Synthetic Datasets as an Ethical Laboratory: Systematically Auditing AI Robustness, Fairness, and Explainability

Alexandre Costa
Artificial Intelligence and Society course FEUP/FCUP
(Dated: January 2, 2026)

This study explores the utility of synthetic datasets as a controlled "ethical laboratory" for the systematic auditing of machine learning models. Unlike real-world data, which often contains entangled and unobservable latent biases, synthetic data allows for the precise modulation of dataset properties, including class imbalance, feature noise, and bias. By subjecting models to these controlled environments, we quantitatively evaluate the models performance in these three critical trustworthy AI pillars: robustness, fairness, and explainability. Our results demonstrate that datasets that own these properties are automatically affected on the AI pillars.

I. INTRODUCTION AND MOTIVATION

The rapid integration of Artificial Intelligence into critical decision-making sectors, such as finance, healthcare, and criminal justice has shifted the focus from pure predictive power to the socio-technical implications of algorithmic behavior. While learning models achieve high accuracy on benchmark datasets, they often inherit and amplify historical biases, lack transparency, and prove fragile when encountering real-world data perturbations.

These flaws in the models can lead to catastrophic consequences in the real world. For example, in a healthcare system, an AI model used for triage might exhibit measurement bias, where medical devices or data collection methods are less accurate for certain demographic groups. If the model is not robust to this noise, it could systematically deprioritize minority patients for life-saving interventions. Similarly, in criminal justice, a lack of explainability in "black-box" risk assessment tools can lead to opaque sentencing decisions that deny individuals their "Right to Explanation" and perpetuate systemic injustice.

The problem is that evaluating these risks in real-world environments is often impossible due to confounded variables and ethical constraints. We cannot "break" a live healthcare system just to see how it handles noise.

By identifying which data parameters affect specific model properties, we are moving toward a more transparent and safer deployment cycle. By treating synthetic data as a sandbox, this project provides a methodology for "Pre-deployment Auditing," allowing developers to understand and mitigate harmful model behaviors before they impact society.

II. BACKGROUND AND RELATED WORK

The evaluation of Trustworthy AI has traditionally been fragmented into isolated domains: robustness, fairness, and explainability. However, recent socio-technical research suggests these pillars are deeply interconnected. This section provides an overview of the existing methodologies and the tools utilized in this study to bridge these gaps.

A. Algorithmic Fairness and Bias Taxonomy

Algorithmic fairness research focuses on detecting and mitigating disparities in model outcomes across protected groups. Others have contributed to this problem by defining a taxonomy of bias. Neri [1] and others distinguish between Historical Bias (systemic inequalities in ground truth), Representation Bias (under-sampling of certain populations), and Measurement Bias (differential noise in feature collection). While tools like *IBM AI Fairness 360* provide metrics to detect these, they often rely on static datasets where these biases are already "baked in" and inseparable. Our work builds on this by using BiasOnDemand[2] to isolate these biases at the generation stage.

B. Model Robustness and Calibration

Robustness is typically evaluated through a model's sensitivity to perturbations. Guo *et al.*[3] introduced the **Expected Calibration Error (ECE)** as a measure of model reliability, arguing that a robust model must not only be accurate but also "honest" about its uncertainty. Furthermore, the concept of randomized smoothing[4] suggests that stable models should maintain consistent predictions within a small distance of the input. We integrate these techniques to measure how synthetic noise affects the "trustworthiness" of a model's probability scores.

C. Explainability and the DoX Framework

Explainability techniques, such as SHAP (SHapley Additive exPlanations) or LIME (local interpretable model-agnostic explanations), have been widely adopted to provide "local" explanations for individual predictions. However, the field has struggled to quantify the *quality* of these explanations. Sovrano and Vitali [5] proposed the

Degree of Explainability (DoX) framework, which moves beyond simple feature importance to evaluate if an explanation can answer "Why" (Clarity), "How" (Distinctiveness), and "What" (Coverage).

D. Synthetic Data as an Audit Tool

Traditional AI auditing often relies on "black-box" testing of live APIs or static, real-world datasets, which are frequently constrained by confounding variables that obscure the true drivers of algorithmic failure. To overcome these limitations, recent research has shifted toward **Experimental AI Auditing**, a framework that leverages synthetic data to create a controlled "laboratory environment" where specific variables can be isolated and systematically tested. For instance, recent work by[6] demonstrates the effectiveness of this approach by using the BiasOnDemand[2] package to uncovering fairness through data complexity as an early indicator.

III. METHODOLOGY

The experimental framework of this study is built upon a modular pipeline that treats synthetic data generation as a controlled laboratory environment. This section details the data schema, the parametric definitions of the audit engine, and the specific scenarios used to stress-test model behavior.

A. Synthetic Data Generation

To systematically investigate the "breaking points" of model trustworthiness, ten distinct suites of experimental datasets were generated II. Each suite consists of multiple datasets (with 10000 or 15000 samples) where a single parameter λ is varied across a standardized range $[0.0, 1.0]$ (0.1 variation), which led to the generation of 108 datasets. The generation follows a structured causal model comprising four primary variables and their noisy proxies, as detailed in Table I. This isolation of variables allows us to map the precise relationship between data defects and model degradation.

B. Fairness Metrics

To evaluate model fairness, an **Overall Fairness Score** ($\mathcal{F}$) was created based on multiple metrics III. This composite score is defined as:

$$\mathcal{F} = 1.5 \times \frac{1}{N} \sum_{i=1}^{N} w_i \cdot |M_i - \text{Ideal}_i| \tag{1}$$

where M_i is the metric value and w_i is its weight. A score of 0.0 represents perfect fairness.

TABLE I: Causal Variables in the Synthetic Laboratory

Var	Type	Description
A	Binary	**Sensitive Attribute:** $A = 0$ (Reference) vs. $A = 1$ (Protected/Subject to bias).
R	Cont.	**Legitimate Feature:** Causally influenced by A; represents variables like income or experience.
Q	Cont.	**Performance Feature:** Influenced by A and R; represents credit scores or ratings.
Y	Binary	**Target Outcome:** The prediction target (e.g., loan approval) based on Q, R, and A.
P	Cont.	**Proxy:** A noisy version of R, introduced during measurement bias.

C. Explainability and the DoX Framework

The quality of model explanations is quantified using the **Degree of Explainability (DoX)** framework (the framework is not the original used in Sovrano and Vitali[5], but a calculated approximation). Unlike simple feature importance, DoX evaluates the utility of an explanation by answering three human-centric questions: *Why* (**Clarity**), *How* (**Distinctiveness**), and *What* (**Coverage**). These were aggregated into a composite score IV:

$$DoX = (0.40 \cdot \mathcal{C}) + (0.35 \cdot \mathcal{D}) + (0.25 \cdot \mathcal{V}) \tag{2}$$

where $\mathcal{C}$, $\mathcal{D}$, and $\mathcal{V}$ represent Clarity, Distinctiveness, and Coverage respectively.

This framework allows the "Ethical Laboratory" to detect if data bias leads to "fuzzy" logic, where the model relies on a high-entropy distribution of weak signals rather than clear, distinct predictors.

D. Model Robustness Metrics

Robustness is defined as the model's ability to maintain performance and consistency under uncertainty. We evaluate this through three dimensions: **Stability**, **Resilience**, and **Reliability** V. The composite **Overall Robustness Score** ($\mathcal{R}$) is defined as:

$$\mathcal{R} = (0.40 \cdot S) + (0.30 \cdot R_e) + (0.30 \cdot R_l) \tag{3}$$

Expected Calibration Error (ECE) is calculated by grouping predictions into B confidence bins and measuring the weighted gap between accuracy (acc) and con-

TABLE II: Detailed Taxonomy of Synthetic Bias Datasets

Category	Dataset Series	Path	Description
Historical Bias	1. Bias on R	$A \to R$	Structural influence of A on resources (e.g., income). Q and Y inherit disparity even if learning is unbiased.
	2. Bias on Q	$A \to Q$	Environmental factors correlated with A affect auxiliary variable Q, propagating milder bias to the target Y.
	3. Bias on Y	$A \to Y$	Explicit discrimination in historical labels. The target encodes group-dependent decisions directly.
	4. Interaction Proxy	$A \to P_R \to Q \to Y$	Proxy P_R leaks sensitive info through interactions, enabling the exploitation of spurious correlations.
Measurement Bias	5. Bias on R	$R \to P_R \leftarrow A$	True R is hidden; proxy P_R is contaminated by A, corrupting the semantic meaning of explanatory variables.
	6. Linear Y	$Y \to P_Y \leftarrow A$	Y is replaced by proxy P_Y with proportional mislabeling based on A. Reflects biased evaluation.
	7. Non-Linear Y	$Y \to P_Y \leftarrow A$	Complex, conditional labeling errors in P_Y. Robustness/explainability suffer more than selection fairness.
Imbalance & Noise	8. Representation	$P(A) \neq unif.$	Altered sampling distribution makes one group underrepresented, causing severe fairness degradation (DI, EOD).
	9. Undersampling	-	Uniform data reduction. Causal paths remain intact; metrics for fairness and robustness stay stable.
	10. Label Noise	$Y \to \tilde{Y}$	Stochastic perturbation of target. Weakens input-output link without introducing group-specific mechanisms.

TABLE III: Fairness Metric Definitions and Weights

Name	Formula	Weight	Description
DI	$\frac{P(\hat{Y}=1\|A=1)}{P(\hat{Y}=1\|A=0)}$	1.00	Legal standard for adverse impact.
SP	$\Delta P(\hat{Y}=1\|A)$	0.90	Measure absolute outcome inequality.
EOD	$\frac{\|\Delta TPR\|+\|\Delta FPR\|}{2}$	0.90	Balance absolute error across groups.
AOD	$\frac{\Delta FPR+\Delta TPR}{2}$	0.70	Identify systematic directional bias.
PE	ΔFPR	0.60	Avoid disproportionate false penalties.
EOR_T	TPR_1/TPR_0	0.50	Ensure equal recall for qualified candidates.
ACC	Acc_1/Acc_0	0.30	Verify overall performance consistency.
EOR_F	FNR_1/FNR_0	0.20	Ensure missed opportunities are balanced.

fidence ($conf$):

$$ECE = \sum_{i=1}^{B} \frac{n_i}{N} |acc(i) - conf(i)| \tag{4}$$

This framework detects models that are "brittle",

TABLE IV: Explainability Metrics

Dimension	Weight	Metric Definition and Logic
Clarity ($\mathcal{C}$)	0.40	**Concentration:** Measured via Shannon Entropy and Feature Dominance. High scores indicate a few clear drivers standout from background noise.
Distinctness ($\mathcal{D}$)	0.35	**Separability:** Uses the Coefficient of Variation ($CV = \sigma/\mu$) and Range. High scores ensure a clear hierarchy of importance between features.
Coverage ($\mathcal{V}$)	0.25	**Completeness:** The ratio of Top-K importance to total importance. Uses a sigmoid scaling to penalize explanations where Top-K features account for $< 80\%$ of the decision. (K=1 for this experiment)

TABLE V: Robustness Metric Definitions and Weights

Name	Formula	Weight	Description
Stability (S)	$\frac{1}{N}\sum \mathbb{I}(\hat{y}_i = \hat{y}'_i)$	0.40	**Flicker Test:** Ensure consistent predictions under input jitter ($\sigma = 0.05$).
Resilience (R_e)	$\min(\frac{Acc_{corr}}{Acc_{clean}}, 1.0)$	0.30	**Stress Test:** Measure accuracy retention under significant noise ($\sigma = 0.1$).
Reliability (R_l)	$1.0 - \text{ECE}$	0.30	**Honesty Test:** Ensure confidence scores match empirical accuracy.

those that achieve high accuracy on paper but provide inconsistent or overconfident predictions when subjected to real-world data perturbations.

IV. RESULTS AND DISCUSSION

For each dataset group (from 0.1 to 1 bias on each factor) it was performed a analysis for each of the AI pillars. The project contains graphical comparison between each metric inside each pillar.

A. Fairness Results

Figure 1 reports the evolution of Disparate Impact (DI), Equalized Odds Difference (EOD), and the Overall Fairness Score across different bias conditions and intensities.

a. Disparate Impact (DI). For most bias conditions, DI remains close to its ideal value, indicating that group-wise selection rates are largely preserved. This is expected whenever the causal influence of the sensitive attribute A on the decision Y is either weak or fully me-

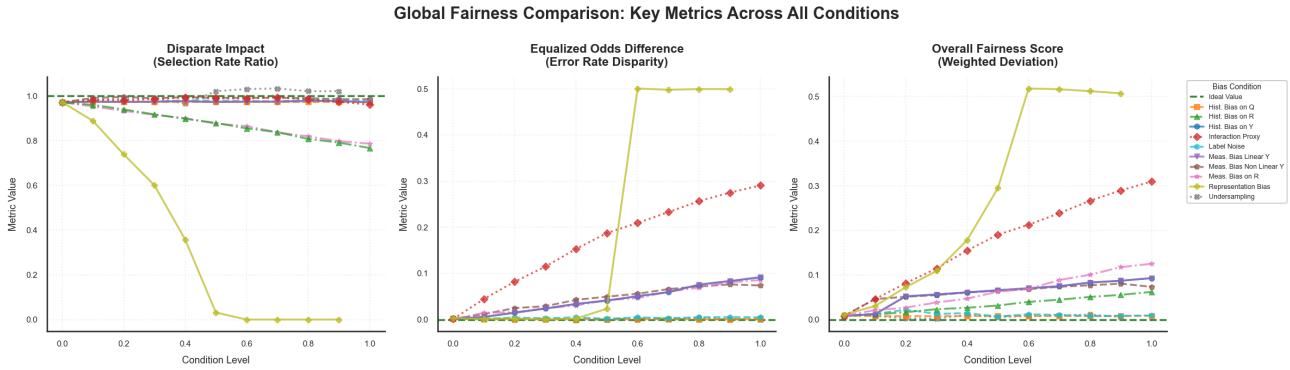

FIG. 1: Comprehensive analysis of the 10 experimental suites: Comparative impact of bias intensity on the three pillars of Trustworthy AI (Fairness, Robustness, and Explainability).

diated by legitimate variables that remain correctly observed. However, both historical bias on R and measurement bias on R lead to a clear degradation in DI. In these settings, A affects Y through the primary causal pathway $A \to R \to Q \to Y$, and distortions applied directly to R, either by altering its distribution across groups or by replacing it with a biased proxy, amplify group-dependent differences in downstream variables. As a result, the model learns systematically different decision thresholds for different values of A, leading to unequal selection rates.

The most severe DI violation occurs under representation bias. Although the structural relationships among variables remain unchanged, under-sampling one group alters the empirical distribution of A, impairing the model's ability to estimate group-conditional decision boundaries. This demonstrates that even in the absence of explicit causal bias, structural data imbalance alone can invalidate selection parity, highlighting DI's sensitivity to representation effects.

b. Equalized Odds Difference (EOD). EOD remains low at small bias intensities but increases sharply when representation bias exceeds moderate levels. This behavior reflects a transition in which insufficient group representation leads to unreliable estimates of error rates, producing large disparities in both false positives and false negatives across groups. Unlike DI, which depends on marginal selection rates, EOD is sensitive to conditional performance, making it particularly vulnerable to estimation errors under data imbalance.

Interaction proxy bias exhibits a strong linear increase in EOD. In this case, the proxy variable P_R, which is contaminated by A, interacts with other features used to predict Y. This creates indirect pathways through which sensitive information influences predictions, enabling the model to exploit false correlations. Consequently, group-specific error rates diverge even when overall selection rates remain relatively stable.

In a lesser extent, measurement bias on R also leads to increasing EOD, as it distorts the primary causal pathway $A \to R \to Q \to Y$. By replacing the true resource variable with a biased proxy, group-dependent differences in R are amplified and propagated downstream, resulting in unequal error rates. Also, historical and measurement bias on Y (both linear and non-linear) increase EOD, these biases primarily distort the target variable rather than restructuring the causal influence of A on Y, leading to noisier but less systematically group-conditioned prediction errors.

c. Overall Fairness Score. The Overall Fairness Score integrates these effects and clearly identifies representation bias as the most harmful condition. By simultaneously degrading both selection parity and error-rate equality, representation bias heavily impacts the highest-weighted fairness metrics. Interaction proxy bias follows, driven by the compounding effect of proxy leakage along causal pathways that allow sensitive information to influence predictions indirectly.

Measurement biases rank next, as they corrupt semantically meaningful variables, particularly R, thereby amplifying group-dependent effects without altering the underlying causal graph. Historical bias produces slightly lower but persistent fairness degradation due to its indirect propagation through legitimate features. In contrast, historical bias on Q, undersampling, and label noise have minimal impact on the overall score. Although Q lies on the causal path from A to Y, bias affecting Q does not distort or replace the primary mediator R, and therefore does not substantially amplify group-conditioned differences in downstream predictions. Undersampling and label noise primarily introduce stochastic variability or reduce data efficiency without reinforcing sensitive-attribute-dependent mechanisms, resulting in limited effects on aggregate fairness metrics.

Overall, these results confirm that fairness violations are most severe when bias aligns with or creates causal pathways from sensitive attributes to the decision variable, rather than merely introducing noise or reducing data availability.

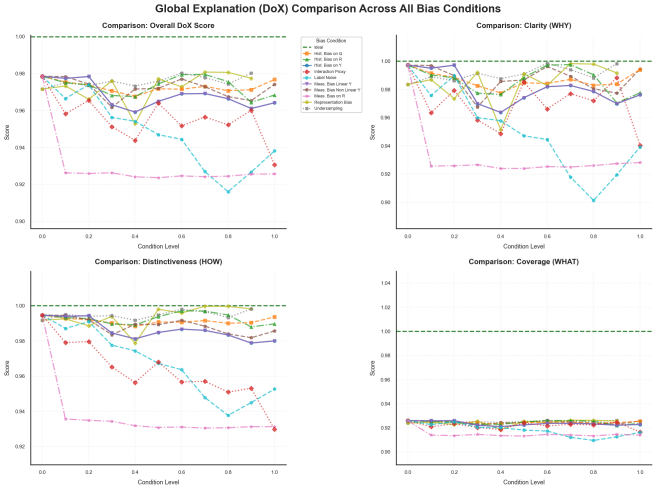

FIG. 2: Comprehensive analysis of the 10 experimental suites: Comparative impact of bias intensity on the three pillars of Trustworthy AI (Fairness, Robustness, and Explainability).

