, 2017, Sundararajan and Najmi, 2020].

A Taxonomy of Semantic Variants. Equation (1) highlights that Shapley values are a family of definitions induced by the choice of background distribution $p(\cdot)$. This choice reflects the tension between being "true to the model" versus "true to the data" [Chen et al., 2020] and can produce divergent attribution patterns [Albini et al., 2022, Datta et al., 2016, Janzing et al., 2020, Lundberg et al., 2019] with direct implications for human-in-the-loop workflows [Chen et al., 2023]. We focus our audit on definitions compatible with the low-latency, high-throughput requirements of production tabular systems, summarized in Table 1. In Appendix A we offer an extended discussion.

2.1 Amortization as an Experimental Control

Exact computation of Equation (2) requires evaluating 2^d feature coalitions, which is intractable for all but very small d. The SHAP framework [Lundberg and Lee, 2017] addresses this by recasting Shapley estimation as fitting an additive surrogate model $g(S) = v_{\boldsymbol{x}}(\emptyset) + \sum_{i\in S} \phi_i$, subject to an

[1] https://anonymous.4open.science/r/SHAP-Value-Function-Evaluation-16B4

efficiency constraint $f(\boldsymbol{x}) = g(\mathcal{F})$. In this setting, KernelSHAP [Lundberg and Lee, 2017] estimates ϕ_i for a given instance $\boldsymbol{x} \in \mathcal{X}$ by solving a weighted least-squares regression over sampled coalitions:

$$\underset{\phi_1,\ldots,\phi_d}{\arg\min} \sum_{S \subseteq \mathcal{F}} \pi(S)\big(v_{\boldsymbol{x}}(S) - g(S)\big)^2, \tag{3}$$

with kernel weights $\pi(S) = (d-1)\big[\binom{d}{|S|}|S|(d-|S|)\big]^{-1}$ emphasizing small and large coalitions. In practice, Monte Carlo approximations of Equation (3) introduce variance and bias [Goldwasser and Hooker, 2024]. Consequently, a fragmented landscape of model-specific heuristic optimizations has emerged [Lundberg et al., 2019, Muschalik et al., 2024, Yang, 2021, Yu et al., 2022], which confound comparisons across Shapley definitions through implementation artifacts, approximations, and heuristics rather than semantics.

To isolate semantic effects from implementation differences, we utilize **surrogates** or **amortizers** [Covert et al., 2024, Jethani et al., 2022] as a unified experimental control. Instead of solving (3) independently for each $\boldsymbol{x}$, a parametric universal approximator $\hat{\phi}_\theta(\boldsymbol{x})$ is trained to minimize the expected attribution loss over the data distribution:

$$\min_\theta \mathbb{E}_{\boldsymbol{X}}\left[\sum_{S \subseteq \mathcal{F}} \pi(S)\big(v_{\boldsymbol{X}}(S) - \hat{g}_\theta(S;\boldsymbol{X})\big)^2\right], \quad \text{where} \quad \hat{g}_\theta(S;\boldsymbol{x}) = v_{\boldsymbol{x}}(\emptyset) + \sum_{i \in S} \hat{\phi}_{i,\theta}(\boldsymbol{x}). \tag{4}$$

By construction, the surrogate $\hat{g}_\theta$ satisfies the efficiency constraint $\sum_i \hat{\phi}_{i,\theta}(\boldsymbol{x}) = f(\boldsymbol{x}) - v_{\boldsymbol{x}}(\emptyset)$. This reduces inference to a single forward pass, enabling the computational efficiency required for high-throughput systems [Covert et al., 2024, Dyer et al., 2024, Meyer et al., 2023]. Crucially, this separation of computational mechanics from theoretical semantics enables like-for-like comparisons across Shapley formulations under identical conditions.

3 Dimensions of XAI Evaluation

Prior research has established broad desiderata for explanation quality [Doshi-Velez and Kim, 2017, Saeed and Omlin, 2023, Vilone and Longo, 2021], converging on *functionally grounded properties* that characterize faithfulness to the predictive model, robustness to perturbations, selectivity, or truthfulness (see Appendix B for a detailed taxonomy). However, these properties are frequently treated as ends in themselves [Canha et al., 2025], often under the implicit assumption that optimizing for technical proxies leads to better human-AI outcomes. In this work, we demonstrate that statistical alignment between proxies and human utility should not be taken for granted [Buçinca et al., 2020, Ribeiro et al., 2016]. We evaluate Shapley attributions along two complementary axes: (i) *model-based* quantitative proxies, and (ii) *behavioral utility* measured through human interaction.

3.1 Quantitative Evaluation and Proxies

We operationalize a compact set of metrics to capture core functional properties prevalent in XAI literature. These proxies evaluate whether attributions exhibit behaviors theoretically associated with "good" explanations, independent of human judgment.

Deletion and Insertion AUC assess *faithfulness* by measuring how predictive entropy $H(\cdot)$ changes as features are removed or reintroduced in order of importance [Perez et al., 2022]. Let $\boldsymbol{x}^{(k)}$ denote an input with top-k features modified. Faithful explanations identify features whose removal quickly degrades model confidence and whose reintroduction rapidly restores it. For Deletion, this is captured by the area under the normalized entropy curve:

$$1 - \frac{1}{d}\sum_{k=1}^{d} \frac{H(\boldsymbol{x}^{(k)}) - H(\boldsymbol{x}^{(0)})}{H(\boldsymbol{x}^{(d)}) - H(\boldsymbol{x}^{(0)})}. \tag{5}$$

Analogously, Insertion AUC is computed by progressively adding top-ranked features to a baseline.

Perturbation Sensitivity measures local *stability* relative to model output changes under small perturbations $\epsilon > 0$ [Yeh et al., 2019]:

$$\|\hat{\phi}_\theta(\boldsymbol{x}+\epsilon) - \hat{\phi}_\theta(\boldsymbol{x})\|_2 \;/\; \left[|f(\boldsymbol{x}+\epsilon) - f(\boldsymbol{x})| + \delta\right], \tag{6}$$

where $\delta > 0$ ensures numerical stability. Lower values indicate smoother and robust explanations. Conversely, *Counterfactual Contrastivity* evaluates if explanations meaningfully differ across perturbed instances $\boldsymbol{x}'$ that cross the decision boundary ($f(\boldsymbol{x}) \neq f(\boldsymbol{x}')$):

$$\|\hat{\phi}_\theta(\boldsymbol{x}) - \hat{\phi}_\theta(\boldsymbol{x}')\|_2 \;/\; [|f(\boldsymbol{x}) - f(\boldsymbol{x}')| + \delta]\,. \tag{7}$$

Higher values indicate explanations that better reflect decision-relevant model changes.

Finally, *Sparsity* (L_1/L_2) quantifies the concentration of attribution mass across features, a property assumed to reduce cognitive load:

$$\|\hat{\phi}_\theta(\boldsymbol{x})\|_1 \;/\; \|\hat{\phi}_\theta(\boldsymbol{x})\|_2. \tag{8}$$

A lower ratio indicates selective attributions concentrated on fewer features, which is hypothesized to support human attention on a smaller subset of high-impact features [Canha et al., 2025].

3.2 Human-in-the-Loop Utility

To evaluate the operational impact of XAI, we use behavioral proxies that measure how effectively human decision makers perform AI-supported tasks when paired with Shapley attributions. *Decision Accuracy* ($\mathbb{I}\{y = \hat{y}\}$) and *Decision Time* ($\log(1 + t)$) serve as objective measures of performance and task efficiency. In addition, Perceived *Confidence* and *Clarity* quantify the subjective utility of explanations.