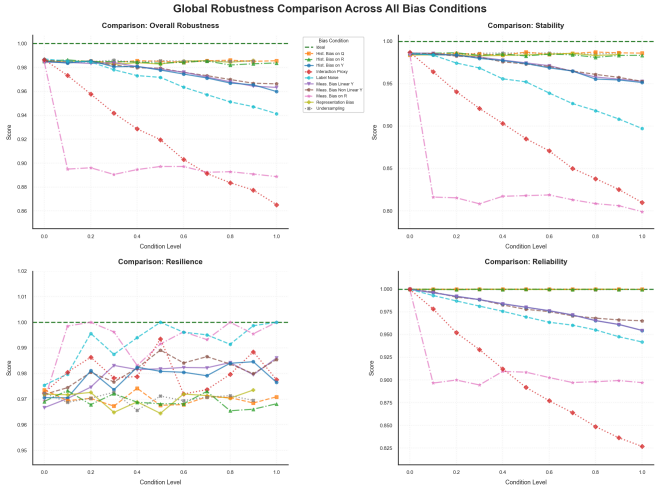

FIG. 3: Comprehensive analysis of the 10 experimental suites: Comparative impact of bias intensity on the three pillars of Trustworthy AI (Fairness, Robustness, and Explainability).

B. Explainability Results

Figure 2 reports the behavior of the overall Degree of Explainability (DoX) score and its components—Clarity (WHY), Distinctiveness (HOW), and Coverage (WHAT)—across all bias conditions and intensities.

Clarity and Distinctiveness exhibit closely aligned trends across bias conditions, indicating that factors reducing the semantic coherence of explanations simultaneously affect both the interpretability of model reasoning and the separability of explanations across instances. In contrast, Coverage remains largely stable, suggesting that the model continues to draw on a similar set of features even when the quality of explanations degrades. The few conditions that reduce Coverage are those that also dominate the other DoX components, pointing to global disruptions rather than metric-specific effects.

Measurement bias on R produces an immediate decline in explainability. Although the same causal pathway is preserved, the observed feature no longer faithfully represents the underlying resource, causing explanations to rely on distorted signals whose meaning varies across groups. This loss of semantic consistency reduces both clarity and distinctiveness, even at low bias levels.

Undersampling and representation bias have limited impact on explainability. While representation bias affects group proportions, it leaves the functional dependencies between variables unchanged within the observed data, allowing explanation structure to remain stable. Similarly, undersampling reduces sample size without altering feature semantics, preserving explanation quality.

In contrast, interaction proxy bias and label noise substantially degrade explainability. Proxy interactions elevate false feature importance, leading to explanations that are less interpretable and harder to differentiate across instances. Label noise weakens the consistency between features and outcomes, increasing ambiguity in the model's decision logic. Overall, explainability is most affected by biases that alter feature meaning or target consistency, rather than those that primarily affect data availability or group representation.

C. Robustness Results

Figure 3 presents the behavior of the Overall Robustness score and its components: Stability, Reliability, and Resilience under the usual circumstances.

Stability and Reliability exhibit nearly identical trends across all bias conditions, indicating that factors degrading prediction consistency under perturbations also reduce the dependability of model outputs. Measurement bias on R, interaction proxy bias, and label noise consistently lead to the strongest degradation. These conditions alter the effective signal used by the model, either by bending the feature space (measurement bias and proxy interactions) or by weakening the alignment between inputs and outputs (label noise), resulting in unstable and less reliable predictions.

Resilience displays a markedly different pattern. Under label noise and measurement bias on R, resilience reaches a perfect score at the highest bias levels, indicating that the model successfully adapts to systematic distortions encountered during training. Rather than resisting these perturbations, the model internalizes them, recovering performance within the biased distribution. This explains the divergence between resilience and the other robustness components, as resilience captures adaptation to persistent bias rather than invariance to

change.

The Overall Robustness score reflects this separation of effects. Historical bias on R and Q has minimal impact, as these biases preserve stable functional relationships between observed features and the target. In contrast, measurement and historical bias on Y moderately degrade robustness by directly corrupting the supervision signal. Overall, robustness is most compromised by biases that disrupt feature semantics or target consistency, while biases that preserve functional structure, even if ethically problematic, employ limited influence on robustness metrics.

D. Limitations and Ethical Reflections

While the proposed framework enables a systematic analysis of fairness, explainability, and robustness, several limitations must be considered when interpreting the results. First, the aggregation of metrics within each pillar relies on predefined weighting schemes. Although these weights are motivated by legal, ethical, and practical considerations, they inevitably reflect normative design choices. Alternative weightings or stakeholder-specific priorities may lead to different trade-offs and conclusions.

Second, the experimental setting assumes a simplified causal structure with a single binary sensitive attribute and clearly separated legitimate, proxy, and target variables. Real-world decision systems often involve multiple protected attributes, intersectional effects, temporal dependencies, and feedback loops that are not captured by this model. Moreover, the use of controlled, synthetic bias injection, while valuable for isolating effects, may not fully represent the complexity and entanglement of biases observed in deployed socio-technical systems.

A further limitation arises in the interpretation of robustness metrics. In particular, resilience reflects the model's ability to adapt to biased or noisy data distributions rather than its resistance to them. As observed in the results, high resilience can coexist with degraded fairness and explainability, raising ethical concerns about systems that perform consistently yet reinforce inequitable or opaque decision-making.

From an ethical standpoint, the results underline that trustworthiness cannot be inferred from any single pillar in isolation. Fairness failures are most severe when bias operates along causal pathways linked to protected attributes, explainability degrades when semantic meaning is distorted, and robustness may remain high even under ethically problematic conditions. These observations highlight the risk of partial evaluations and motivate the need for integrated, causally informed assessment frameworks.

V. CONCLUSIONS AND FUTURE WORK

This work presents a unified experimental framework for analyzing trustworthiness in machine learning systems through the combined lenses of fairness, explainability, and robustness. By systematically introducing different bias mechanisms aligned with a causal data-generating process, we demonstrate that each pillar responds to distinct types of bias and captures complementary failure modes. Representation and interaction-based biases emerge as the most harmful for fairness, measurement bias on legitimate features critically undermines explainability, and robustness metrics reveal that stable or adaptive behavior does not necessarily imply ethical acceptability.

Future work will focus on extending this framework to more complex and realistic settings, including multiple and intersecting sensitive attributes, temporal dynamics, and feedback effects between model decisions and data generation. Incorporating real-world datasets and domain-specific constraints will further validate the proposed approach. Additionally, exploring adaptive or stakeholder-driven weighting strategies and integrating counterfactual or causal fairness notions represent promising directions toward more context-aware and ethically grounded trustworthiness assessments. Code is available at the repository[7].

---

[1] F. Neri, *Bias and Fairness in Machine Learning*, Springer Briefs in Applied Sciences and Technology (Springer Nature, 2020).

[2] V. Contributors, "Biasondemand: A causal framework for synthetic bias generation," https://github.com/bias-on-demand (2024), accessed: 2025-12-09.

[3] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, Proceedings of the 34th International Conference on Machine Learning (ICML) **70**, 1321 (2017).

[4] J. Cohen, E. Rosenfeld, and Z. Kolter, in *Proceedings of the 36th International Conference on Machine Learning (ICML)* (2019) pp. 1310–1320.

[5] F. Sovrano and F. Vitali, IEEE Transactions on Knowledge and Data Engineering (2023), 10.1109/TKDE.2023.3242055.

[6] J. C. Juliett Suárez Ferreira, Marija Slavkovik, arXiv preprint arXiv:2504.05923 (2025).

[7] A. Costa, "Synthetic_datasets_as_an_ethical_laboratory," https://github.com/alexandre2004costa/Synthetic_Datasets_as_an_Ethical_Laboratory (2025), accessed: 2025-12-09.

Appendix

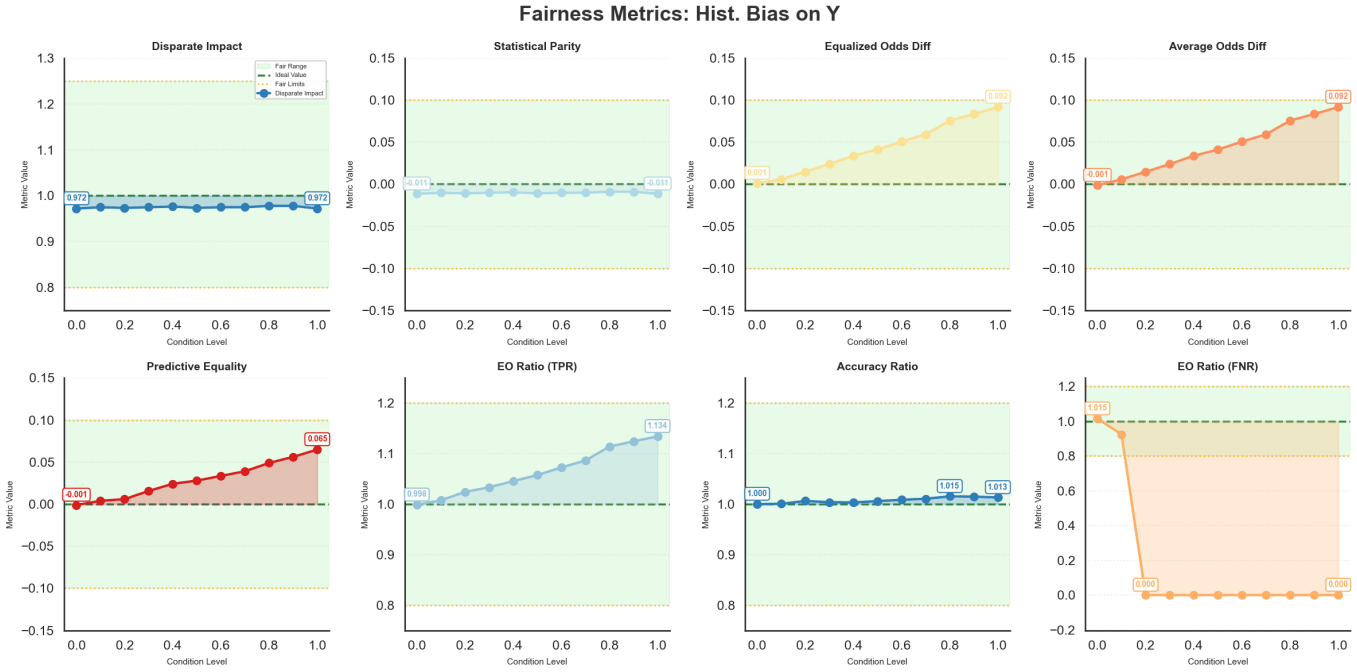
Fairness Metrics: Hist. Bias on Y
Disparate Impact
Statistical Parity
Equalized Odds Diff
Average Odds Diff
Predictive Equality
EO Ratio (TPR)
Accuracy Ratio
EO Ratio (FNR)
Metric Value
Condition Level
Fair Range
Ideal Value
Fair Limits
Disparate Impact
0.972
0.972
-0.011
-0.011
0.092
-0.001
0.092
-0.001
0.065
-0.001
1.134
0.998
1.000
1.015
1.013
1.015
0.000
0.000

FIG. 4

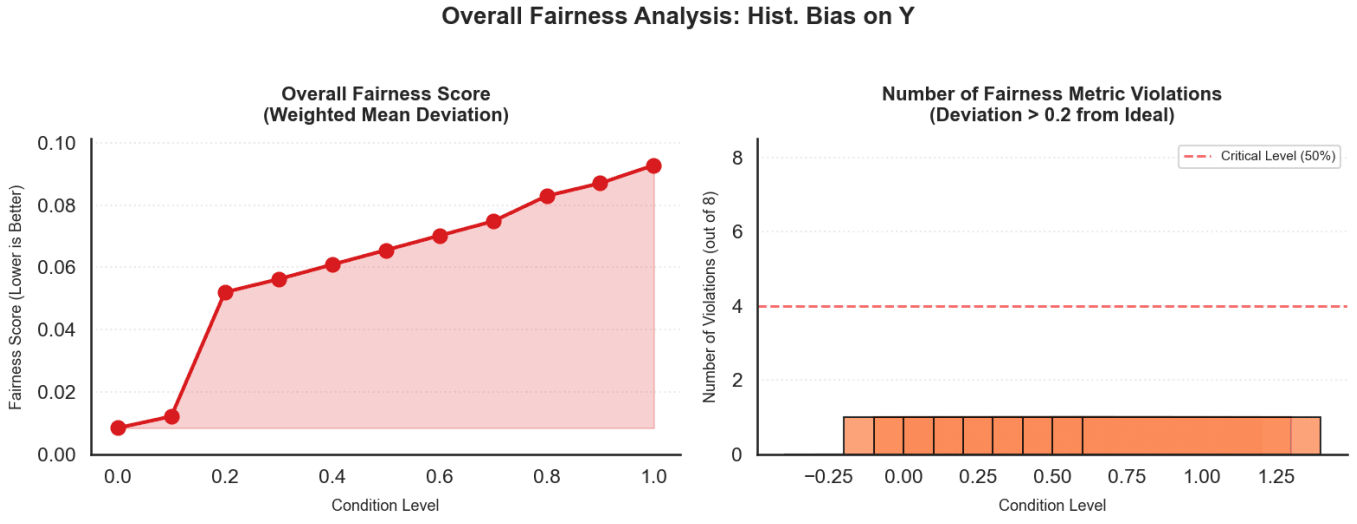
Overall Fairness Analysis: Hist. Bias on Y
Overall Fairness Score
(Weighted Mean Deviation)
Fairness Score (Lower is Better)
Condition Level
Number of Fairness Metric Violations
(Deviation > 0.2 from Ideal)
Number of Violations (out of 8)
Critical Level (50%)
Condition Level

FIG. 5

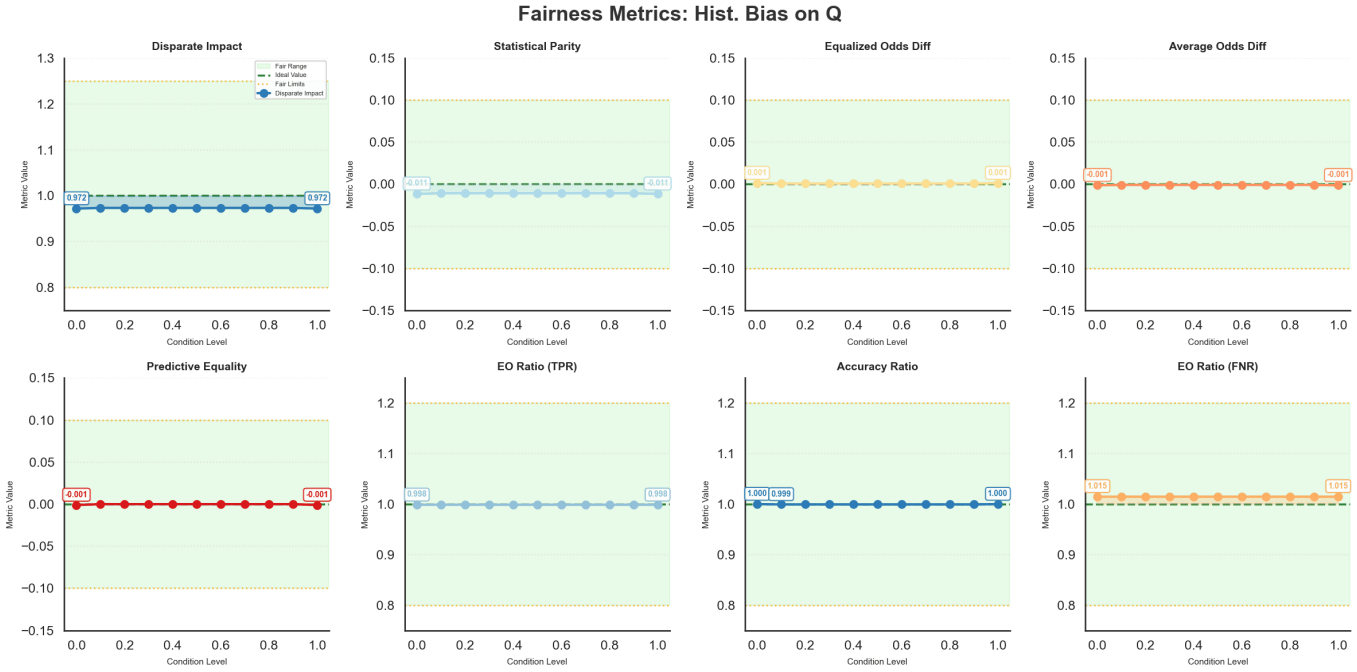
Fairness Metrics: Hist. Bias on Q
Disparate Impact
Statistical Parity
Equalized Odds Diff
Average Odds Diff
Predictive Equality
EO Ratio (TPR)
Accuracy Ratio
EO Ratio (FNR)
Metric Value
Condition Level
Fair Range
Ideal Value
Fair Limits
Disparate Impact
0.972
-0.011
0.001
-0.001
0.998
1.000
0.999
1.015

FIG. 6

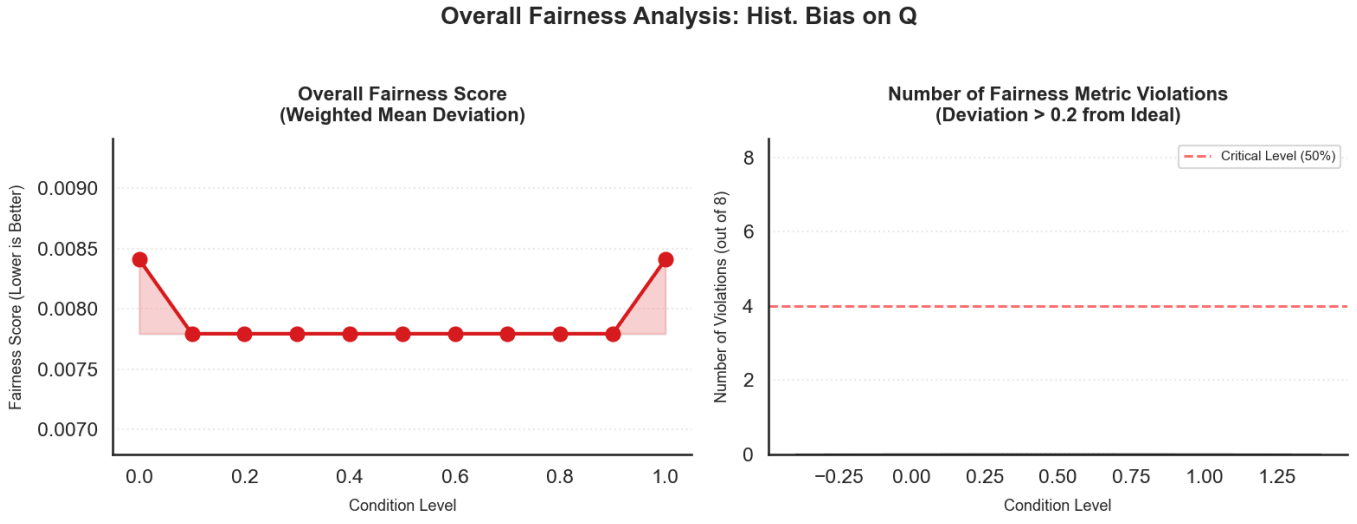
Overall Fairness Analysis: Hist. Bias on Q
Overall Fairness Score
(Weighted Mean Deviation)
Fairness Score (Lower is Better)
Condition Level
Number of Fairness Metric Violations
(Deviation > 0.2 from Ideal)
Number of Violations (out of 8)
Critical Level (50%)