Crucially, by auditing the alignment between self-reported *Confidence* and objective *Accuracy*, we monitor **automation bias**, where explanations increase trust without improving decision quality.

4 Methodology

To establish a rigorous foundation for our audit, we proceed in two stages: (i) a technical validation of the amortized control framework, and (ii) a large-scale experimental study of human-AI alignment.

Our evaluation spans an unusually large-scale experimental matrix, where all formulations are compared under matched data, predictive models, and computational budgets: (i) **5 risk datasets**, including Maternal [Ahmed et al., 2020], Credit [Hofmann, 1994], HELOC [FICO, 2025], Adult [Ding et al., 2021], and a proprietary **real-world** fraud dataset; (ii) **2 industry-standard predictive models** in low-latency systems (Logistic Regression and LightGBM [Ke et al., 2017]); (iii) **8 distinct Shapley formulations** (Table 1); and (iv) **37 analysts**, including professional fraud-review experts. A total of **3735 case-review measurements** are released publicly (details in Appendices C–E).

4.1 Technical Control and Validation

To ensure that the amortizers accurately reconstruct the target Shapley definitions, we utilize two control measures following standard evaluation frameworks [Covert et al., 2024]:

- *Attribution Error*: The mean squared error (MSE) between amortized $\hat{\phi}_{j,\theta}(\boldsymbol{x})$ and a high-sample KernelSHAP ground truth $\phi_j(\boldsymbol{x})$; and
- *Recall@k*: The overlap across the top-k most influential features between amortized and ground truth attributions.

In all cases, KernelSHAP reference values are tailored to each specific Shapley definition. Finally, we measure *Cross-Agreement* (Spearman correlation) between different Shapley formulations; this quantifies how the choice of definition alters the explanation rank-order.

4.2 Analyst Workflow and Interface

Experiments in the audit use a standardized interface (Figure 2) grounded in academic conventions [Lundberg and Lee, 2017, Wysocki et al., 2023] and mirroring professional risk tools [Jesus et al., 2021]. It presents a model score/percentile, a Shapley attribution bar chart, and natural-language reason codes derived from the attributions.

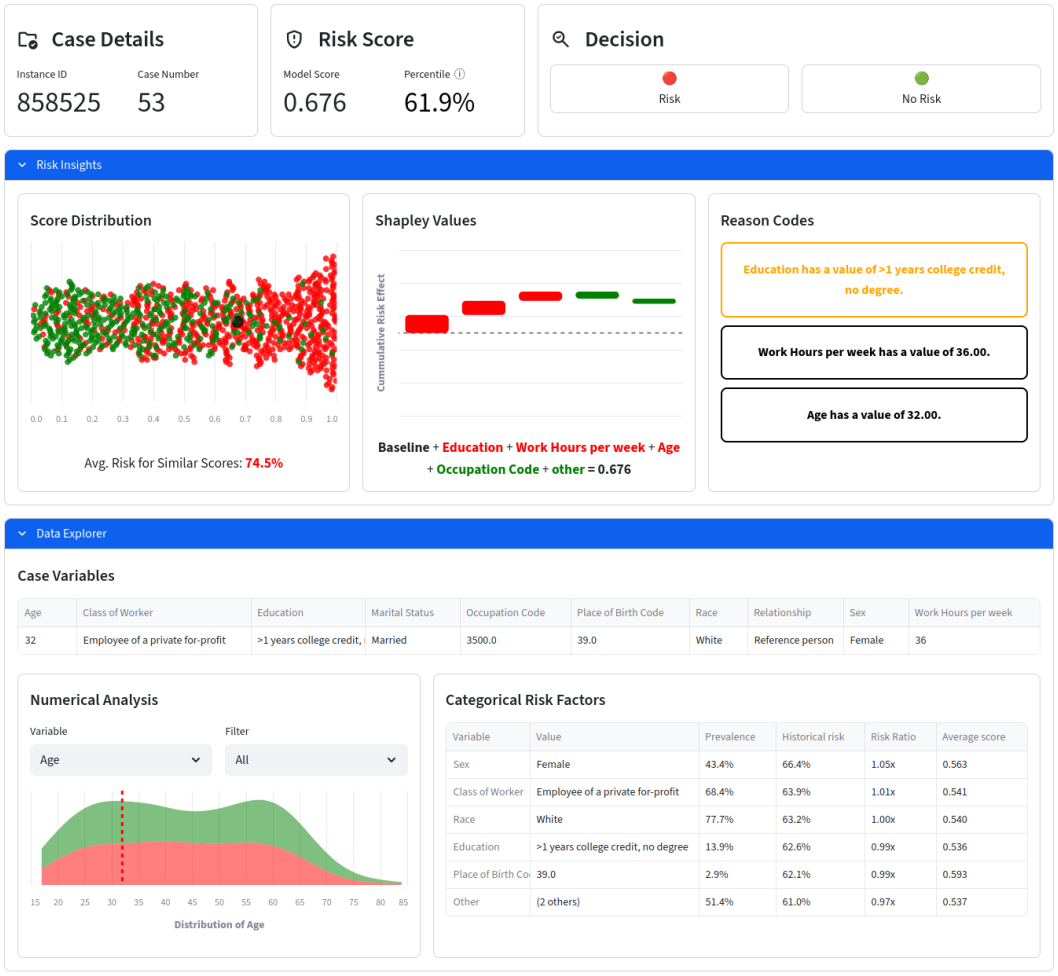

Figure 2: The case review interface integrates model scores, Shapley attributions, and contextual data summaries while holding visual encodings fixed across Shapley formulations.

To isolate the formulation's semantic content as the primary behavioral driver, experiments follow a blinded, randomized *within-subjects* design with a no-explanation control. Participants review cases balanced across operational decision thresholds, while we hold the layout and interaction mechanisms constant. We utilize two deployments: an open-source platform for benchmarks and a sandboxed system integrated into professional fraud review pipelines. Analysts submit binary decisions (*risk*/*no risk*) and report perceived clarity and confidence. Each review produces a structured log of decision outcomes, timestamps, and analyst metadata (details in Appendix E).

4.3 Statistical Framework

We integrate all evaluations within a unified statistical pipeline that combines aggregate quantitative proxies with analyst-level effects under repeated observations.

Aggregation of Quantitative Proxies. We distinguish between *scale-invariant* and *scale-dependent* metrics. Scale-invariant metrics (e.g., Deletion AUC, Recall@k) are averaged directly across dataset–model pairs as they share a consistent interpretation. Conversely, scale-dependent metrics (e.g., sensitivity, sparsity) often exhibit units and variances that are not comparable across different data distributions. For these, we convert raw scores into *ordinal rankings* within each dataset–model pair before aggregation. All results are reported as point estimates with standard errors obtained via bootstrapping (50 subsamples per pair).

Table 2: Quantitative results across scale-invariant and dependent metrics (bootstrapped standard deviations). Scale-dependent metrics use relative Rank statistics. **Bold** indicates best outcomes.