FIG. 7

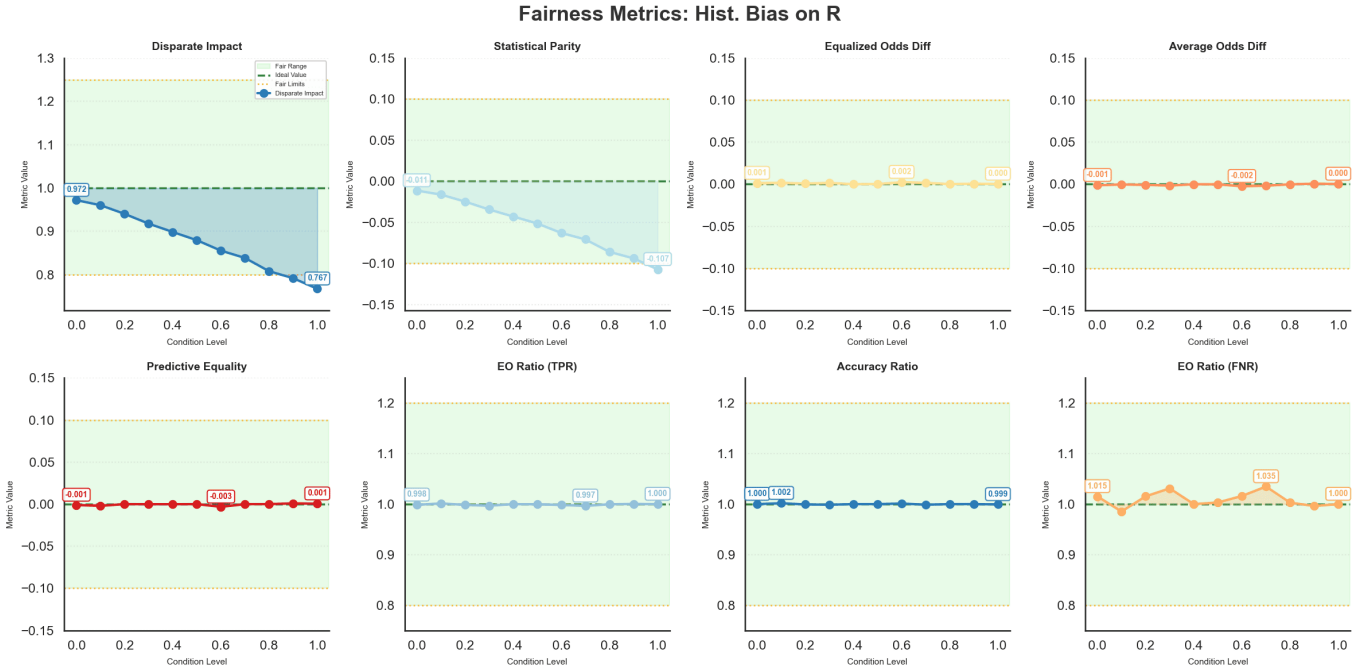
Fairness Metrics: Hist. Bias on R
Disparate Impact
Statistical Parity
Equalized Odds Diff
Average Odds Diff
Predictive Equality
EO Ratio (TPR)
Accuracy Ratio
EO Ratio (FNR)
Metric Value
Condition Level
Fair Range
Ideal Value
Fair Limits
Disparate Impact

FIG. 8

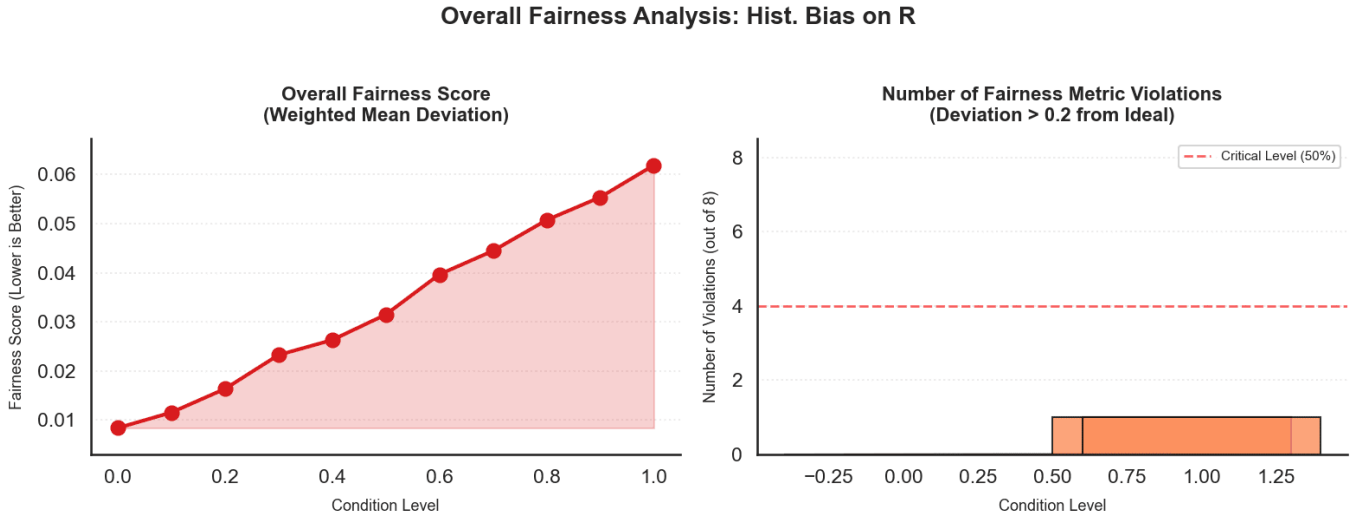
Overall Fairness Analysis: Hist. Bias on R
Overall Fairness Score
(Weighted Mean Deviation)
Fairness Score (Lower is Better)
Condition Level
Number of Fairness Metric Violations
(Deviation > 0.2 from Ideal)
Number of Violations (out of 8)
Critical Level (50%)
Condition Level

FIG. 9

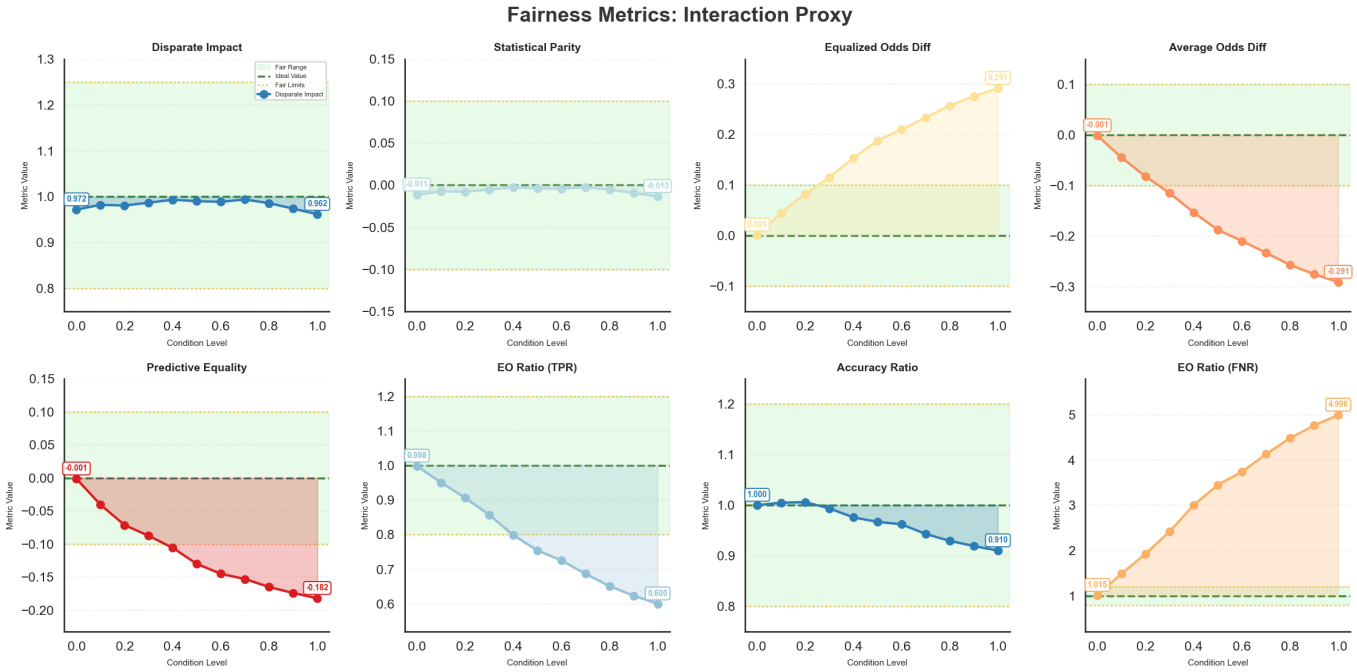
Fairness Metrics: Interaction Proxy
Disparate Impact
Statistical Parity
Equalized Odds Diff
Average Odds Diff
Predictive Equality
EO Ratio (TPR)
Accuracy Ratio
EO Ratio (FNR)
Metric Value
Condition Level
Fair Range
Ideal Value
Fair Limits
Disparate Impact

FIG. 10

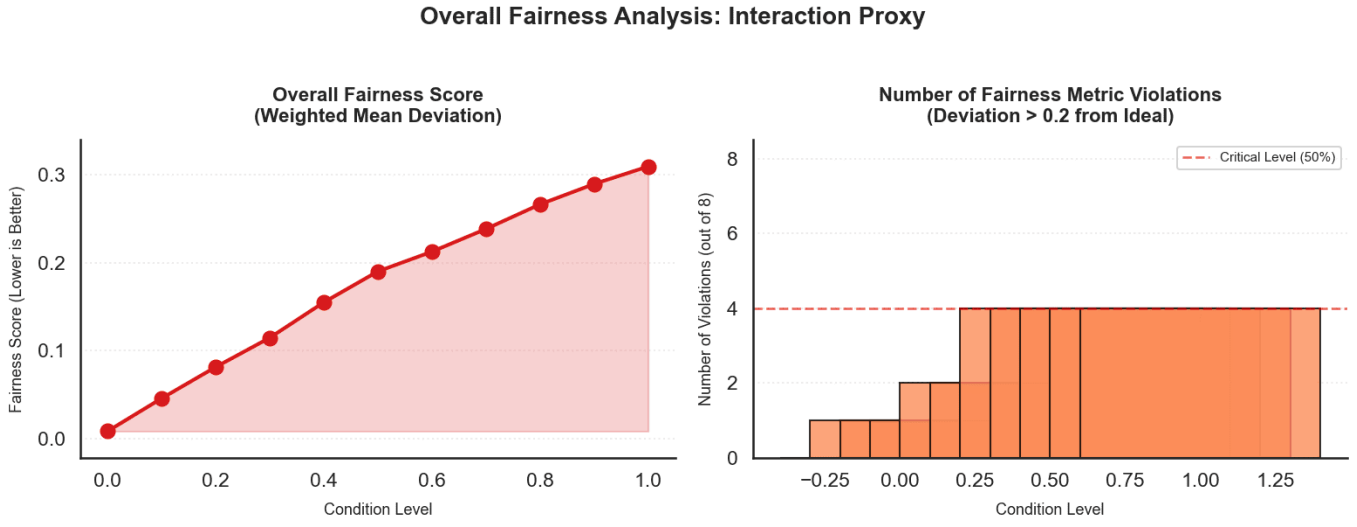
Overall Fairness Analysis: Interaction Proxy
Overall Fairness Score
(Weighted Mean Deviation)
Fairness Score (Lower is Better)
Condition Level
Number of Fairness Metric Violations
(Deviation > 0.2 from Ideal)
Number of Violations (out of 8)
Critical Level (50%)
Condition Level

FIG. 11

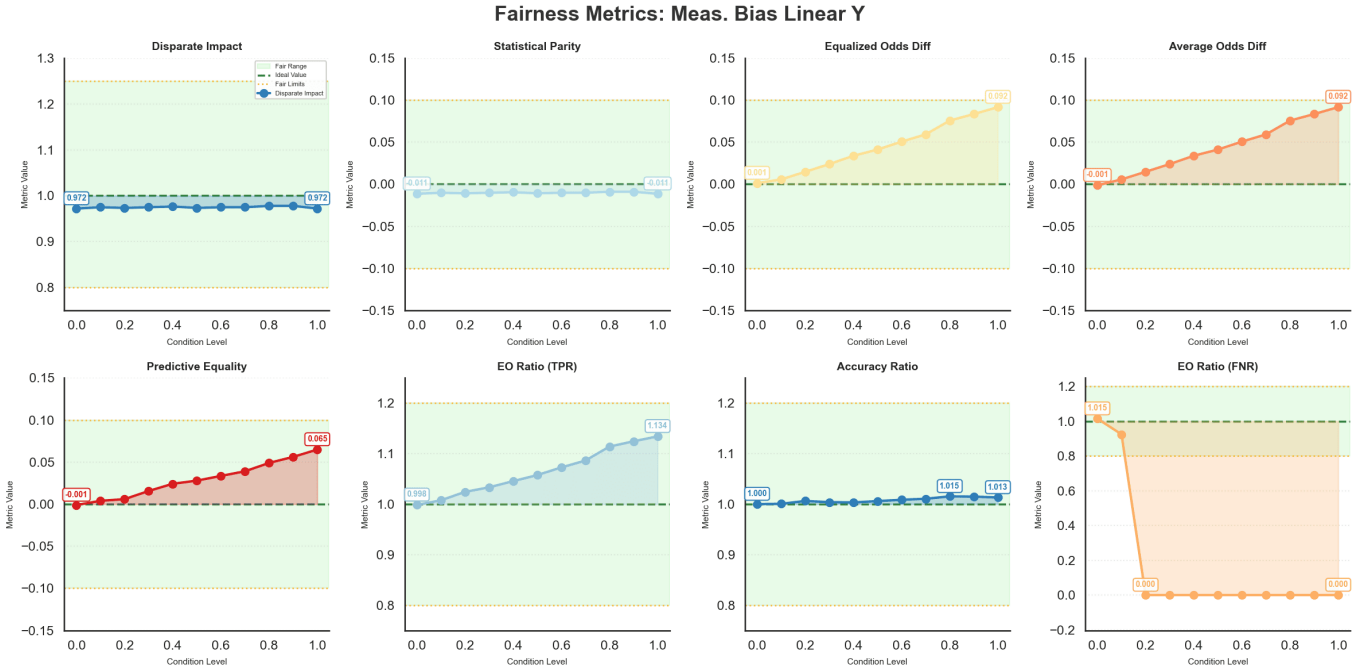
Fairness Metrics: Meas. Bias Linear Y
Disparate Impact
Statistical Parity
Equalized Odds Diff
Average Odds Diff
Predictive Equality
EO Ratio (TPR)
Accuracy Ratio
EO Ratio (FNR)
Metric Value
Condition Level
Fair Range
Ideal Value
Fair Limits
Disparate Impact
0.972
0.972
-0.011
-0.011
0.091
0.092
-0.001
0.092
-0.001
0.065
0.998
1.134
1.000
1.015
1.013
1.015
0.000
0.000

FIG. 12

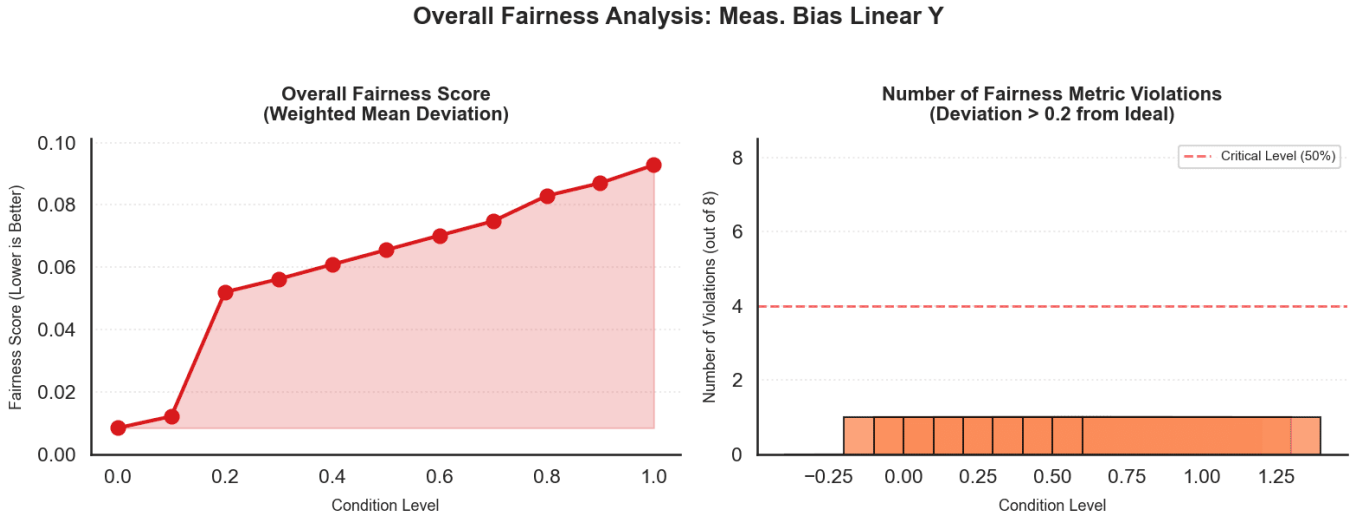
Overall Fairness Analysis: Meas. Bias Linear Y
Overall Fairness Score
(Weighted Mean Deviation)
Fairness Score (Lower is Better)
Condition Level
Number of Fairness Metric Violations
(Deviation > 0.2 from Ideal)
Number of Violations (out of 8)
Critical Level (50%)
Condition Level

FIG. 13

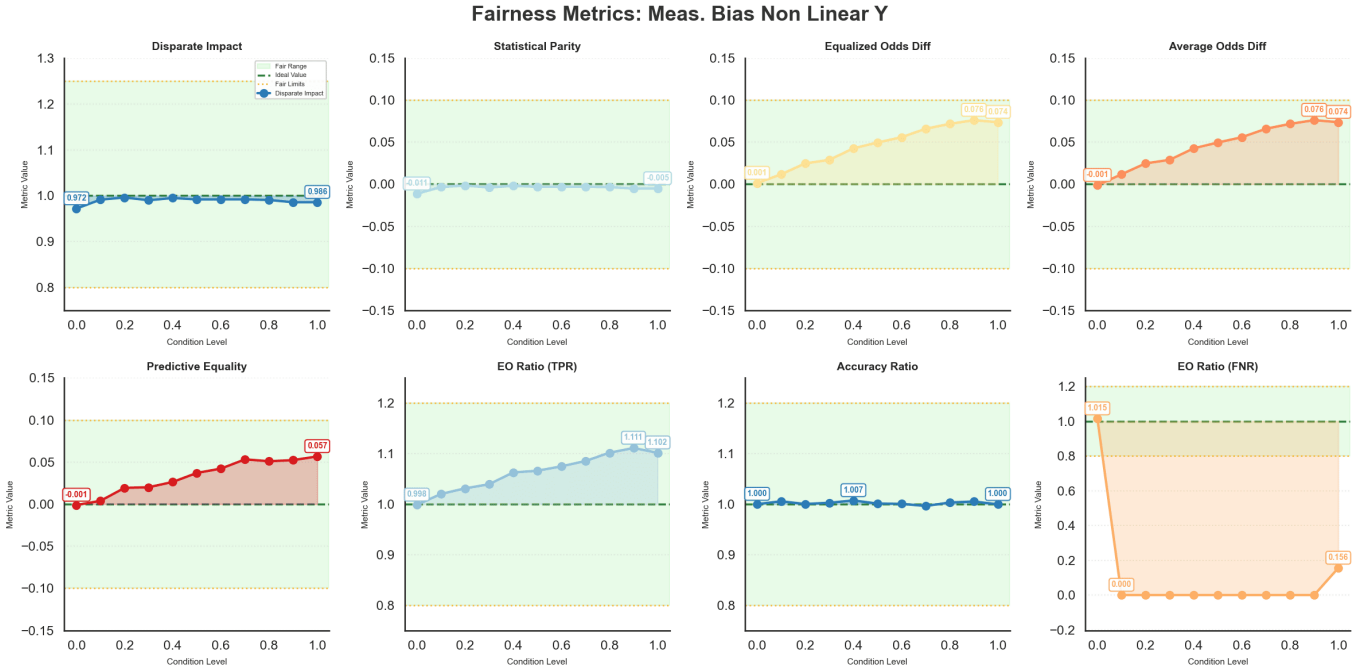
Fairness Metrics: Meas. Bias Non Linear Y
Disparate Impact
Statistical Parity
Equalized Odds Diff
Average Odds Diff
Predictive Equality
EO Ratio (TPR)
Accuracy Ratio
EO Ratio (FNR)
Metric Value
Condition Level
Fair Range
Ideal Value
Fair Limits
Disparate Impact

FIG. 14

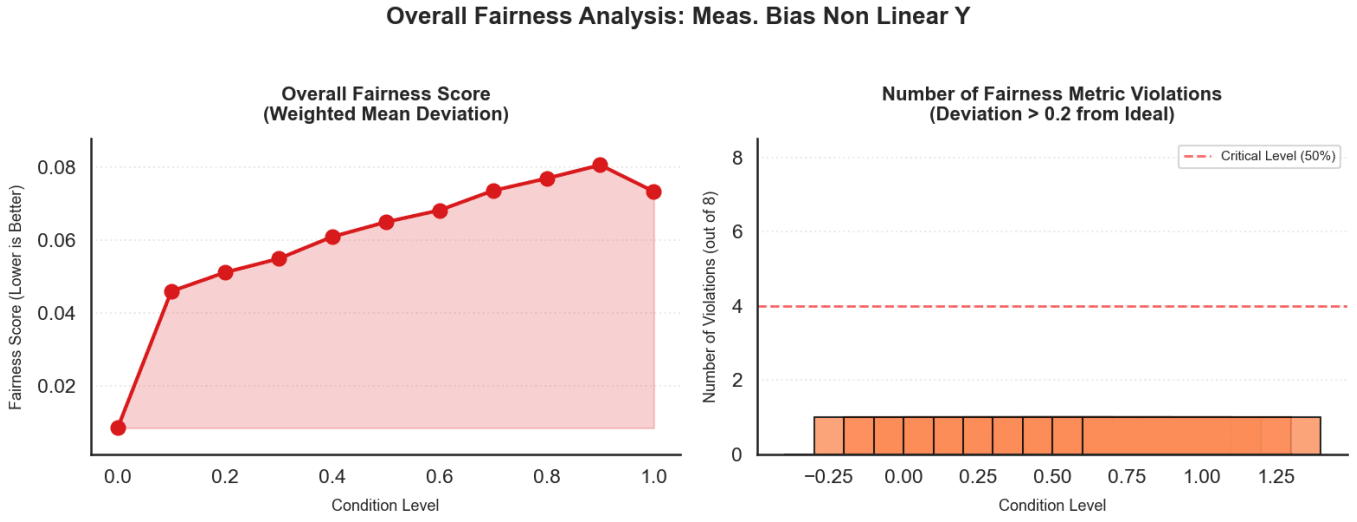
Overall Fairness Analysis: Meas. Bias Non Linear Y
Overall Fairness Score
(Weighted Mean Deviation)
Fairness Score (Lower is Better)
Condition Level
Number of Fairness Metric Violations
(Deviation > 0.2 from Ideal)
Number of Violations (out of 8)
Critical Level (50%)
Condition Level

FIG. 15

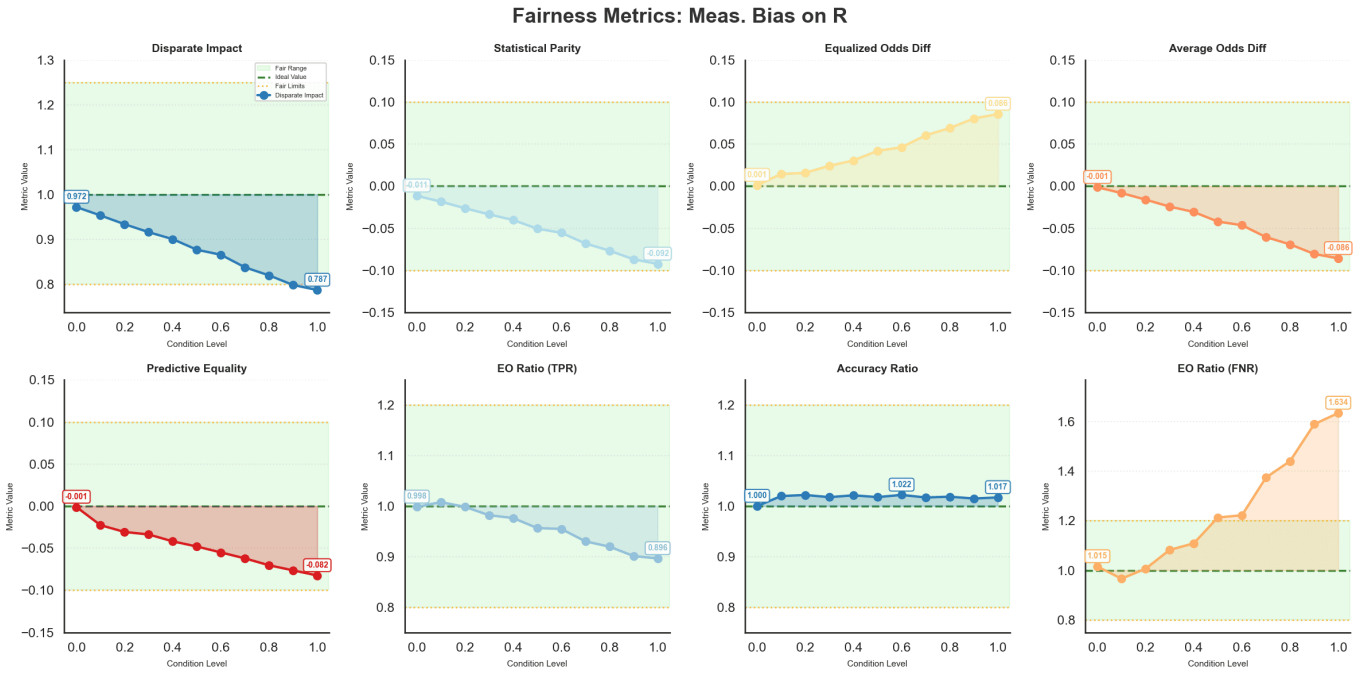
Fairness Metrics: Meas. Bias on R
Disparate Impact
Statistical Parity
Equalized Odds Diff
Average Odds Diff
Predictive Equality
EO Ratio (TPR)
Accuracy Ratio
EO Ratio (FNR)
Metric Value
Condition Level
Fair Range
Ideal Value
Fair Limits
Disparate Impact
0.972
0.787
-0.011
-0.092
0.001
0.086
-0.001
-0.086
-0.001
-0.082
0.998
0.896
1.000
1.022
1.017
1.015
1.634

FIG. 16

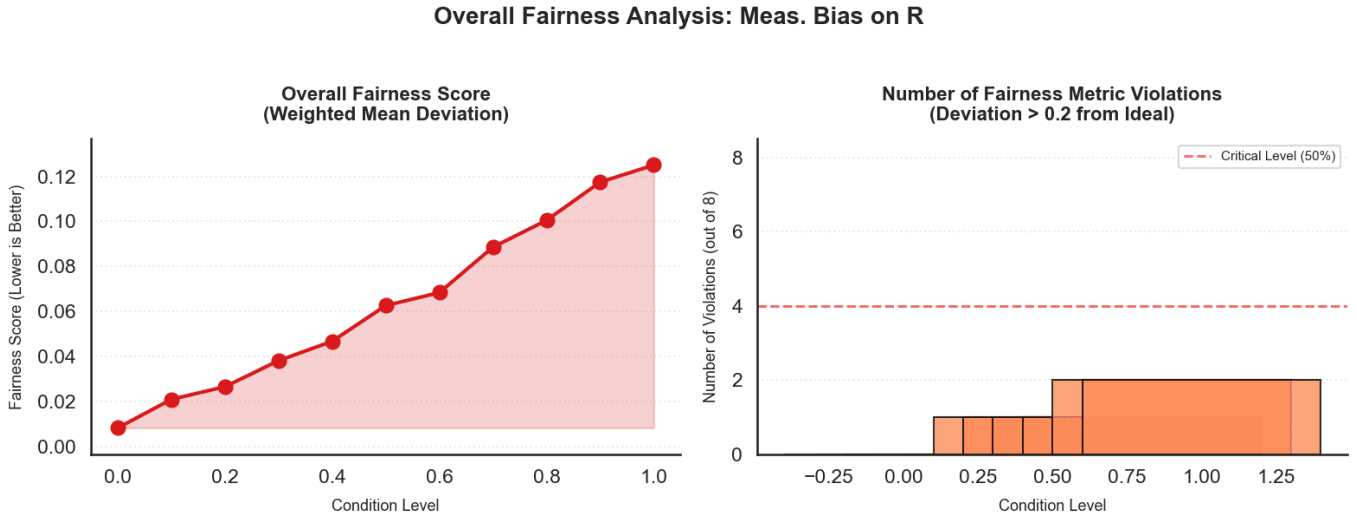
Overall Fairness Analysis: Meas. Bias on R
Overall Fairness Score
(Weighted Mean Deviation)
Fairness Score (Lower is Better)
Condition Level
Number of Fairness Metric Violations
(Deviation > 0.2 from Ideal)
Number of Violations (out of 8)
Critical Level (50%)
Condition Level

FIG. 17

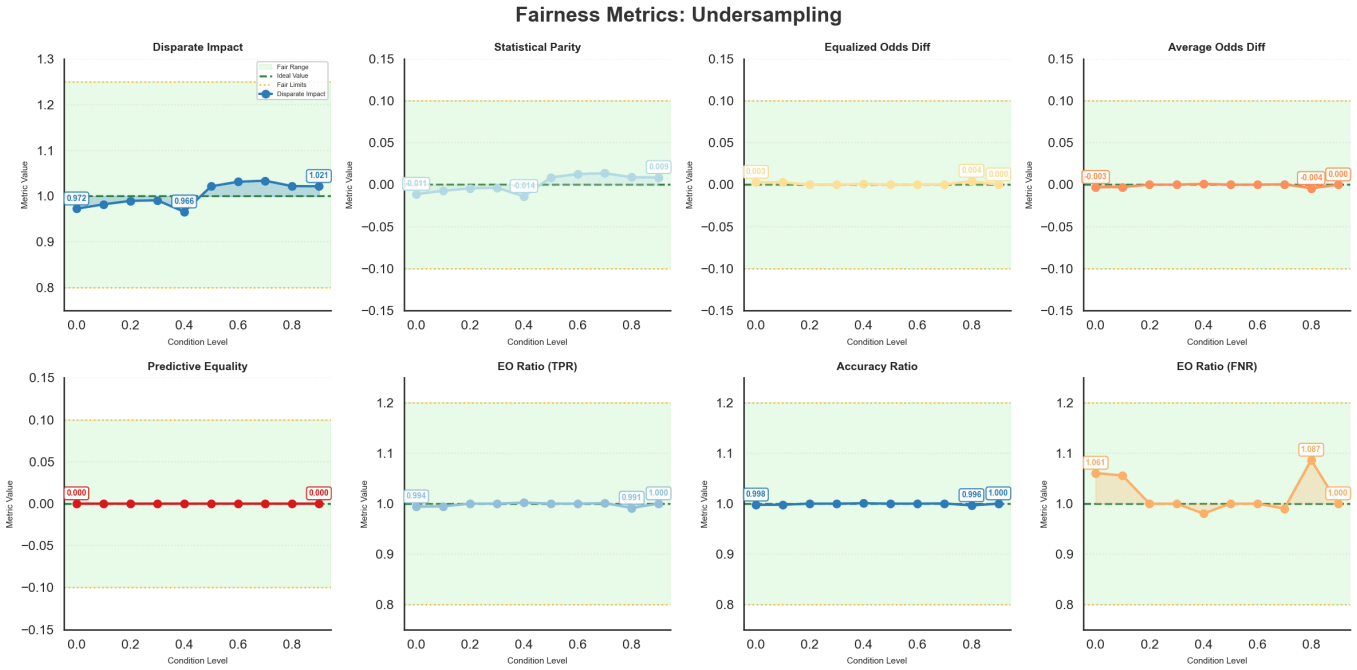
Fairness Metrics: Undersampling
Disparate Impact
Statistical Parity
Equalized Odds Diff
Average Odds Diff
Predictive Equality
EO Ratio (TPR)
Accuracy Ratio
EO Ratio (FNR)
Metric Value
Condition Level
Fair Range
Ideal Value
Fair Limits
Disparate Impact

FIG. 18

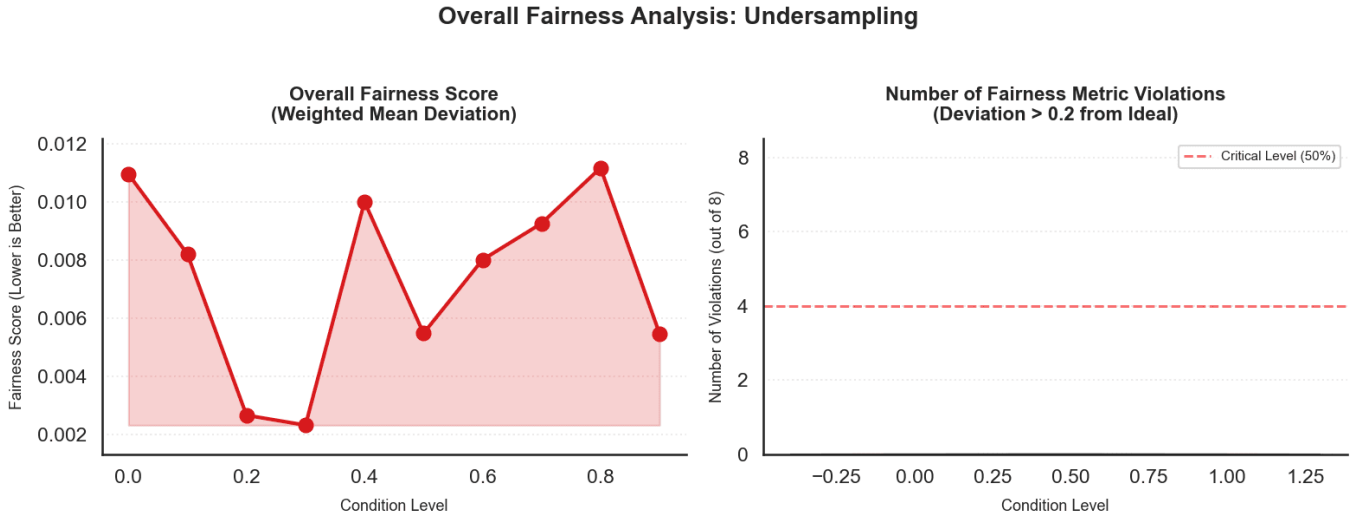
Overall Fairness Analysis: Undersampling
Overall Fairness Score
(Weighted Mean Deviation)
Fairness Score (Lower is Better)
Condition Level
Number of Fairness Metric Violations
(Deviation > 0.2 from Ideal)
Number of Violations (out of 8)
Critical Level (50%)
Condition Level