Value F.	Rank Scale Dependent Metrics				Scale Invariant Metrics				
	Spar.	Sensit.	Contr.	Attr. Er.	Del. AUC	Ins. AUC	Recall@1	Recall@3	Recall@5
Base. Zero	3.9 (2.7)	4.1 (2.8)	4.4 (2.9)	3.4 (2.5)	**0.11 (0.08)**	0.88 (0.20)	**0.90 (0.07)**	**0.90 (0.06)**	**0.91 (0.05)**
Base. Mean	4.3 (2.3)	5.2 (1.8)	4.6 (1.5)	3.5 (2.1)	0.19 (0.13)	0.83 (0.13)	0.83 (0.11)	0.85 (0.06)	0.87 (0.07)
Uniform	5.2 (2.3)	**2.8 (1.7)**	6.0 (2.2)	5.0 (1.8)	0.15 (0.24)	0.76 (0.14)	0.86 (0.10)	0.85 (0.06)	0.86 (0.06)
Marginal	4.4 (1.3)	3.6 (1.5)	4.1 (1.9)	3.5 (1.1)	0.16 (0.10)	0.75 (0.09)	0.80 (0.12)	0.84 (0.07)	0.86 (0.08)
Joint-Marg.	4.9 (1.9)	3.5 (1.5)	5.3 (1.2)	**3.1 (1.2)**	0.16 (0.10)	0.75 (0.09)	0.82 (0.08)	0.82 (0.07)	0.84 (0.07)
Conditional	5.6 (1.6)	4.7 (1.6)	3.9 (1.1)	6.2 (1.5)	0.16 (0.11)	0.77 (0.09)	0.63 (0.24)	0.72 (0.10)	0.78 (0.09)
Counterf.	5.2 (2.2)	5.2 (2.6)	5.9 (1.9)	7.4 (0.8)	0.20 (0.08)	0.77 (0.09)	0.53 (0.10)	0.63 (0.10)	0.70 (0.13)
Filt. Cond.	**2.1 (1.3)**	6.7 (1.4)	**1.5 (0.7)**	3.7 (2.0)	0.18 (0.08)	**0.95 (0.05)**	0.90 (0.08)	0.88 (0.06)	0.89 (0.06)
	Lower is better (↓)					**Higher is better** (↑)			

Inferential Modeling. We isolate explanation effects from confounders like case difficulty, analyst expertise, or repeated exposure, using *mixed-effects modeling* [Fitzmaurice et al., 2012, Nelder and Wedderburn, 1972, Pinheiro and Bates, 2000]. We fit a global model across datasets, predictive models, and analysts, to account for structured heterogeneity and maximize statistical power for generalizable inference. We model effects relative to a no-explanation baseline for *accuracy* (logistic), *confidence* (ordinal), and *time* (log-linear). Perceived *clarity* is modeled relative to the population mean, excluding no-explanation cases. Fixed effects include the Shapley formulation, *model type*, *dataset*, *model entropy*, *prediction error*, and *analyst experience*, with further controls for self-reported *subject domain*, *ML*, and *Shapley* familiarity. Results are reported as odds ratios or multiplicative effects with 95% confidence intervals (details in Appendix E).

Synthesis. Finally, we address a core Research Question: *To what extent do offline functional proxies predict online human utility?* To answer this, we extend the mixed-effects framework at the *case level*, using quantitative explanation properties as predictors for perceived *clarity* and *confidence*, applying the same rigorous confounder controls. Thus, we statistically determine if offline proxies are reliable predictors of human utility in operational settings.

5 Results

We present results in three stages: a technical benchmark of formulation properties, a behavioral analysis of analyst utility, and a synthesis of the alignment between the two. For technical implementation choices during the amortization of formulations we refer the reader to Appendix A.

5.1 Quantitative Benchmarks

Benchmarks in Table 2 reveal that no formulation dominates across all functional properties. Instead, we observe a structured trade-off between *stability* and *selectivity*. High-sparsity formulations like Filtered Conditional and Zero-baseline achieve superior contrastivity and insertion AUC, reflecting selective attributions, but increase *sensitivity* to small input perturbations. Conversely, *smoother* formulations like Uniform are stable but produce dense, less selective explanations.

Amortization as Experimental Control. Across formulations, amortization achieves high fidelity to the *ground truth* reference, with low Attribution Error and high Recall@k. Cross-alignment analysis in Figure 1 (right) further shows that formulations cluster into coherent families. Thus, results below reflect the semantic intent of Shapley formulations rather than approximation artifacts.

Key trends across formulation families include:

- **Fixed baselines:** A Zero baseline yields strong Deletion AUC (0.11), identifying *must-have* features for prediction, while the Uniform variant trades stability (Sensitivity Rank 2.8) for

Table 3: Fixed effects on human outcomes. Decision time reported as multiplicative effect; Accuracy, Clarity, and Confidence as Odds Ratios. Standard errors retrieved via the Delta Method. Bold values indicate statistical significance ($p < 0.05$).

Value Func.	Decision Time (↓)			Odds Ratio (↑)		
	p2.5	p50	p97.5	Accuracy	Clarity	Confidence
Base. Zero	**1.29 (0.09)**	**1.16 (0.05)**	1.03 (0.07)	1.04 (0.20)	**0.57 (0.06)**	1.20 (0.15)
Base. Mean	0.99 (0.07)	1.03 (0.04)	0.88 (0.06)	1.05 (0.21)	1.17 (0.14)	**1.49 (0.19)**
Uniform	1.11 (0.08)	**1.11 (0.05)**	1.02 (0.08)	0.84 (0.16)	**0.79 (0.09)**	1.02 (0.13)
Marginal	0.97 (0.07)	1.01 (0.04)	1.08 (0.08)	0.76 (0.14)	1.18 (0.14)	1.11 (0.14)
Joint-Marg.	**1.17 (0.08)**	1.06 (0.05)	0.99 (0.07)	1.19 (0.23)	**1.52 (0.19)**	1.20 (0.16)
Conditional	0.98 (0.07)	1.05 (0.04)	0.94 (0.07)	0.99 (0.19)	1.23 (0.14)	**1.58 (0.20)**
Counterf.	1.08 (0.07)	1.00 (0.04)	1.15 (0.08)	1.02 (0.20)	1.25 (0.16)	**1.55 (0.20)**
Filt. Cond.	1.03 (0.07)	1.04 (0.05)	0.98 (0.07)	0.79 (0.15)	**0.68 (0.08)**	**1.30 (0.16)**
Log Count	**0.84 (0.02)**	**0.79 (0.01)**	**0.80 (0.02)**	0.93 (0.06)	1.01 (0.07)	0.94 (0.05)
Model Entropy	**1.25 (0.03)**	**1.23 (0.02)**	**1.07 (0.03)**	0.95 (0.08)	**0.68 (0.05)**	**0.44 (0.05)**
Score Error	1.00 (0.02)	1.00 (0.02)	0.99 (0.02)	**0.19 (0.02)**	**0.86 (0.05)**	**0.80 (0.03)**
Professional Analyst	**0.85 (0.05)**	**0.88 (0.04)**	1.10 (0.07)	0.94 (0.17)	1.08 (0.17)	1.10 (0.13)

reduced sparsity and contrastivity. On the other hand, the Mean baseline is closely aligned to popular empirical formulations (Figure 1, right), but delivers no standout advantages.

- **Empirical formulations:** Marginal and Joint-Marginal variants show mutual agreement and balanced performance, with low sensitivity (Ranks 3.5–3.6), along with moderate sparsity (4.4–4.9) and contrastivity (4.1–5.3). This yields low cognitive load paired with low volatility. In contrast, conditional Shapley produces dense, volatile attributions that reflect *data correlations rather than model behavior* [Chen et al., 2020].
- **Counterfactual formulations:** The Filtered Conditional variant maximizes sparsity (Rank 2.1), contrastivity (1.5), and Insertion AUC, prioritizing the *why not* of a decision. However, this is most unstable under perturbations (Sensit. Rank 6.7). Fully counterfactual explanations are similarly dense and unstable.