FIG. 19

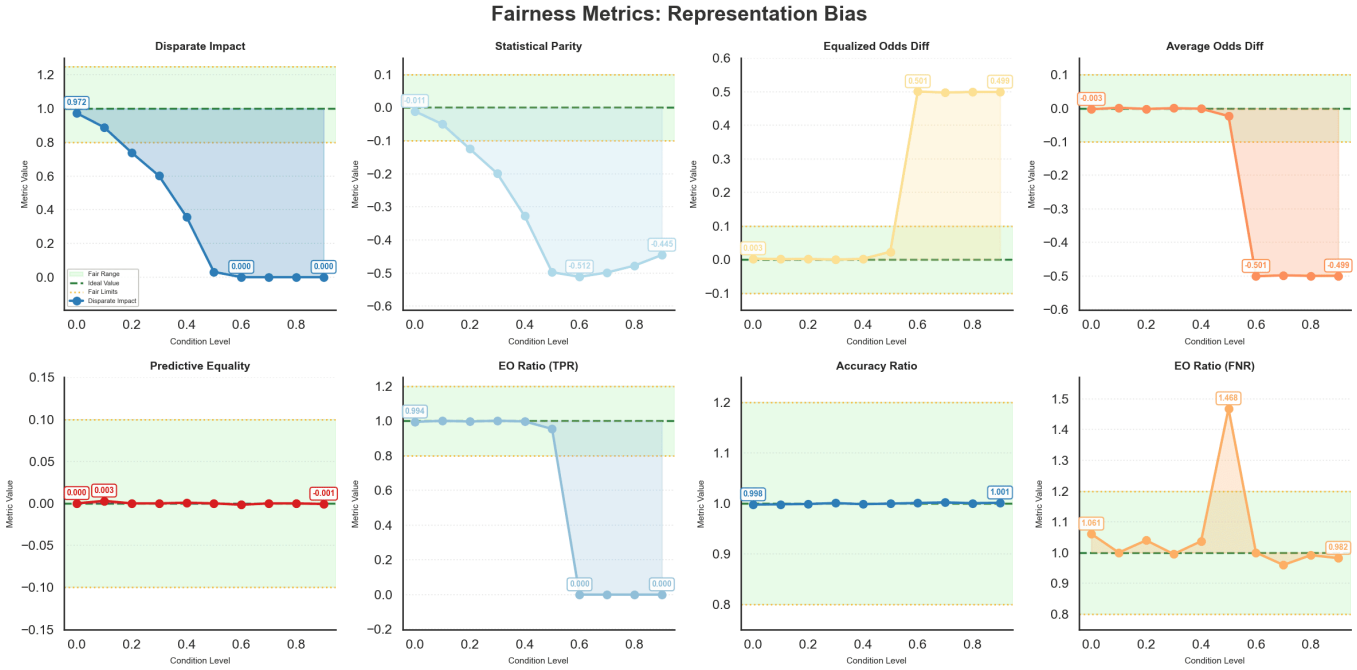
Fairness Metrics: Representation Bias
Disparate Impact
Statistical Parity
Equalized Odds Diff
Average Odds Diff
Predictive Equality
EO Ratio (TPR)
Accuracy Ratio
EO Ratio (FNR)
Metric Value
Condition Level
Fair Range
Ideal Value
Fair Limits
Disparate Impact

FIG. 20

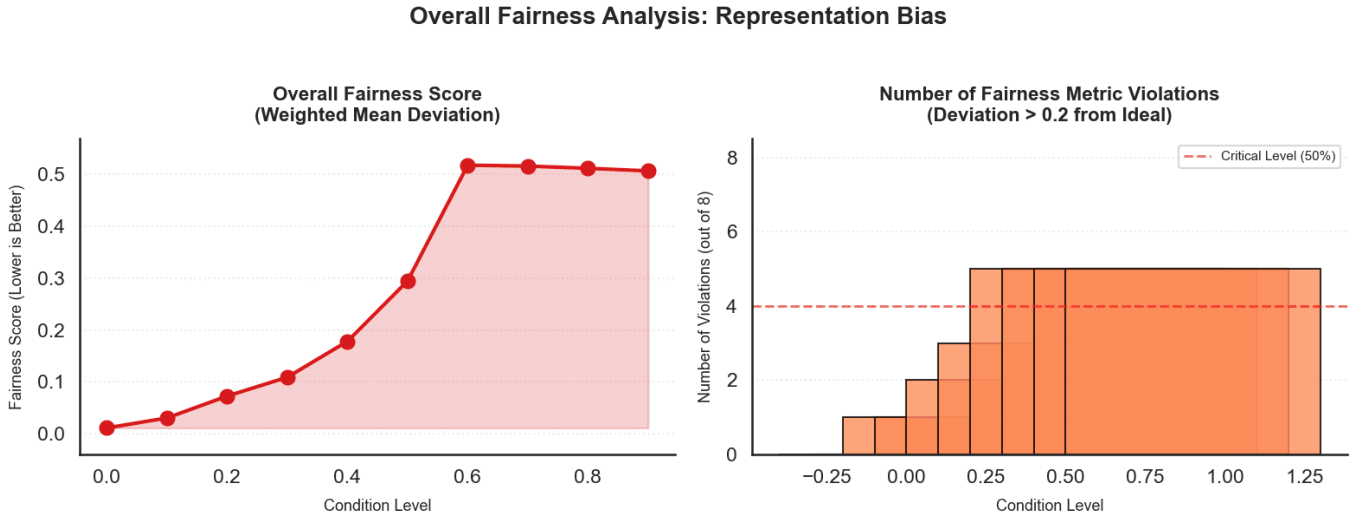
Overall Fairness Analysis: Representation Bias
Overall Fairness Score
(Weighted Mean Deviation)
Fairness Score (Lower is Better)
Condition Level
Number of Fairness Metric Violations
(Deviation > 0.2 from Ideal)
Number of Violations (out of 8)
Critical Level (50%)
Condition Level

FIG. 21

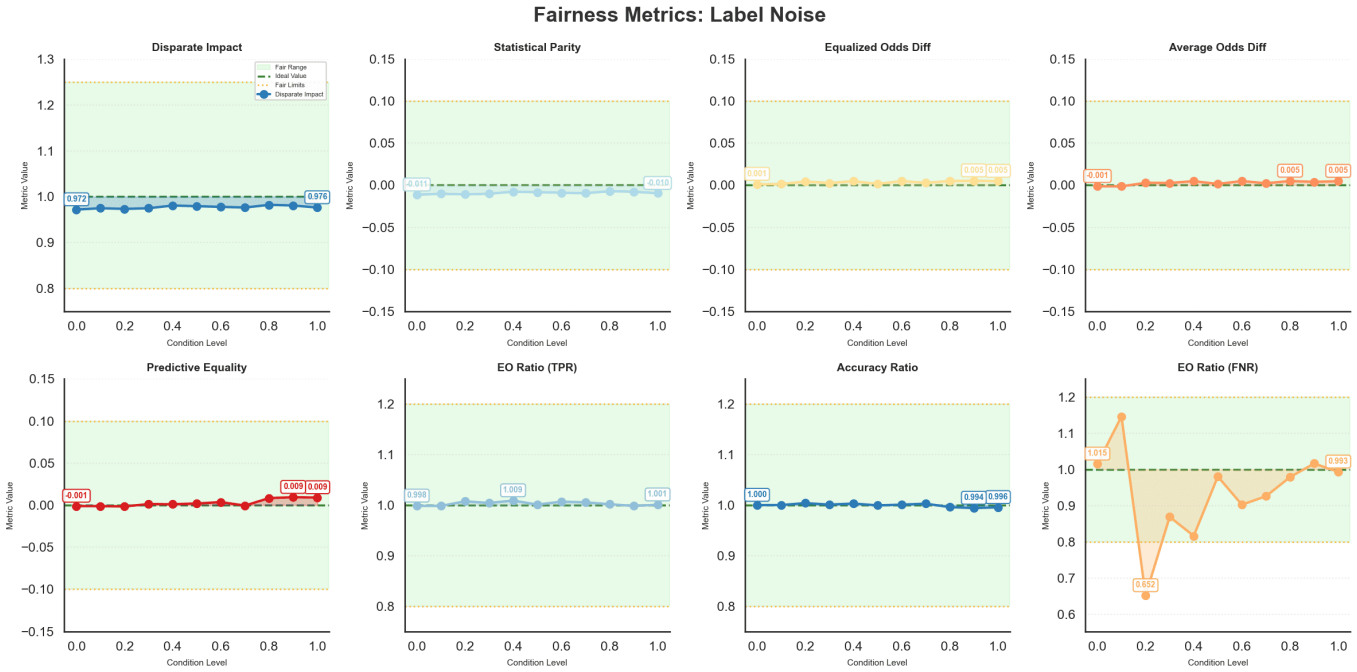
Fairness Metrics: Label Noise
Disparate Impact
Statistical Parity
Equalized Odds Diff
Average Odds Diff
Predictive Equality
EO Ratio (TPR)
Accuracy Ratio
EO Ratio (FNR)
Metric Value
Condition Level

FIG. 22

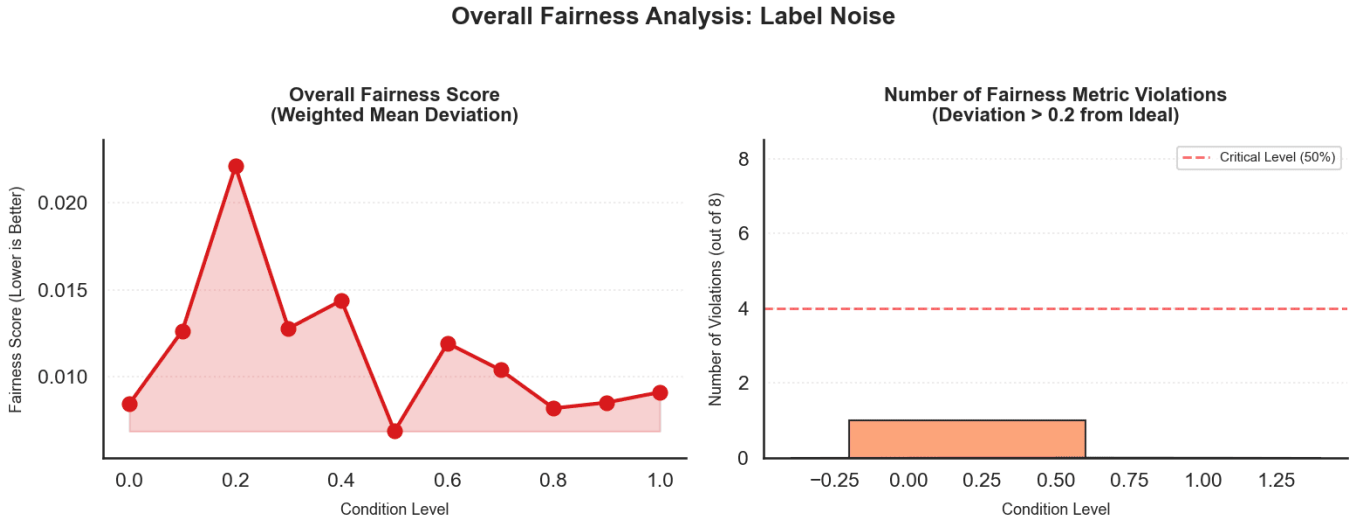
Overall Fairness Analysis: Label Noise
Overall Fairness Score
(Weighted Mean Deviation)
Fairness Score (Lower is Better)
Condition Level
Number of Fairness Metric Violations
(Deviation > 0.2 from Ideal)
Number of Violations (out of 8)
Critical Level (50%)
Condition Level

FIG. 23

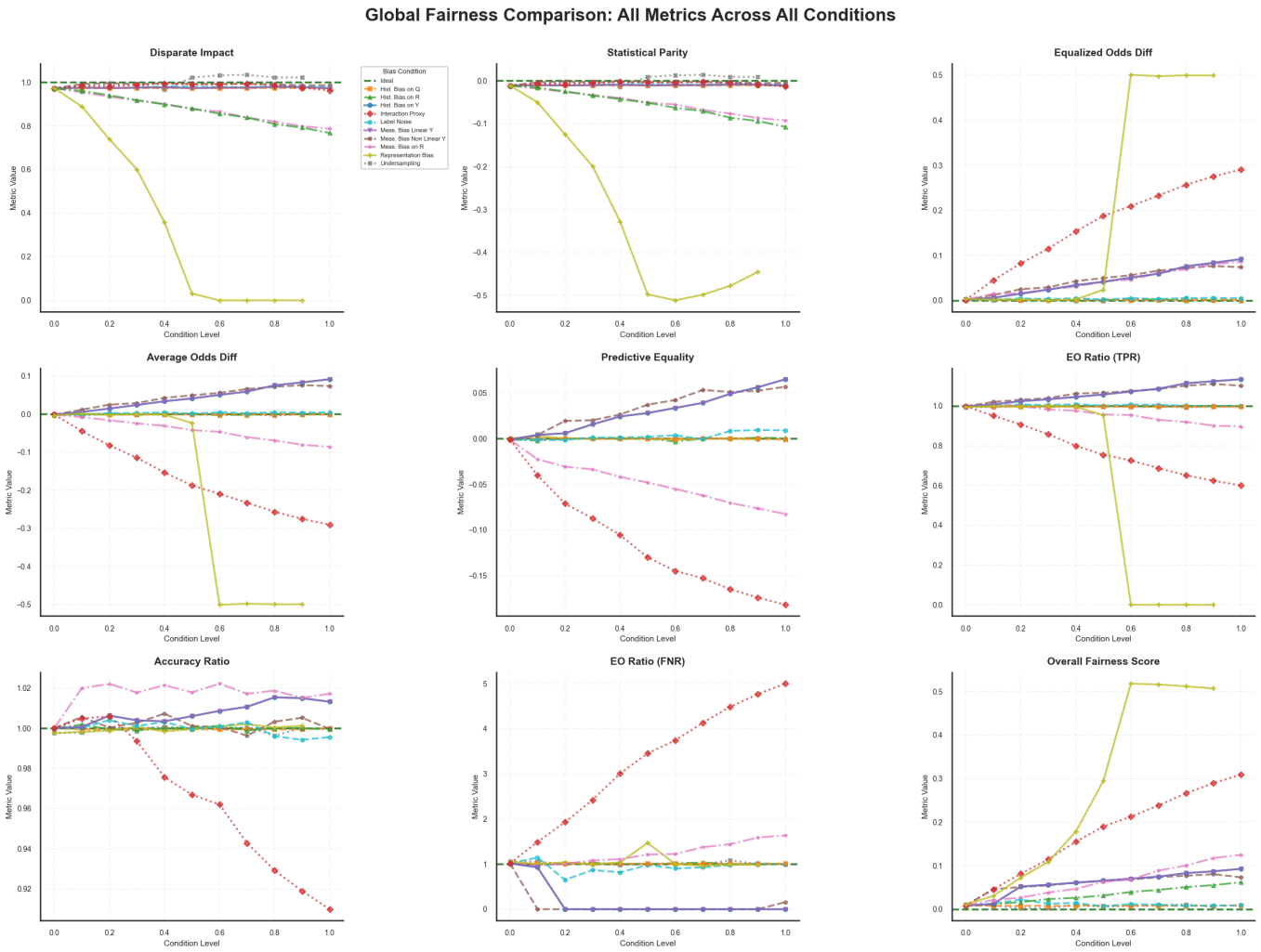
Global Fairness Comparison: All Metrics Across All Conditions
Disparate Impact
Statistical Parity
Equalized Odds Diff
Average Odds Diff
Predictive Equality
EO Ratio (TPR)
Accuracy Ratio
EO Ratio (FNR)
Overall Fairness Score
Metric Value
Condition Level
Bias Condition
Ideal
Hist. Bias on Q
Hist. Bias on R
Hist. Bias on Y
Interaction Proxy
Label Noise
Meas. Bias Linear Y
Meas. Bias Non Linear Y
Meas. Bias on R
Representation Bias
Undersampling

FIG. 24

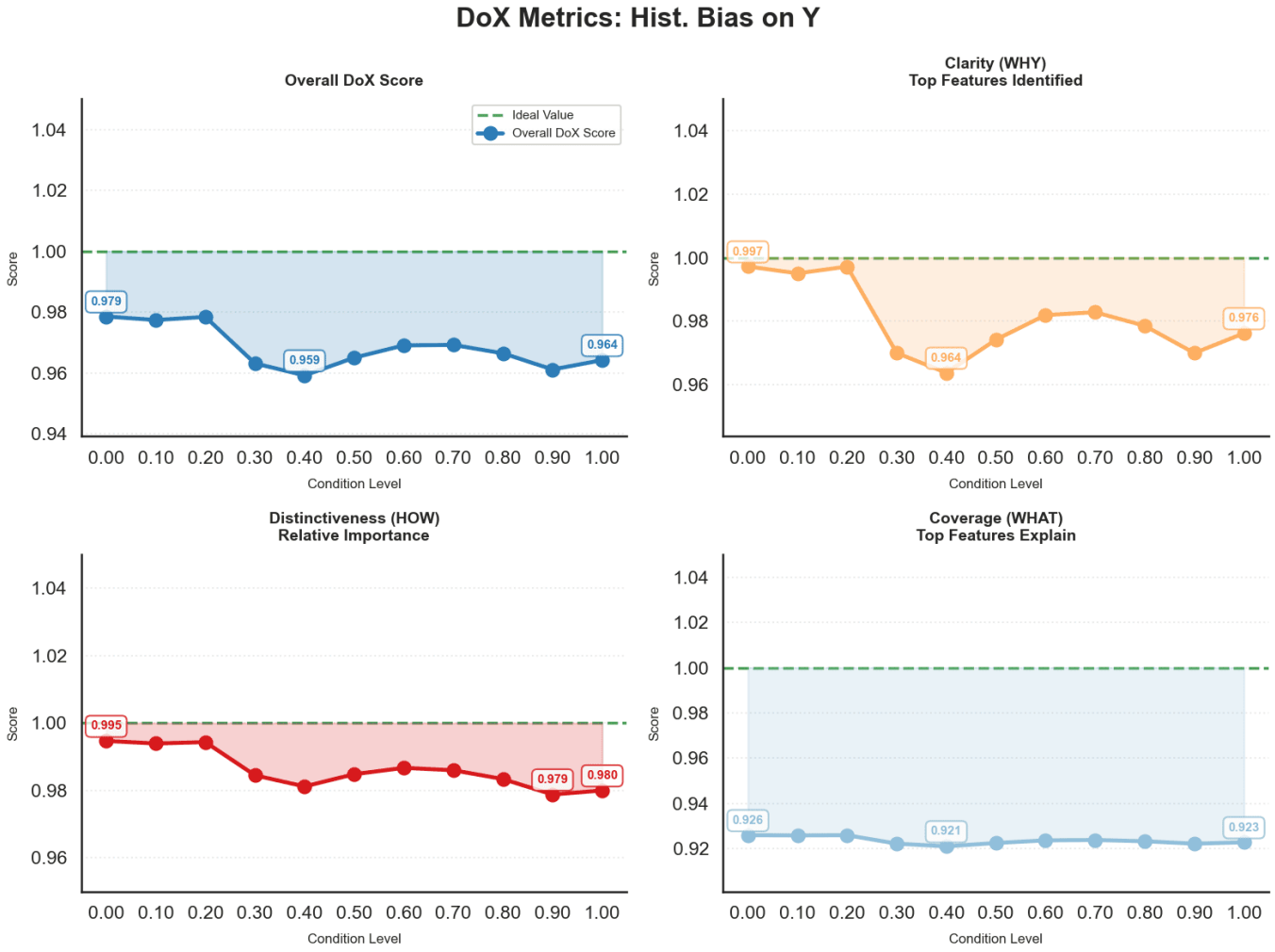
DoX Metrics: Hist. Bias on Y
Overall DoX Score
Ideal Value
Overall DoX Score
Score
Condition Level
0.979
0.959
0.964
Clarity (WHY)
Top Features Identified
0.997
0.964
0.976
Distinctiveness (HOW)
Relative Importance
0.995
0.979
0.980
Coverage (WHAT)
Top Features Explain
0.926
0.921
0.923

FIG. 25

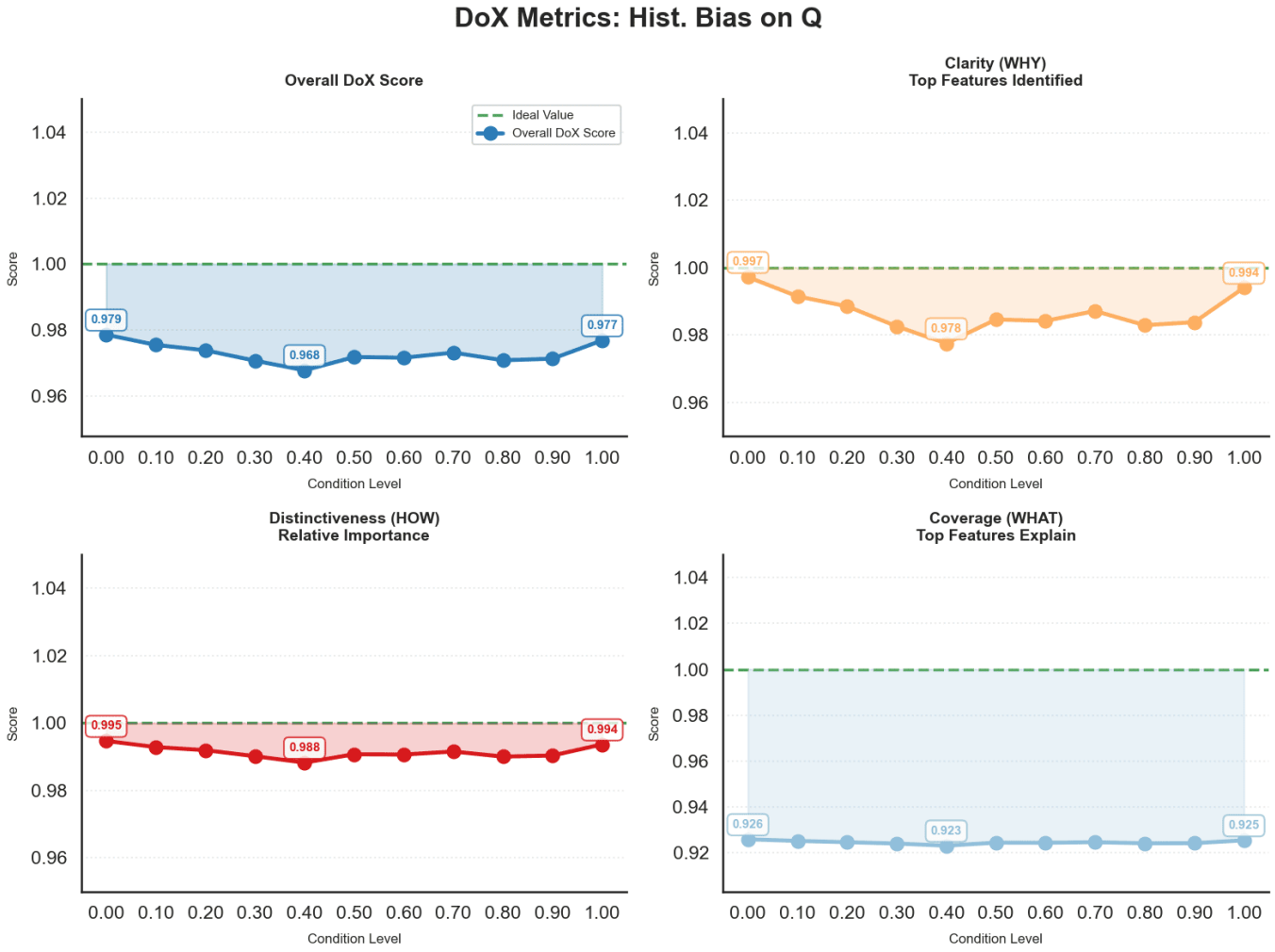
DoX Metrics: Hist. Bias on Q
Overall DoX Score
Ideal Value
Overall DoX Score
Score
Condition Level
0.979
0.968
0.977
Clarity (WHY)
Top Features Identified
0.997
0.978
0.994
Distinctiveness (HOW)
Relative Importance
0.995
0.988
0.994
Coverage (WHAT)
Top Features Explain
0.926
0.923
0.925

FIG. 26

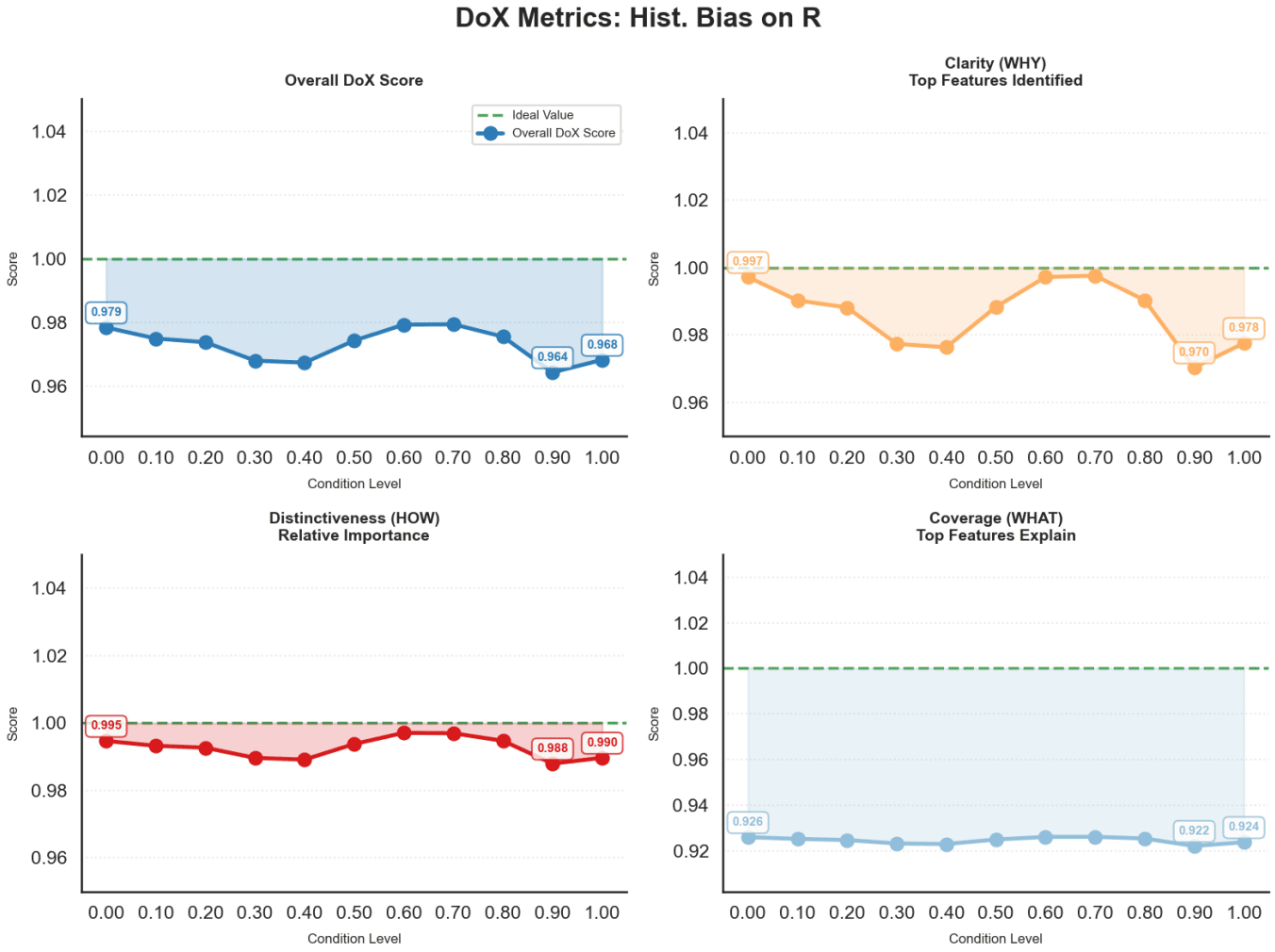
DoX Metrics: Hist. Bias on R
Overall DoX Score
Ideal Value
Overall DoX Score
Score
Condition Level
0.979
0.964
0.968
Clarity (WHY)
Top Features Identified
0.997
0.970
0.978
Distinctiveness (HOW)
Relative Importance
0.995
0.988
0.990
Coverage (WHAT)
Top Features Explain
0.926
0.922
0.924

FIG. 27

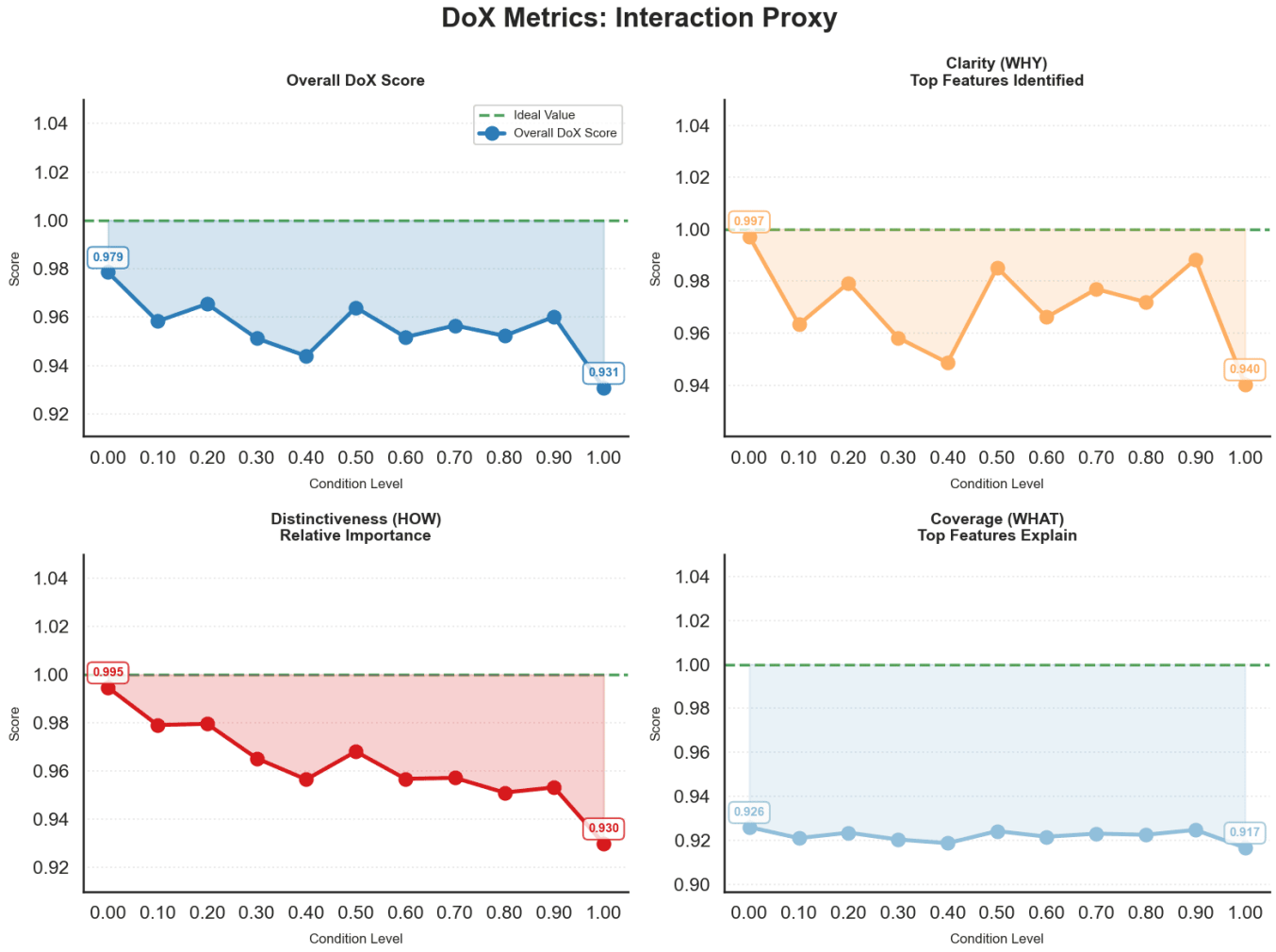
DoX Metrics: Interaction Proxy
Overall DoX Score
Ideal Value
Overall DoX Score
0.979
0.931
Clarity (WHY)
Top Features Identified
0.997
0.940
Distinctiveness (HOW)
Relative Importance
0.995
0.930
Coverage (WHAT)
Top Features Explain
0.926
0.917
Score
Condition Level
0.00 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00

FIG. 28

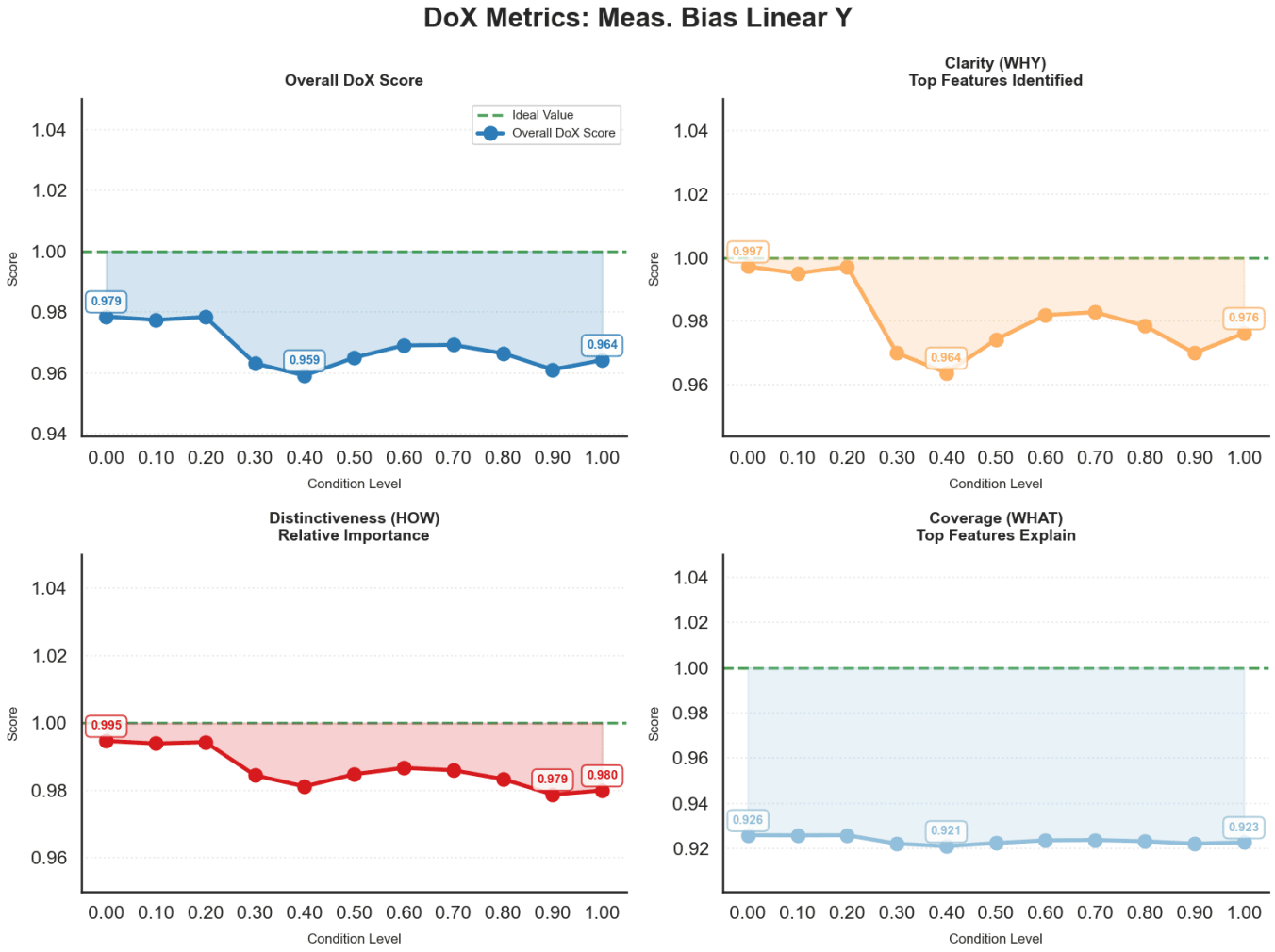
DoX Metrics: Meas. Bias Linear Y
Overall DoX Score
Ideal Value
Overall DoX Score
Score
Condition Level
0.979
0.959
0.964
Clarity (WHY)
Top Features Identified
0.997
0.964
0.976
Distinctiveness (HOW)
Relative Importance
0.995
0.979
0.980
Coverage (WHAT)
Top Features Explain
0.926
0.921
0.923