5.2 Human-in-the-Loop Utility

Analyst behavior results in Table 3 and Figure 3 (left) reveal a critical asymmetry: formulation choice does not significantly alter *what* analysts decide, but strongly influences *how they feel* about those decisions. Across formulations:

- **Objective performance:** No Shapley definition reliably improves accuracy or complex-case ($p_{97.5}$) decision time. Instead, effects are driven by case-level factors such as *Score Error* (Accuracy OR 0.19) and *Repeated Exposure* (Decision time Log Count OR 0.95). This highlights that even high-quality explanations do not easily override task difficulty.
- **Perceived clarity:** Clarity is highly sensitive to formulation choice. Joint-Marginal explanations, which balance stability and selectivity, significantly improve clarity (OR 1.52). Conversely, sparse formulations such as the Zero baseline (OR 0.57) and Filtered Conditional (OR 0.68) are consistently rated as confusing.
- **Subjective confidence:** Several explainers (Conditional, Counterfactual, and Mean) substantially inflated analyst confidence (OR > 1.49). Coupled with the lack of accuracy gains, this signals a risk of **over-reliance** and **automation bias**, where explanations encourage trust even when analyst predictions are erroneous.

Ultimately, explanations modulate analyst perception, but **do not override task difficulty**, aligning with recent literature on XAI and complementary performance [Bansal et al., 2021, Poursabzi-Sangdeh et al., 2021, Sivaraman et al., 2023].

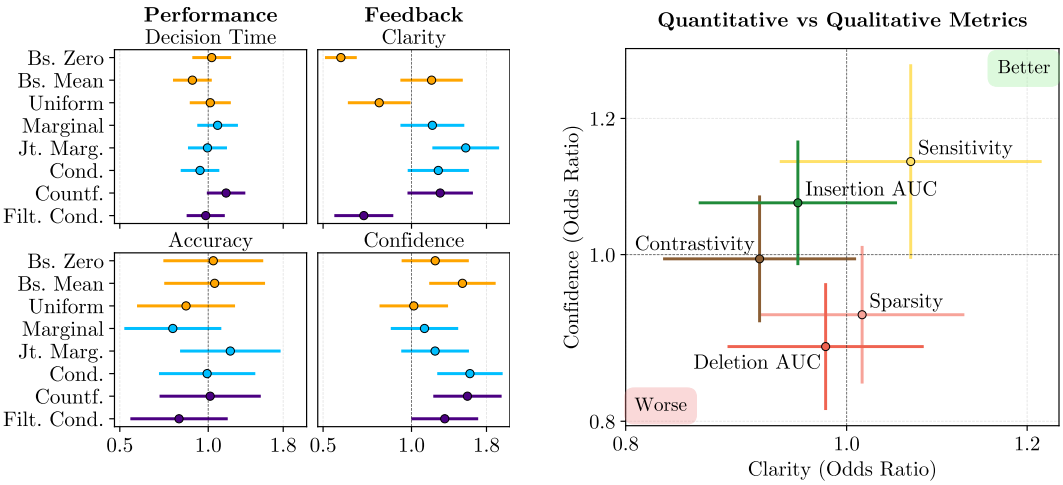

Figure 3: **Left:** Effect sizes on human outcomes, with 95% confidence intervals, across accuracy, decision time ($p_{97.5}$), confidence and clarity. X-axis is in log-scale. **Right:** Effect of quantitative explanation metrics on clarity and decision confidence, with 95% confidence intervals.

5.3 Alignment: Do Proxies Predict Human Outcomes?

We finally examine the alignment between quantitative proxies and human outcomes (Figure 3, right) by using functional properties as predictors for *clarity* and *confidence*. In all cases, we control for case difficulty, prediction error, analyst experience, and expertise.

Results reveal a **decoupling** between quantitative metrics and perceived utility: standard proxies assumed to associate with *good* explanations show **no positive relationship** with reported clarity or confidence. Notably, *sparsity* may be **negatively** associated with confidence, contradicting the assumption that low-dimensional explanations enhance intelligibility [Canha et al., 2025]. All in all, widely used benchmarks capture structural properties but fail to predict human perception. While valuable for debugging, they are **not substitutes for behavioral evaluation** in operational settings, which demonstrates the necessity of large-scale expert interaction studies and datasets, as contributed here, to construct a basis for auditing XAI utility.

6 Conclusion

This paper has conducted a unified large-scale empirical study and audit of Shapley value formulations, testing the alignment between functional quantitative proxies and human utility in high-stakes risk workflows. By evaluating eight semantic variants under strictly matched data, model, interfaces, and computational conditions, we proved fundamental **decoupling** in XAI evaluation: standard quantitative benchmarks capture structural properties of explanations but fail to predict how they are perceived or used by final users. Furthermore, while no Shapley formulation improved objective accuracy, several significantly inflated analyst confidence. This uncovers a systemic risk of **automation bias** where explanations encourage over-reliance without providing cognitive gains.

The results have implications for both research and practice. Based on our audit and the accompanying 3,735 measurements, we offer the following guidance for the deployment and evaluation of XAI:

- Practitioners must **decouple algorithms from users**, and treat proxy metrics (e.g., deletion AUC, sensitivity) as debugging tools for model-faithfulness, not as proxies for human interpretability or trust.
- System designers should **prioritize empirical formulations** for clarity (Marginal, Joint-Marginal, Conditional), while remaining wary of unstable and highly sensitive *Conditional* or *Counterfactual* variants that inflate confidence without commensurate gains in accuracy.

Limitations and Scope. Our audit focused on low-latency, tabular risk models as a backbone to financial, e-commerce or medical decision-making [Van Breugel and Van Der Schaar, 2024].

However, results may differ in vision or language domains where feature semantics follow different dynamics. Additionally, our experiments were conducted in controlled settings and cannot capture longer-term effects such as learning, adaptation, or changes in institutional decision norms.

Ultimately, our results motivate a shift in the XAI community toward behaviorally grounded evaluation; until metrics are proven to align with human outcomes, they cannot serve as substitutes for rigorous user studies in high-stakes production systems.

References

K. Aas, M. Jullum, and A. Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502, 2021.

M. Ahmed, M. A. Kashem, M. Rahman, and S. Khatun. Review and analysis of risk factor of maternal health in remote area using the internet of things (iot). In *InECCE2019*, pages 357–365. Springer Singapore, 2020. ISBN 978-981-15-2317-5.

T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, 2019. ISBN 9781450362016.

E. Albini, J. Long, D. Dervovic, and D. Magazzeni. Counterfactual shapley additive explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1054–1070. Association for Computing Machinery, 2022.

S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805, 2023.

K. Amarasinghe, K. T. Rodolfa, S. Jesus, V. Chen, V. Balayan, P. Saleiro, P. Bizarro, A. Talwalkar, and R. Ghani. On the importance of application-grounded experimental design for evaluating explainable ml methods. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38: 20921–20929, 2024.

G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–16, 2021.

U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, and P. Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, page 648–657. Association for Computing Machinery, 2020.

Z. Buçinca, P. Lin, K. Z. Gajos, and E. L. Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th international conference on intelligent user interfaces*, pages 454–464, 2020.

M. Bücker, G. Szepannek, A. Gosiewska, and P. Biecek. Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society*, 73(1): 70–90, 2022.

D. Canha, S. Kubler, K. Främling, and G. Fagherazzi. A functionally-grounded benchmark framework for xai methods: Insights and foundations from a systematic literature review. *ACM Comput. Surv.*, 57(12), 2025.

H. Chen, J. D. Janizek, S. Lundberg, and S.-I. Lee. True to the model or true to the data? *arXiv preprint arXiv:2006.16234*, 2020.

H. Chen, I. C. Covert, S. M. Lundberg, and S.-I. Lee. Algorithms to estimate shapley value feature attributions. *Nature Machine Intelligence*, 5(6):590–601, 2023.