FIG. 29

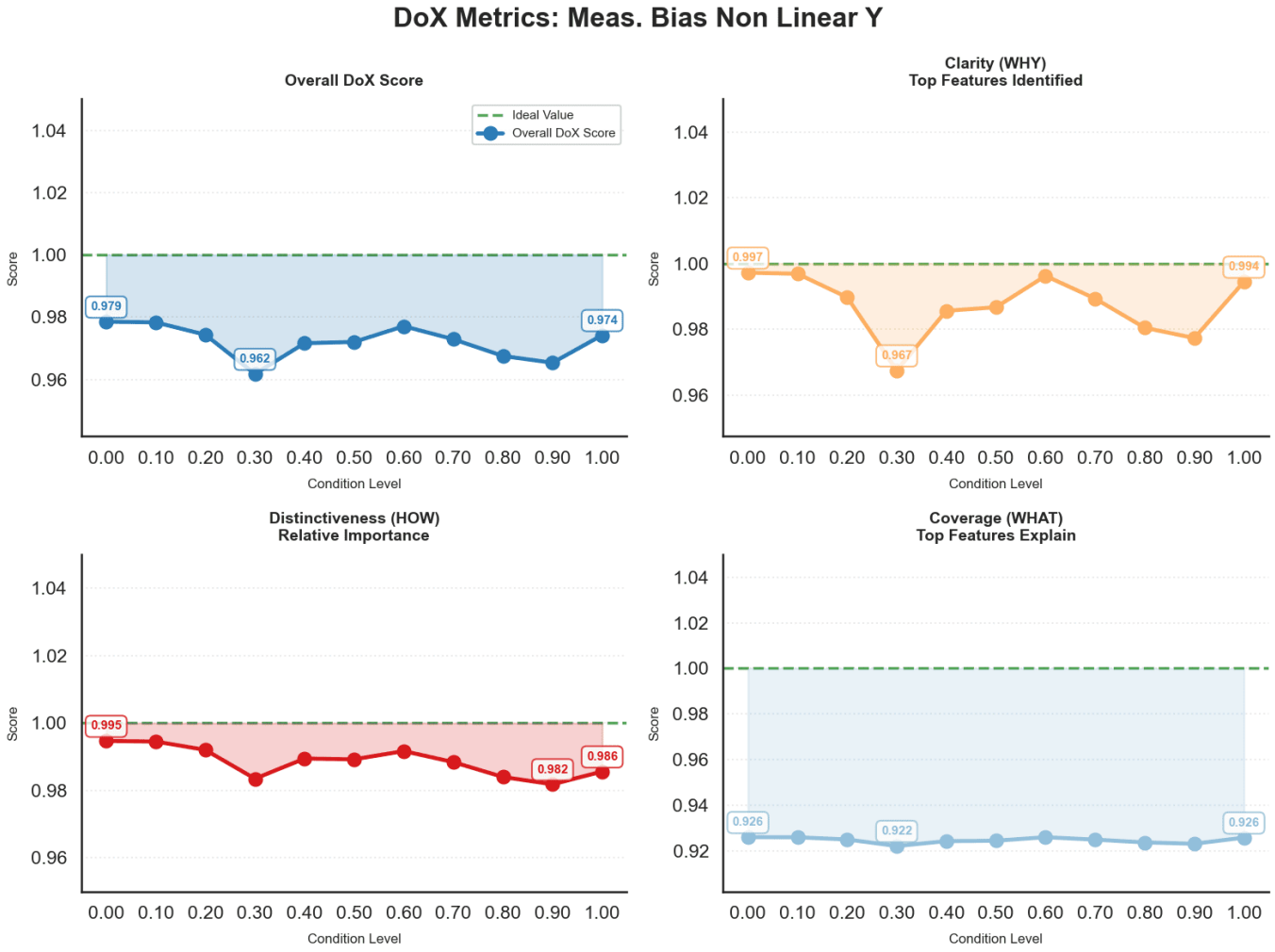
DoX Metrics: Meas. Bias Non Linear Y
Overall DoX Score
Ideal Value
Overall DoX Score
0.979
0.962
0.974
Clarity (WHY)
Top Features Identified
0.997
0.967
0.994
Distinctiveness (HOW)
Relative Importance
0.995
0.982
0.986
Coverage (WHAT)
Top Features Explain
0.926
0.922
0.926
Score
Condition Level
0.00 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00

FIG. 30

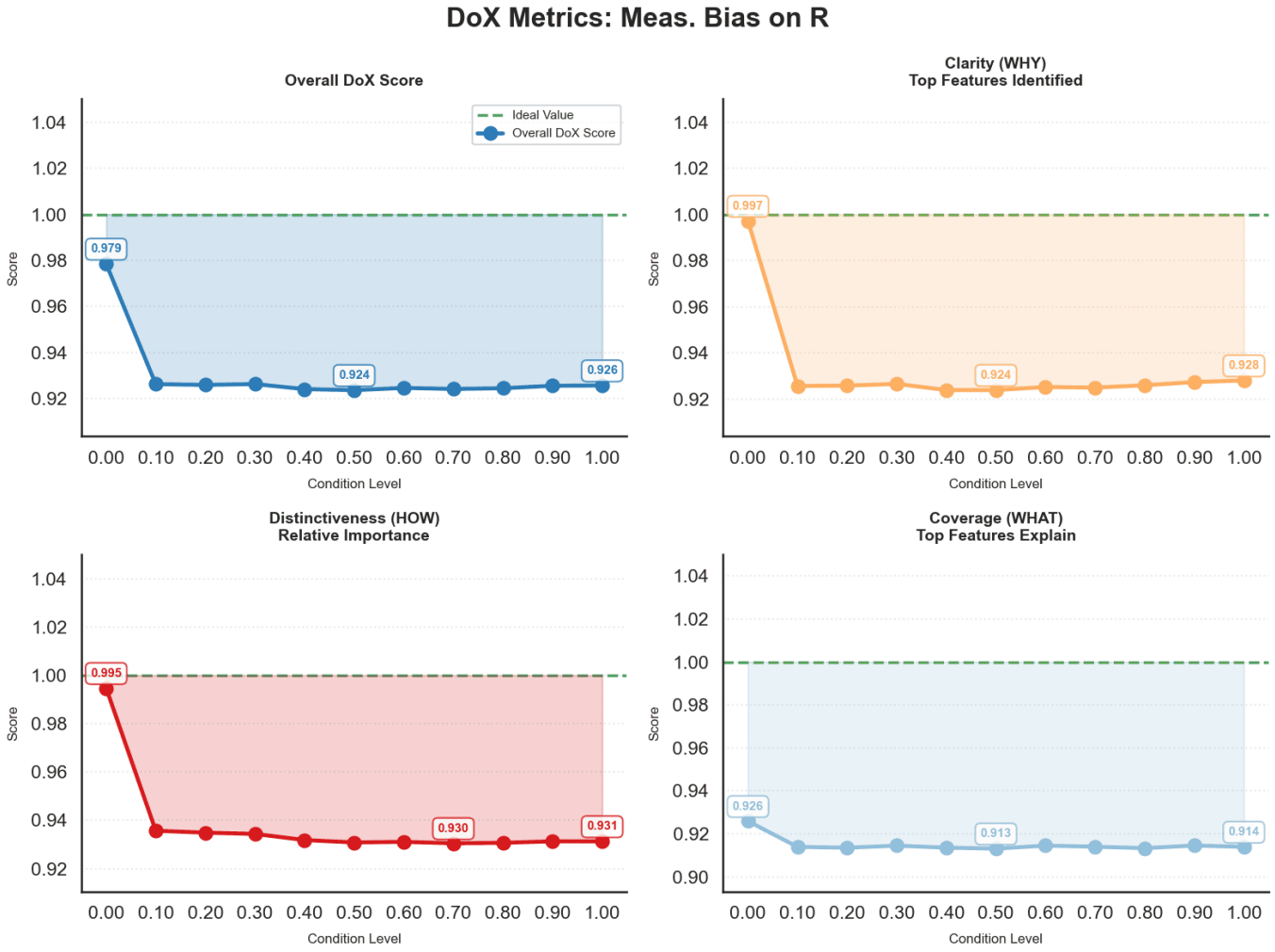
DoX Metrics: Meas. Bias on R
Overall DoX Score
Ideal Value
Overall DoX Score
0.979
0.924
0.926
Clarity (WHY)
Top Features Identified
0.997
0.924
0.928
Distinctiveness (HOW)
Relative Importance
0.995
0.930
0.931
Coverage (WHAT)
Top Features Explain
0.926
0.913
0.914
Score
Condition Level

FIG. 31

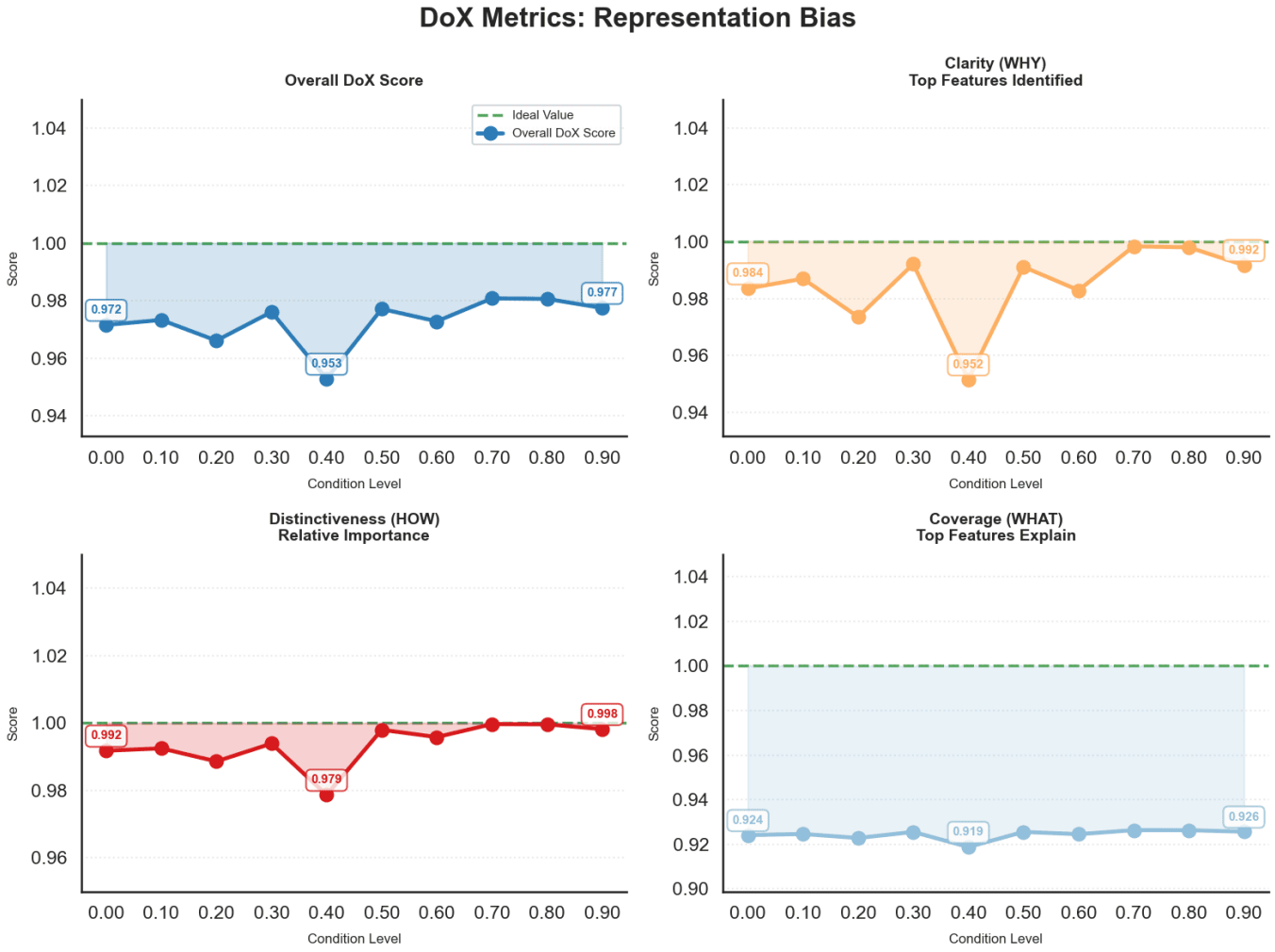
DoX Metrics: Representation Bias
Overall DoX Score
Ideal Value
Overall DoX Score
0.972
0.953
0.977
Clarity (WHY)
Top Features Identified
0.984
0.952
0.992
Distinctiveness (HOW)
Relative Importance
0.992
0.979
0.998
Coverage (WHAT)
Top Features Explain
0.924
0.919
0.926
Score
Condition Level
1.04
1.02
1.00
0.98
0.96
0.94
0.92
0.90
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90

FIG. 32

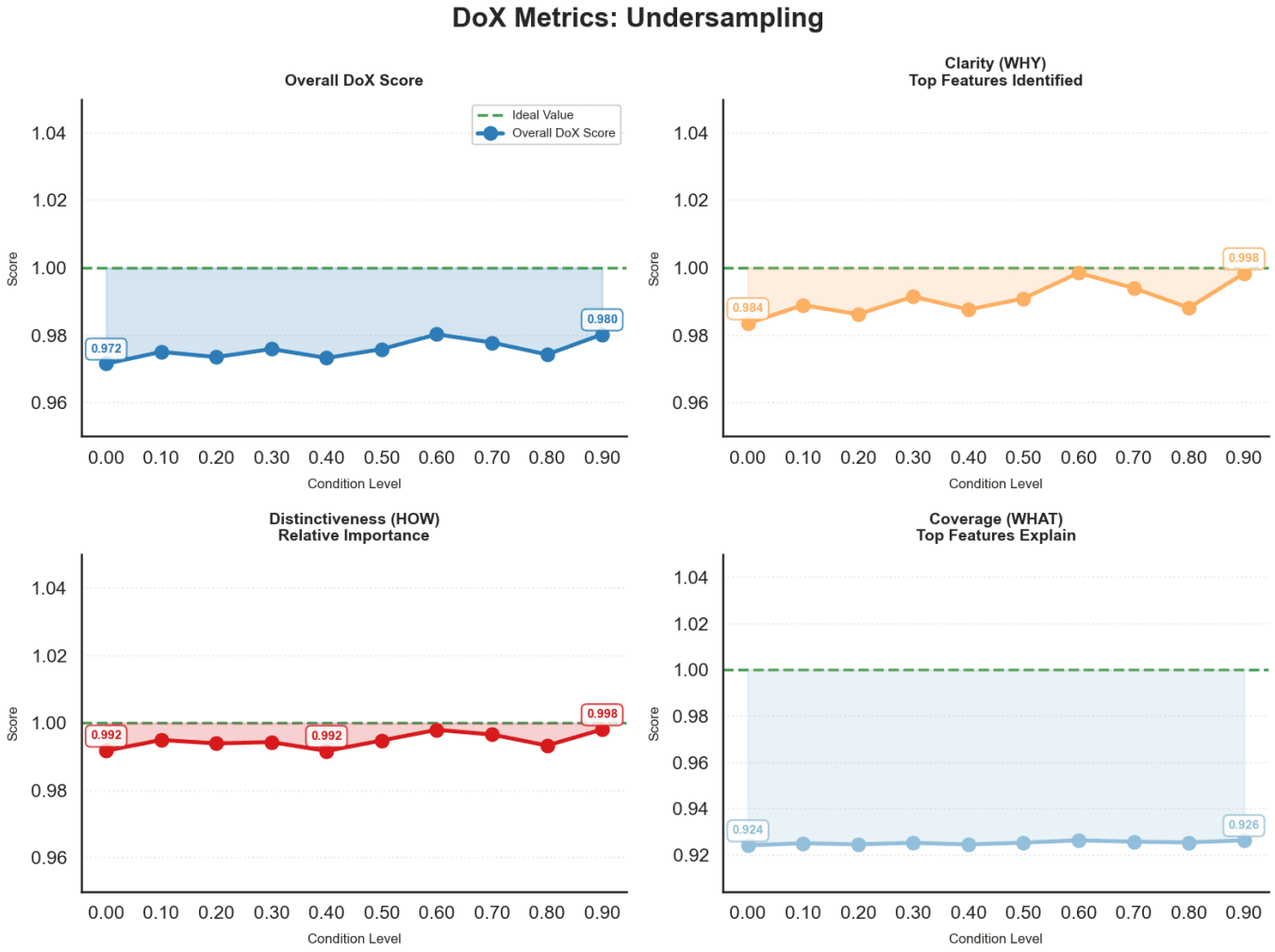
DoX Metrics: Undersampling
Overall DoX Score
Ideal Value
Overall DoX Score
0.972
0.980
Clarity (WHY)
Top Features Identified
0.984
0.998
Distinctiveness (HOW)
Relative Importance
0.992
0.992
0.998
Coverage (WHAT)
Top Features Explain
0.924
0.926
Score
Condition Level
1.04
1.02
1.00
0.98
0.96
0.94
0.92
0.00 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90

FIG. 33

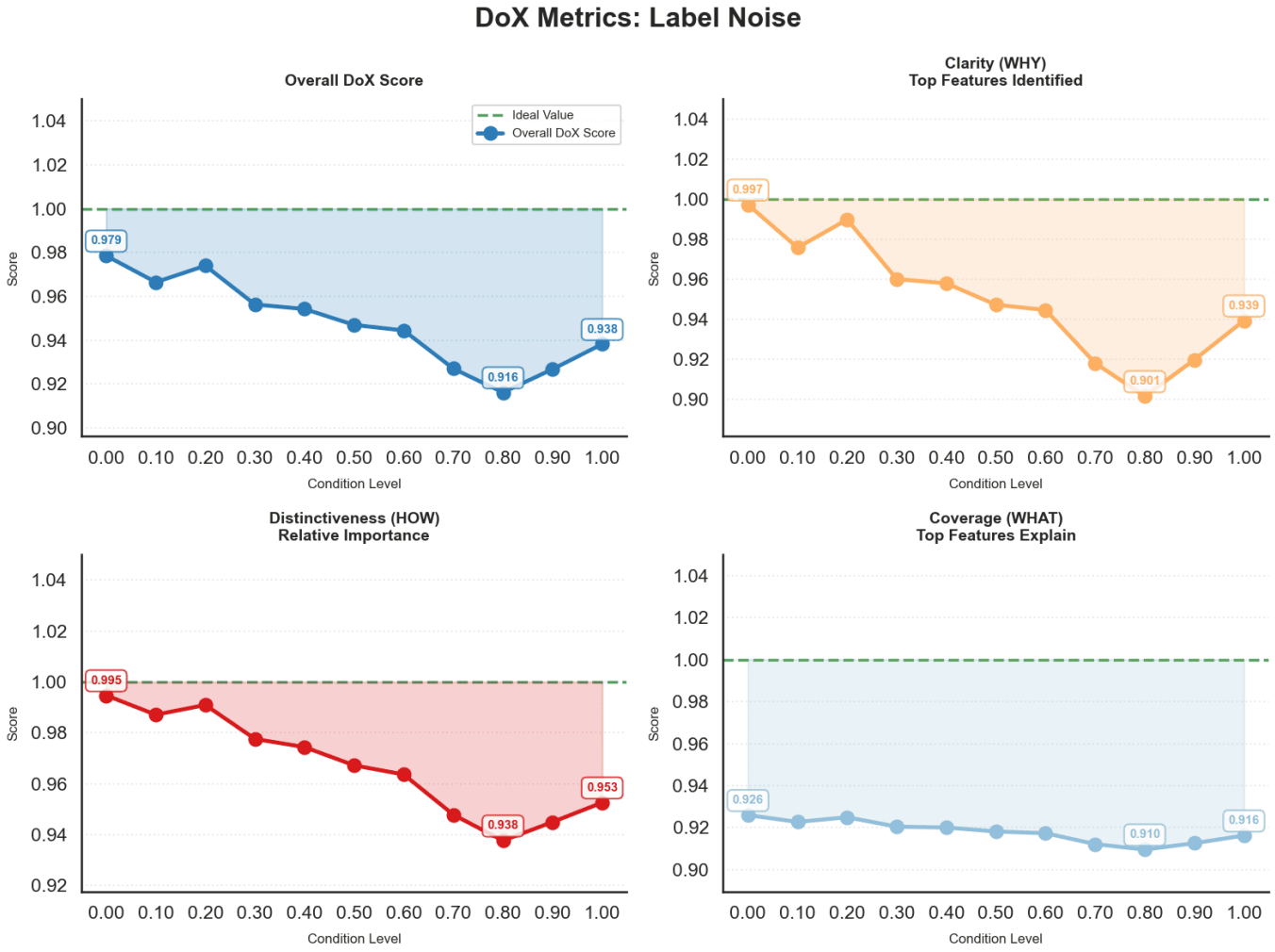
DoX Metrics: Label Noise
Overall DoX Score
Ideal Value
Overall DoX Score
0.979
0.916
0.938
Clarity (WHY)
Top Features Identified
0.997
0.901
0.939
Distinctiveness (HOW)
Relative Importance
0.995
0.938
0.953
Coverage (WHAT)
Top Features Explain
0.926
0.910
0.916
Score
Condition Level
1.04
1.02
1.00
0.98
0.96
0.94
0.92
0.90
0.00 0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90 1.00