I. Covert and S.-I. Lee. Improving kernelshap: Practical shapley value estimation using linear regression. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3457–3465. PMLR, 2021.

I. Covert, S. Lundberg, and S.-I. Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021.

I. Covert, C. Kim, S.-I. Lee, J. Zou, and T. Hashimoto. Stochastic amortization: a unified approach to accelerate feature and data attribution. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24. Curran Associates Inc., 2024.

A. Datta, S. Sen, and Y. Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 598–617. IEEE Computer Society, 2016.

F. Ding, M. Hardt, J. Miller, and L. Schmidt. Retiring adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 6478–6490. Curran Associates, Inc., 2021.

F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning, 2017.

J. Dyer, N. Bishop, Y. Felekis, F. M. Zennaro, A. Calinescu, T. Damoulas, and M. Wooldridge. Interventionally consistent surrogates for complex simulation models. In *Advances in Neural Information Processing Systems*, volume 37, pages 21814–21841. Curran Associates, Inc., 2024.

FICO. Fico expands educational analytics challenge program with three new historically black colleges and universities to educate aspiring data scientists, 2025. Accessed: 2026-01-02.

G. M. Fitzmaurice, N. M. Laird, and J. H. Ware. *Applied longitudinal analysis*. John Wiley & Sons, 2012.

J. Goldwasser and G. Hooker. Stabilizing estimates of shapley values with control variates. In *Explainable Artificial Intelligence*, pages 416–439. Springer Nature Switzerland, 2024.

L. Grinsztajn, E. Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520, 2022.

S. Gupte and J. Paparrizos. Understanding the black box: A deep empirical dive into shapley value approximations for tabular data. *Proceedings of the ACM on Management of Data*, 3(3):1–31, 2025.

T. Heskes, E. Sijben, I. G. Bucur, and T. Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In *Advances in Neural Information Processing Systems*, volume 33, pages 4778–4789. Curran Associates, Inc., 2020.

H. Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994.

D. Janzing, L. Minorics, and P. Bloebaum. Feature relevance quantification in explainable ai: A causal problem. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2907–2916. PMLR, 2020.

S. Jesus, C. Belém, V. Balayan, J. a. Bento, P. Saleiro, P. Bizarro, and J. a. Gama. How can i choose an explainer? an application-grounded evaluation of post-hoc explanations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2021. ISBN 9781450383097.

S. Jesus, P. Saleiro, I. O. e Silva, B. M. Jorge, R. P. Ribeiro, J. Gama, P. Bizarro, and R. Ghani. Aequitas flow: Streamlining fair ml experimentation. *Journal of Machine Learning Research*, 25: 1–7, 2024.

N. Jethani, M. Sudarshan, I. C. Covert, S.-I. Lee, and R. Ranganath. FastSHAP: Real-time shapley value estimation. In *International Conference on Learning Representations*, 2022.

A.-H. Karimi, B. Schölkopf, and I. Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 353–362. Association for Computing Machinery, 2021.

G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

F. S. Khan, S. S. Mazhar, K. Mazhar, D. A. AlSaleh, and A. Mazhar. Model-agnostic explainable artificial intelligence methods in finance: a systematic review, recent developments, limitations, challenges and future directions. *Artificial Intelligence Review*, 58(8):232, 2025.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 4768–4777. Curran Associates Inc., 2017.

S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent individualized feature attribution for tree ensembles, 2019.

L. Merrick and A. Taly. The explanation game: Explaining machine learning models using shapley values. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 17–38. Springer International Publishing, 2020.

L. T. Meyer, M. Schouler, R. A. Caulk, A. Ribes, and B. Raffin. Training deep surrogate models with large scale online learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 24614–24630. PMLR, 2023.

I. D. Mienye, G. Obaido, N. Jere, E. Mienye, K. Aruleba, I. D. Emmanuel, and B. Ogbuokiri. A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges. *Informatics in Medicine Unlocked*, 51:101587, 2024.

C. Molnar, G. Casalicchio, and B. Bischl. Interpretable machine learning – a brief history, state-of-the-art and challenges. In *ECML PKDD 2020 Workshops*, pages 417–431. Springer International Publishing, 2020.

R. K. Mothilal, A. Sharma, and C. Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, 2020. Association for Computing Machinery.

M. Muschalik, F. Fumagalli, B. Hammer, and E. Hüllermeier. Beyond treeshap: efficient computation of any-order shapley interactions for tree ensembles. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'24. AAAI Press, 2024.

M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, and C. Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Comput. Surv.*, 55(13s), 2023.

J. A. Nelder and R. W. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135:370–384, 1972.

L. H. B. Olsen and M. Jullum. Improving the weighting strategy in kernelshap. In *Explainable Artificial Intelligence*, pages 194–218. Springer Nature Switzerland, 2026.

I. Perez, P. Skalski, A. Barns-Graham, J. Wong, and D. Sutton. Attribution of predictive uncertainties in classification models. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 1582–1591. PMLR, 2022.

J. C. Pinheiro and D. M. Bates. *Mixed-effects models in S and S-PLUS*. Springer, 2000.

F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52, 2021.

M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

W. Saeed and C. Omlin. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273, 2023.

M. Sahakyan, Z. Aung, and T. Rahwan. Explainable artificial intelligence for tabular data: A survey. *IEEE access*, 9:135392–135422, 2021.

L. S. Shapley. A value for n-person games. In *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton, 1953.

V. Sivaraman, L. A. Bukowski, J. Levin, J. M. Kahn, and A. Perer. Ignore, trust, or negotiate: understanding clinician acceptance of ai-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2023.

E. Štrumbelj and I. Kononenko. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11(1):1–18, 2010.

M. Sundararajan and A. Najmi. The many shapley values for model explanation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9269–9278. PMLR, 2020.

M. F. Taufiq, P. Blöbaum, and L. Minorics. Manifold restricted interventional shapley values. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 5079–5106. PMLR, 2023.

B. Van Breugel and M. Van Der Schaar. Position: Why tabular foundation models should be a research priority. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 48976–48993. PMLR, 2024.

G. Vilone and L. Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021.

O. Wysocki, J. K. Davies, M. Vigo, A. C. Armstrong, D. Landers, R. Lee, and A. Freitas. Assessing the communication gap between ai models and healthcare professionals: Explainability, utility and trust in ai-driven clinical decision-making. *Artificial Intelligence*, 316:103839, 2023.

J. Yang. Fast treeshap: Accelerating shap value computation for trees. *ArXiv*, abs/2109.09847, 2021.

C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, and P. K. Ravikumar. On the (in)fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

P. Yu, A. Bifet, J. Read, and C. Xu. Linear tree shap. In *Advances in Neural Information Processing Systems*, volume 35, pages 25818–25828. Curran Associates, Inc., 2022.

A. Zern, K. Broelemann, and G. Kasneci. Interventional shap values and interaction values for piecewise linear regression trees. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11164–11173, 2023.

A Extended Overview of Shapley Value Formulations

The definition of the value function $v_{\boldsymbol{x}}(\mathcal{S})$ hinges on how one represents feature **absence**. These formulations are not interchangeable; they encode fundamental assumptions about the data-generating process and the intended use of the explanation. Table 1 summarizes the formulations considered in this work. Here, we extend their definitions with a discussion on use cases and limitations, and discuss technical amortization choices and challenges where necessary.

A.1 Fixed Baseline Shapley Values

The *fixed baseline* formulation replaces absent features with values from a predefined reference input $\boldsymbol{x}'$ [Lundberg and Lee, 2017, Sundararajan and Najmi, 2020]. While computationally efficient (requiring only one model evaluation per coalition), this method is highly sensitive to the choice of $\boldsymbol{x}'$. As there is rarely a "neutral" reference point in tabular data [Chen et al., 2023], the choice of zero, mean, or median can fundamentally shift the attribution logic.

A.2 Marginal (Interventional) Shapley Values

In the *marginal* formulation [Janzing et al., 2020, Sundararajan and Najmi, 2020], absent features are sampled from their global empirical distribution. This corresponds to `do`-intervention under a causal graph with no feature dependencies [Heskes et al., 2020, Janzing et al., 2020]. While it isolates the model's response to specific features, it frequently generates **off-manifold** inputs, i.e. combinations of values that could never occur in reality. This may lead to misleading attributions in highly correlated datasets [Merrick and Taly, 2020].

A.3 Joint-Marginal Shapley Values

The *joint-marginal* variant treats every absent feature as independent, sampling each from its univariate marginal [Datta et al., 2016]. By enforcing independence, it simplifies computation but amplifies the *off-manifold* problem. In our study, this variant was rated as the **clearest** by analysts, suggesting that "true-to-the-model" logic may be easier for humans to parse than complex distributions.

A.4 Conditional Shapley Values

The *conditional* formulation preserves empirical dependencies by conditioning on observed coalition values [Aas et al., 2021, Chen et al., 2023]. This produces realistic *in-manifold* imputations but is computationally intensive and difficult to estimate in high dimensions, where practical implementations rely on approximate matching, nearest neighbours, or parametric density models. During amortization, we implement it using **proximity-weighted sampling** on a background set. Our audit shows this variant significantly **inflates analyst confidence** without improving accuracy.

A.5 Uniform Shapley Values

The *uniform* formulation replaces absent features with samples drawn uniformly from a predefined hyperbox [Merrick and Taly, 2020, Štrumbelj and Kononenko, 2010]. While useful for exploring the model's global behavior, it often results in noisy, confusing explanations for human reviewers.

A.6 Search Counterfactual Shapley Values

This formulation identifies the minimal changes needed to flip a model's prediction [Albini et al., 2022]. We implement this using optimization-based generation via **DiCE** [Mothilal et al., 2020]. It replaces absent features with values from a vector $\boldsymbol{X}^{(c)}$ that reaches a target outcome y^*. This emphasizes **actionability** but is computationally demanding [Karimi et al., 2021], and can be sensitive to chosen distance metrics or plausibility constraints.

A.7 Filtered Conditional Shapley Values

This variant restricts the background distribution to samples where model outputs fall within a specific range $\mathcal{Y}$. We implement this by **filtering the empirical data** to only include instances where the predicted model risk score remains under a low threshold. In our audit, this produced the most **sparse** and **contrastive** explanations, yet was consistently rated as confusing by professional analysts.

A.8 Excluded Formulations

We excluded causal [Heskes et al., 2020] and generative in-manifold [Taufiq et al., 2023] variants due to their high computational overhead and requirement for strong causal assumptions. These are

Table 4: High-level overview of key functional properties of explanation quality [Taufiq et al., 2023], with example attributes.

Property	What it captures	Example quantitative / qualitative attributes
Representativeness	**What** explanation is provided.	Local vs. global, model portability or data coverage.
Structure	**How** is the explanation provided.	Expressive power, graphical integrity, morphological clarity.
Selectivity	**Size** of the explanation.	How concise. Sparsity and top-k relevance.
Contrastivity	**Relational** differences.	Sensitivity to counterfactuals, benchmarks or perturbations.
Interactivity	User **control**.	Interaction options, adjustable parameters or interfaces.
Fidelity	**Surrogate** consistency	Approximation accuracy and model assumptions.
Faithfulness	How **reliable** for the black-box.	Causal correctness, feature removal and white-box checks.
Truthfulness	Domain **alignment**.	Expert plausibility; bias exposure.
Stability	Behavioural **consistency**.	Similarity across perturbations or resampling.
(Un)certainty	How **transparent** is it?	Confidence disclosure, stochastic awareness.
Speed	Computational **efficiency**.	Latency; throughput under fixed budgets.

currently incompatible with the millisecond-level latency required in production risk systems and are thus outside the scope of our operational audit.

B Properties of Explanation Quality

Recent taxonomies of explainability [Canha et al., 2025] identify a broad set of *functionally grounded properties* that characterize a "good" explanation. These span computational, representational, and human-centered dimensions. This appendix refines these concepts for the *Shapley value framework*, distinguishing between properties that are inherent to the framework and those that are formulation-dependent and thus central to our audit. Table 4 provides a high-level overview.

B.1 Properties Invariant Across Shapley Configurations

The following properties are inherent to the Shapley framework itself, and remain constant across all eight variants in our study. Thus, they do not serve as discriminators in our evaluation.

Representativeness. Shapley values are, by construction, *local feature attributions* that are model-agnostic [Aas et al., 2021, Lundberg and Lee, 2017]. All formulations share this scope; therefore, representativeness does not vary between definitions.

Structure. The structural form of a Shapley explanation is always additive. Consequently, the visual format (e.g., bar charts) is independent of the value-function choice. All variants in our study were presented using the exact same visual encoding to ensure structural parity.

Interactivity. Interaction depends on the UI (e.g., toggles, sliders) rather than the mathematical definition. By holding the interface fixed, we ensure interactivity remains a constant across our experimental matrix.

(Un)certainty. While uncertainty estimates can be attached to explanations (e.g., via ensemble bootstraps [Covert et al., 2021]), these quantify the stochasticity of the underlying model or the estimation procedure, not the conceptual choice of the value function.

Speed. While training costs vary, the *inference* speed is standardized by our use of amortized surrogates. This ensures that the operational feasibility of the explanations is consistent across all formulations during the analyst review process.

B.2 Properties Relevant for Shapley Evaluation

These properties depend directly on the choice of value function and form the basis of our empirical audit.

Table 5: Summary statistics for experimental datasets; including sample sizes for training, validation, and A/B testing with amortisers. Prevalence refers to the proportion of positive (risk) instances.

Dataset	Feat.	Cat.	Prev.	Train	Val.	Test
Maternal Risk	6	0	26.8%	812	101	101
German Credit	20	11	30.0%	800	100	100
Adult	10	6	63.1%	998,300	200	200
HELOC	22	0	52.1%	7,029	200	200
Fraud Risk*	54	15	9.1%	16,496	200	200

*Heavily downsampled to achieve a 10-1 risk ratio before amortization.

Selectivity. This captures how concentrated attributions are across features, and determines the amount of information users must process, affecting cognitive load [Canha et al., 2025]. Different formulations produce different sparsity profiles. We quantify this via the *Sparsity* metric (8), testing the assumption that "less is more" for human interpretability.

Contrastivity It captures whether explanations express differences relative to a meaningful reference or baseline [Canha et al., 2025]. Certain Shapley definitions (e.g., counterfactual, causal) are inherently contrastive. Other definitions may conflate average and specific contributions, leading to sensitivity to arbitrary baselines. We operationalize this through sensitivity to counterfactual perturbations (7).

Fidelity Fidelity reflects how accurately the additive decomposition of Shapley attributions reconstruct a black-box prediction. Specifically, how accurately an amortized surrogate reconstructs the "ground truth" for its specific definition. We use *Attribution Error* and *Recall@k* as control measures to ensure that our surrogates are mathematically sound.

Faithfulness It captures whether the explanation preserves the causal influence of features on predictions [Molnar et al., 2020]. Because different value functions define different intervention semantics, faithfulness varies significantly. We evaluate this using *Deletion and Insertion AUC* (5).

Truthfulness Truthfulness connects the explanation to domain reality and expert priors [Canha et al., 2025]. Some value functions can yield explanations that contradict established domain knowledge. We assess this through our human-in-the-loop experiments in Section 3.2, measuring whether an explanation helps or hinders an analyst's decision-making process.

Stability Stability denotes reproducibility under small input or model perturbations. Formulations that integrate over correlated feature distributions tend to produce smoother attributions. We operationalize this via the *Perturbation Sensitivity* metric (6).

C Datasets

We evaluate five datasets commonly used in risk assessment and explainability research, summarized in Table 5. These span healthcare, credit scoring, income prediction, and fraud detection, and vary in feature types, class imbalance, and task complexity.

Maternal Risk Clinical measurements used to predict maternal health outcomes [Ahmed et al., 2020]. The task is to predict whether a pregnancy is associated with elevated health risk. This represents a low-dimensional, fully continuous healthcare task.

German Credit The German Credit dataset [Hofmann, 1994] consists of an anonymized financial and demographic set of features for loan applicants. It combines numerical and categorical attributes describing credit history and employment, frequently used to study the intersection of XAI and algorithmic fairness.

Adult The Adult (Income) dataset [Ding et al., 2021] contains demographic and employment-related attributes used to predict income thresholds. We use a curated subset of features from the Aequitas package [Jesus et al., 2024] to focus on socio-economic indicators common in automated decision-making.

HELOC Credit bureau attributes from FICO [FICO, 2025] used to predict risk for home equity lines of credit. This dataset is a cornerstone for XAI evaluation in the financial sector due to its high-dimensional numerical feature set.

Fraud Risk This is a proprietary, large-scale dataset of card payment transactions collected and downsampled from a real-world financial risk assessment system. Each instance is represented by a heterogeneous set of numerical and categorical features capturing transactional, behavioral, and contextual information available at decision time. The dataset is highly imbalanced, reflecting real operational fraud prevalence.

C.1 Dataset Preprocessing

All datasets undergo a standardized preprocessing pipeline. Categorical features are ordinally encoded to preserve fixed dimensionality across models and explainers. Numerical features are used as provided. Dataset-specific preprocessing is minimal, except for HELOC, which contains missing or invalid sentinel values; these are handled via feature filtering, instance removal, and targeted imputation.

Each dataset is divided into training, validation, and test sets using stratified sampling to preserve class proportions. Validation and test sets are capped at 10% of the dataset size or a maximum of 200 instances each for A/B Testing, whichever is smaller. All splits are generated using a fixed random seed to ensure reproducibility. The final number of features, label prevalence, and split sizes for each dataset are reported in Table 5.

D Extended Experimental Setup

Our audit utilizes a standardized training and optimization pipeline to ensure that variations in explanation quality are not confounded by model performance or optimization noise.

Model Training We train two predictive binary classification risk models per dataset: Logistic Regression and LightGBM [Ke et al., 2017]. Hyperparameters were optimized via *Optuna* [Akiba et al., 2019] with a Tree-structured Parzen Estimator (TPE) sampler and median pruning. We use 3-fold stratified cross-validation and 20 trials per configuration. The optimization objective was set to validation AUC. Final model configurations are summarized in Table 6.

Amortizer Architecture and Training To support the low-latency requirements of the audit, we deploy feed-forward neural network amortizers.

- **Preprocessing:** Numerical features are scaled using robust scaling (IQR-based) with *tanh* saturation to mitigate outliers. Categorical features are passed through learned embedding layers.
- **Architecture:** Layers include fully connected blocks of varying dimensionality, Layer Normalization, LeakyReLU activations, and Dropout (0.1).
- **Optimization:** We employ a two-phase schedule: (1) Adam optimization with a linear warmup, followed by (2) SGD with momentum and cosine decay.
- **Parameters:** Amortizers are trained for 1000 epochs (100 for Adult) with a learning rate of 0.001. We sample four Shapley masks per input during training, with an efficiency loss weight of 0.1.

Implementation of Value Functions We amortize eight distinct value functions (summarized in Appendix A). Sampling-based formulations (Marginal, Conditional, etc.) utilize a background set of

Table 6: Risk model hyperparameter configurations.

Hyperparameter	Maternal	Credit	Adult	HELOC	Fraud
LightGBM					
Num. Leaves	11	82	98	59	6
Learning Rate	0.27	0.10	0.14	0.05	0.10
N-Estimators	258	121	285	187	50
Max. Depth	4	4	6	3	3
Min. Child Samples	25	22	90	100	500
Subsample	0.76	0.65	0.60	0.61	0.50
Bagging Frac.	0.59	0.68	0.96	0.99	0.50
Feature Frac.	0.65	0.86	0.54	0.76	0.50
Colsample/tree	0.72	0.61	0.52	0.89	0.50
Early Stopping	72	132	78	179	20
Logistic Regression					
C (Reg. Strength)	0.13	0.36	0.60	0.59	0.09
Tol	0.00	0.01	0.00	0.00	0.00
Solver	S_1	S_1	S_2	S_1	S_3
Max Iterations	648	237	374	102	785
Class Weight	Bal.	Bal.	Bal.	Bal.	Bal.
Fit Intercept	T	F	T	T	F
Intercept Scaling	0.66	0.82	0.31	1.53	0.19

S_1: newton-cholesky, S_2: sag, S_3: newton-cg, Bal.: Balanced

100 instances specifically generated according to the respective sampling logic. Fixed baseline formulations are implemented using both Zero and Mean reference points to contrast their performance against empirical variants.

Computational Resources All experiments were conducted on a workstation equipped with an Intel Xeon W-2145 CPU (8 cores, 16 threads, 3.70 GHz base clock) and three NVIDIA GeForce RTX 2080 Ti GPUs (11 GB GDDR6 each). While this hardware configuration was used to accelerate training and reduce experimentation time, the computational requirements of our approach are modest; the code can be executed on any relatively recent laptop, albeit with longer runtimes. No specialized hardware or distributed infrastructure is required to reproduce the reported results.

E Human-in-the-Loop A/B Testing Framework Details

To ensure the validity of our audit, we recruited a diverse pool of participants and professional analysts to interact with a standardized risk-assessment interface. By holding the UI, model outputs, and task context constant, we isolated the choice of Shapley formulation as the sole independent variable.

E.1 Participant Profiles

Table 7 summarizes the 37 individuals who contributed to the 3,735 human–AI interactions recorded in this study. The pool includes 5 professional fraud analysts (13.5%) and a high proportion of participants with significant ML and Shapley-specific knowledge, ensuring that the observed "confidence-inflation" was not merely a result of user naivety. Here, **domain knowledge** reflects self-assessed familiarity with the specific dataset being analyzed, presented separately for each dataset included in the study.

Table 7: Summary of participant profiles in the A/B testing experiments.

Total participants		37	
Professional analysts		5 (13.5%)	
	Low	**Moderate**	**High**
ML knowledge	9 (24.3%)	10 (27.0%)	18 (48.6%)
		Yes	**No**
Shapley knowledge		25 (67.6%)	12 (32.4%)
Domain knowledge		**Yes**	**No**
Maternal Risk		3 (12.5%)	21 (87.5%)
UCI German Credit		4 (13.3%)	26 (86.7%)
Adult		3 (12.5%)	21 (87.5%)
FICO HELOC		2 (10.0%)	18 (90.0%)
Fraud Risk		8 (38.1%)	13 (61.9%)

E.2 Experimental Interface and Task Flow

The interface was designed to resemble operational risk analysis tools commonly used in practice. All visual components, layouts, and interaction patterns were held fixed across datasets, models, and Shapley value formulations.

E.2.1 Overview and task flow

The experiment followed a strictly linear workflow to prevent backtracking or cross-case comparison:

1. **Onboarding:** Dataset selection and completion of the analyst profile.
2. **Contextualization:** Review of standardized, dataset-specific task summaries.
3. **Review Loop:** Sequential case analysis and decision-making.

Participants could only advance after submitting a final decision, ensuring that recorded interactions were independent.

E.2.2 Dataset Selection and Analyst Profile

Upon selecting a dataset from a drop-down menu, participants used a panel to self-report their expertise in three areas: domain knowledge for the specific data, general ML familiarity, and prior experience with Shapley-based explanations. These attributes are collected exclusively for post-hoc analysis and stratification and do not influence task assignment, model selection, or explanation configuration. A screenshot of the dataset selection and analyst profile form is shown in Figure 4.

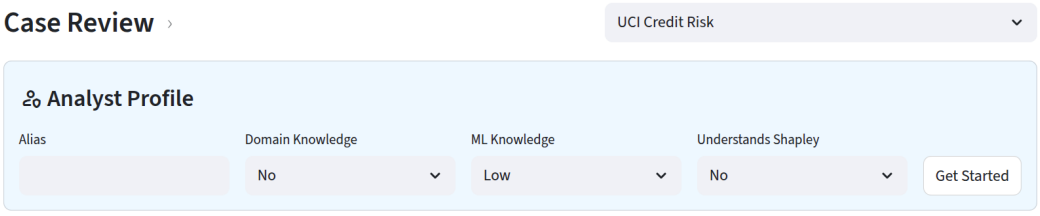

Figure 4: A screenshot showing the initial onboarding form with dropdowns for expertise levels.

E.2.3 Task summary panel

A collapsible summary panel provided a constant reference for the prediction objective, feature definitions (including categorical levels), and the operational meaning of *risk*. This ensured that all participants, regardless of prior expertise, operated under a shared conceptual framework. Figure 5 shows an example of the task summary panel.

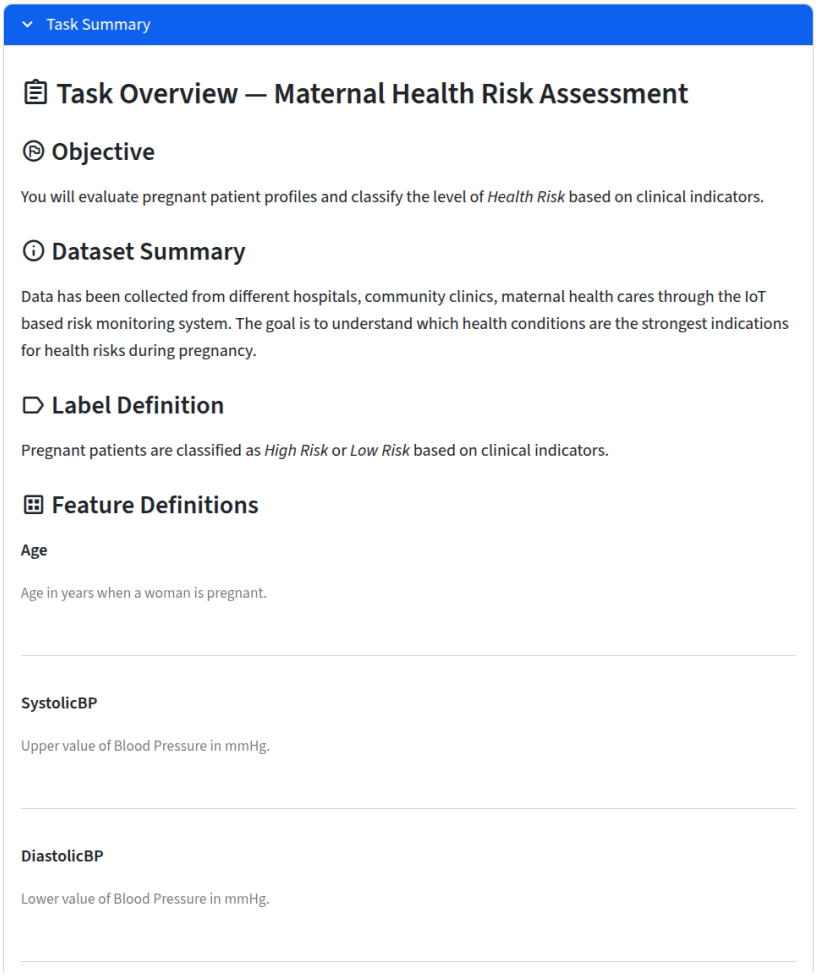

Figure 5: A screenshot of the task summary panel providing analysts with dataset and task context.

E.2.4 Case-review screen

The interface (detailed in Figure 2 of the main text) comprised three primary information blocks:

Model output. The risk score is displayed as a value in $[0, 1]$, together with its empirical percentile within the dataset. A score distribution plot provides additional context by situating the instance relative to the overall population.

Explanation panel. Explanations featured a horizontal bar chart of the four features with the highest absolute Shapley attributions, using consistent color, scale, and ordering conventions across all Shapley formulations. To mirror real-world compliance requirements, the interface displays automatically generated *reason codes*, i.e. short natural-language statements deterministically derived from the Shapley values (e.g., "*Age value of 70 is high*").

Data explorer. A read-only explorer allows participants to inspect the raw feature values of the current instance. Numerical features were accompanied by univariate distributions, while categorical features included prevalence and historical risk statistics. This ensured analysts could verify the data without manipulating it.

E.2.5 Decision and feedback collection

The feedback loop (Figure 6) was designed to capture the core outcome variables of our audit. After reviewing a case, analysts submitted:

- **Decision:** A binary "Risk" or "No Risk" judgment.
- **Confidence:** Self-reported certainty (*Weak, Moderate, Strong*).
- **Clarity:** A subjective assessment of the explanation (*Clear* or *Confusing*).

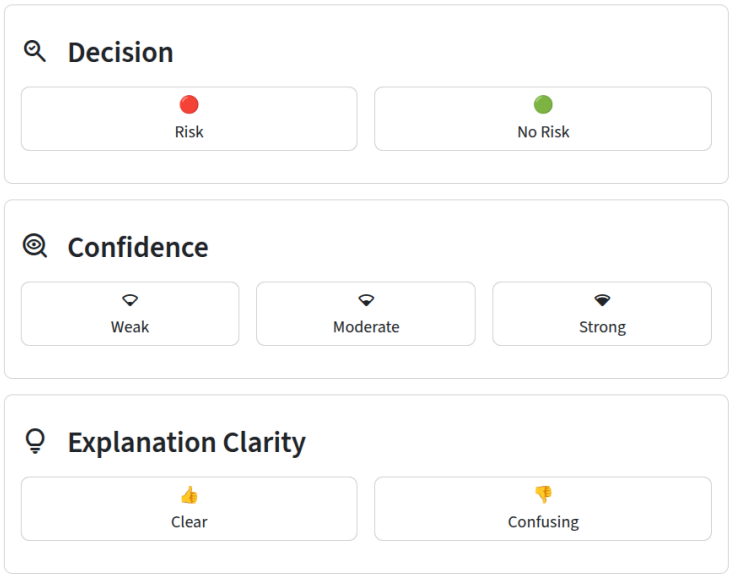

Figure 6: Screenshots of the sequence of questions asked to the analyst for each reviewed case.

No additional controls, filters, or explanation parameters are exposed. This design ensures that observed differences in decision behavior, confidence, and response time arise from explanation content rather than interface affordances or user-driven customization.