FIG. 34

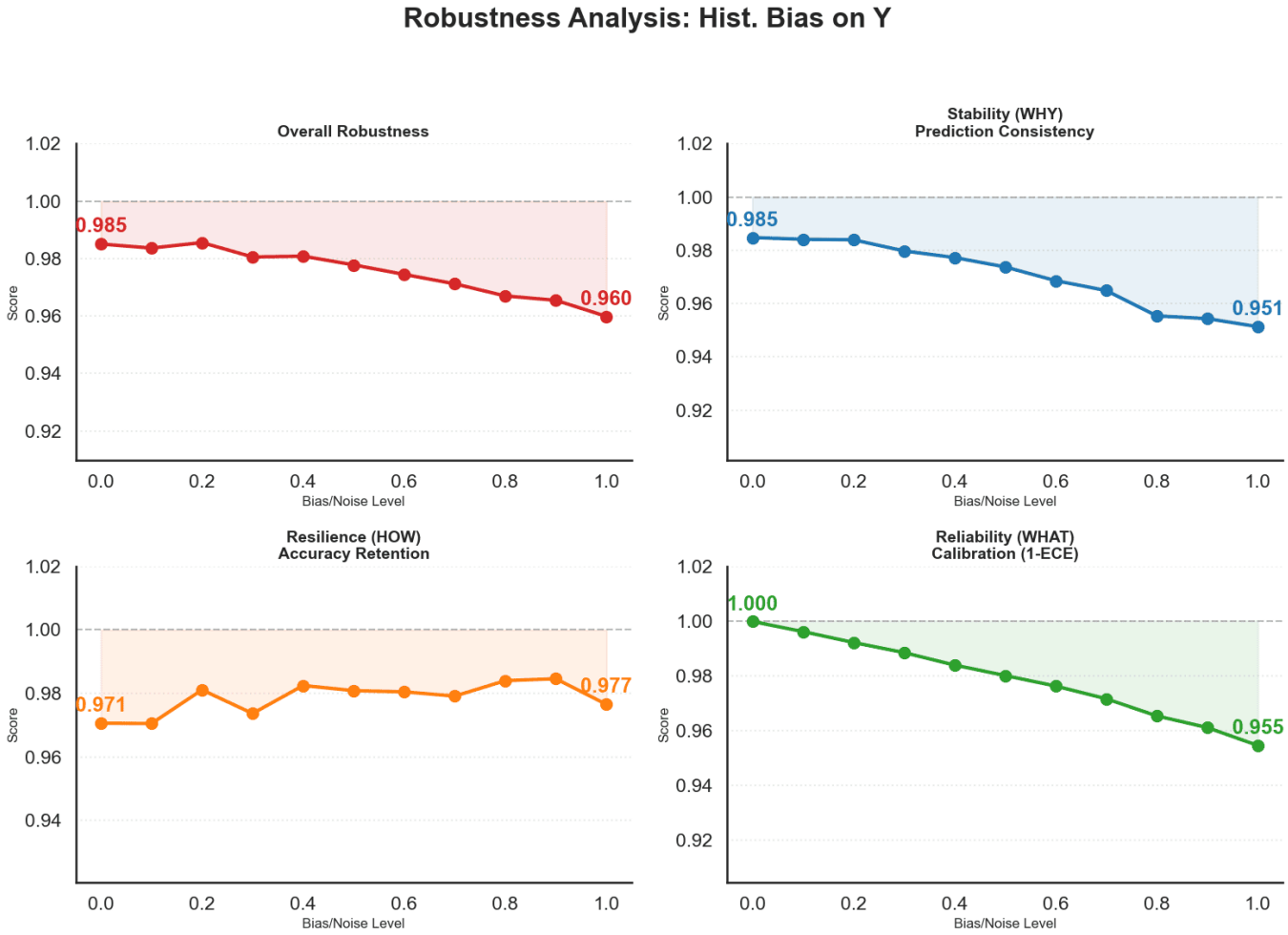
Robustness Analysis: Hist. Bias on Y
Overall Robustness
Stability (WHY)
Prediction Consistency
Resilience (HOW)
Accuracy Retention
Reliability (WHAT)
Calibration (1-ECE)
Score
Bias/Noise Level
0.985
0.960
0.985
0.951
0.971
0.977
1.000
0.955

FIG. 35

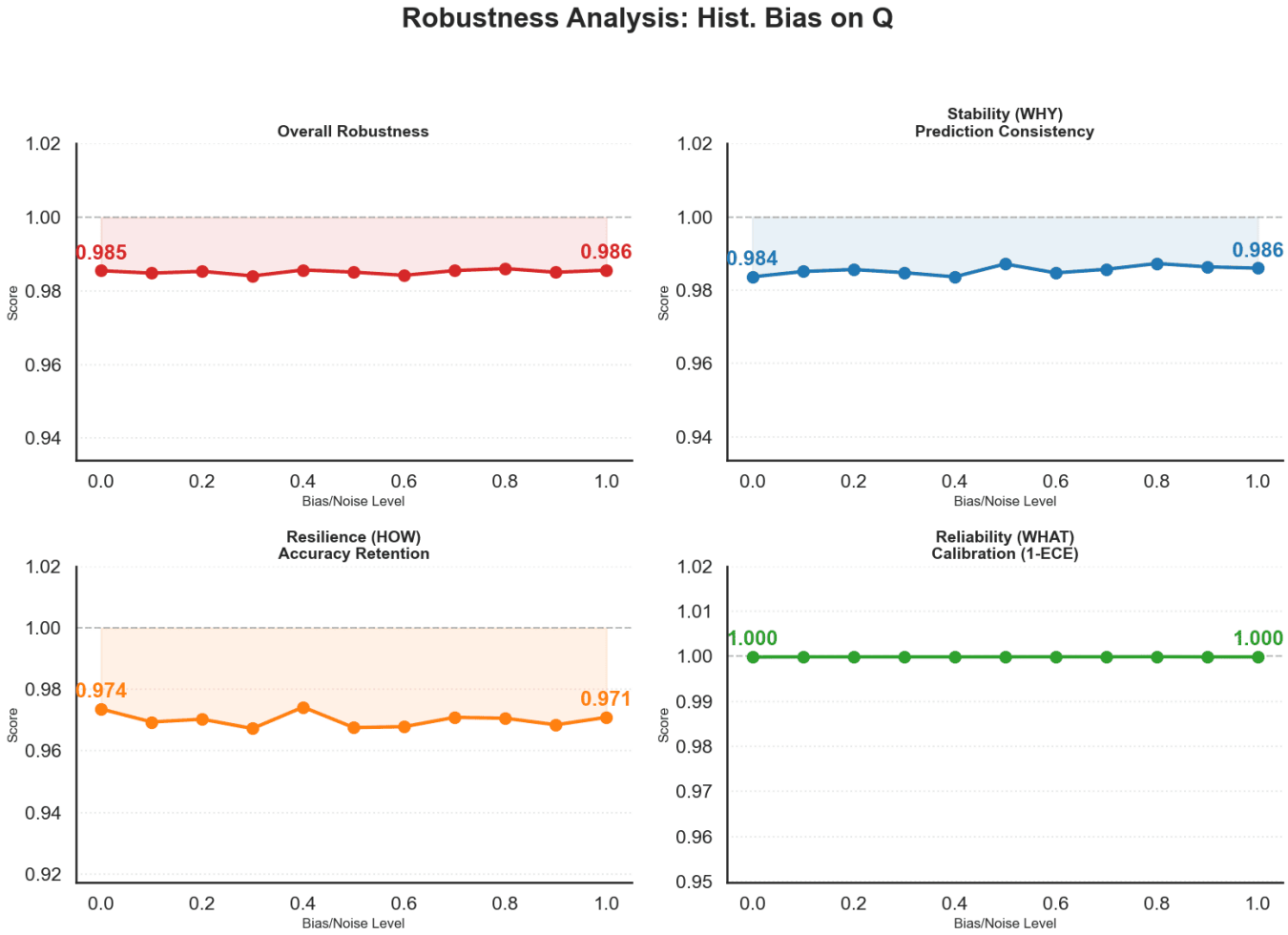
Robustness Analysis: Hist. Bias on Q
Overall Robustness
0.985
0.986
Stability (WHY)
Prediction Consistency
0.984
0.986
Resilience (HOW)
Accuracy Retention
0.974
0.971
Reliability (WHAT)
Calibration (1-ECE)
1.000
1.000
Score
Bias/Noise Level

FIG. 36

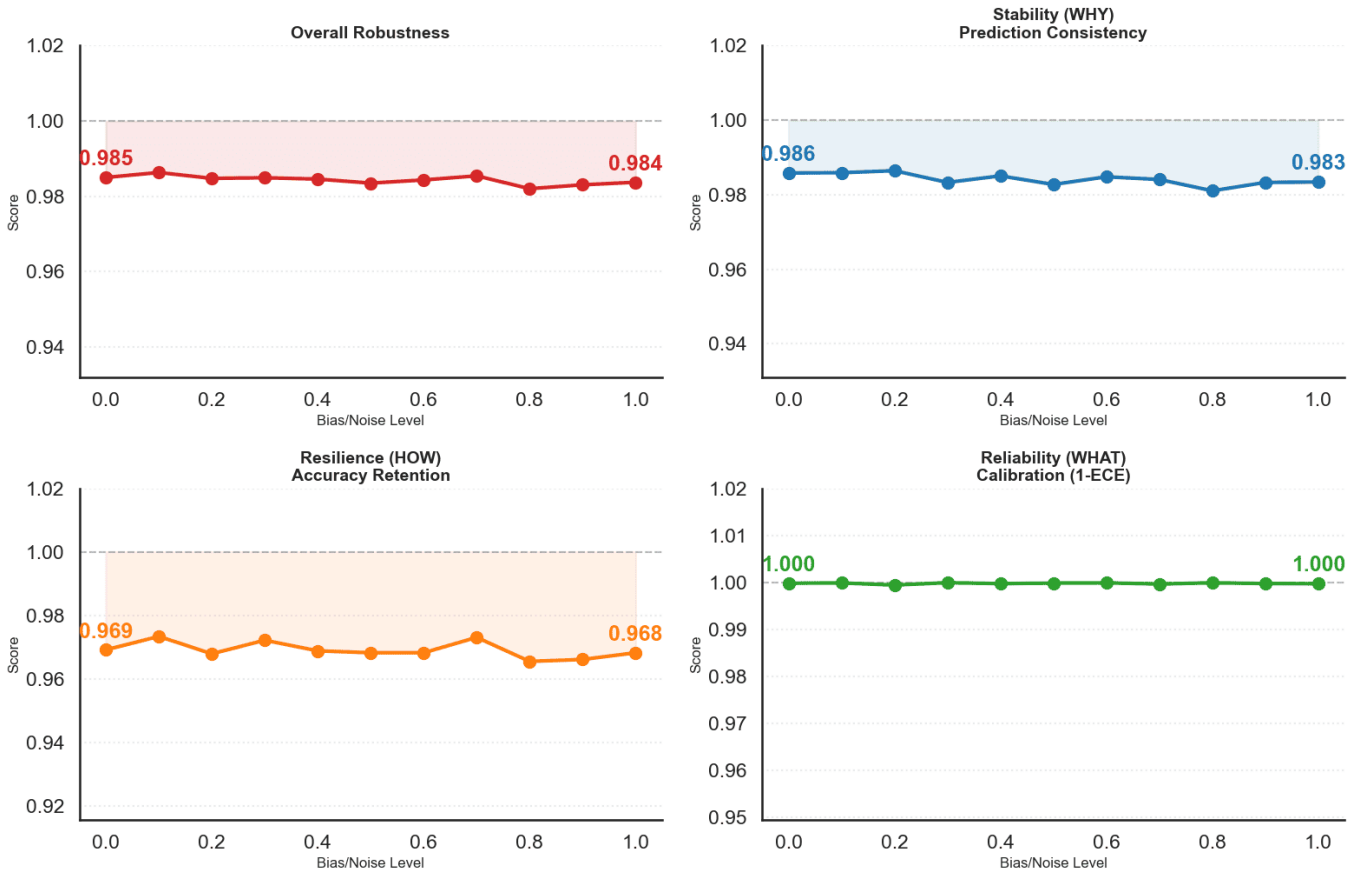
Robustness Analysis: Hist. Bias on R
Overall Robustness
0.985
0.984
Stability (WHY)
Prediction Consistency
0.986
0.983
Resilience (HOW)
Accuracy Retention
0.969
0.968
Reliability (WHAT)
Calibration (1-ECE)
1.000
1.000
Score
Bias/Noise Level

FIG. 37

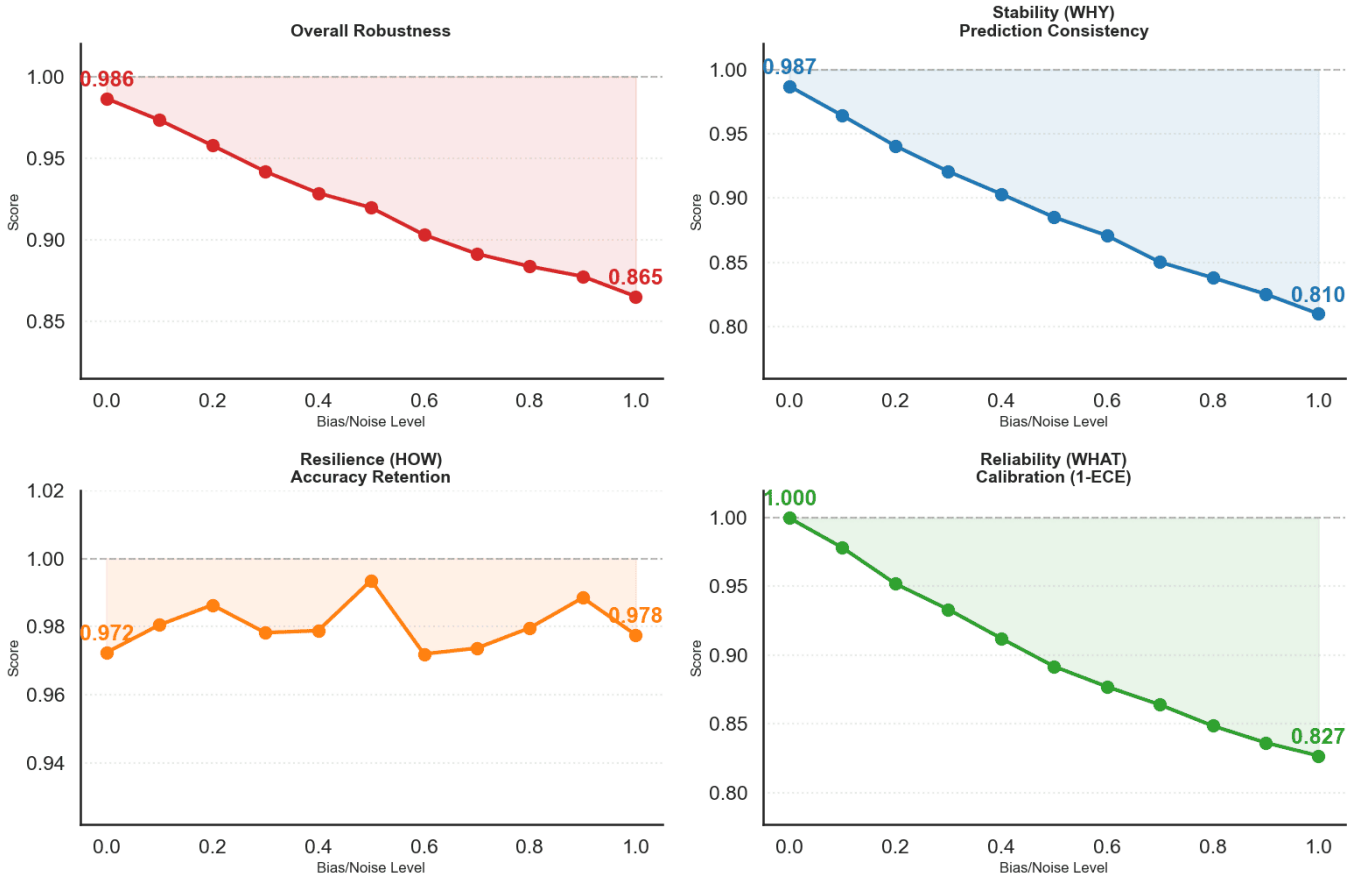
Robustness Analysis: Interaction Proxy
Overall Robustness
0.986
0.865
Score
Bias/Noise Level
Stability (WHY)
Prediction Consistency
0.987
0.810
Score
Bias/Noise Level
Resilience (HOW)
Accuracy Retention
0.972
0.978
Score
Bias/Noise Level
Reliability (WHAT)
Calibration (1-ECE)
1.000
0.827
Score
Bias/Noise Level

FIG. 38

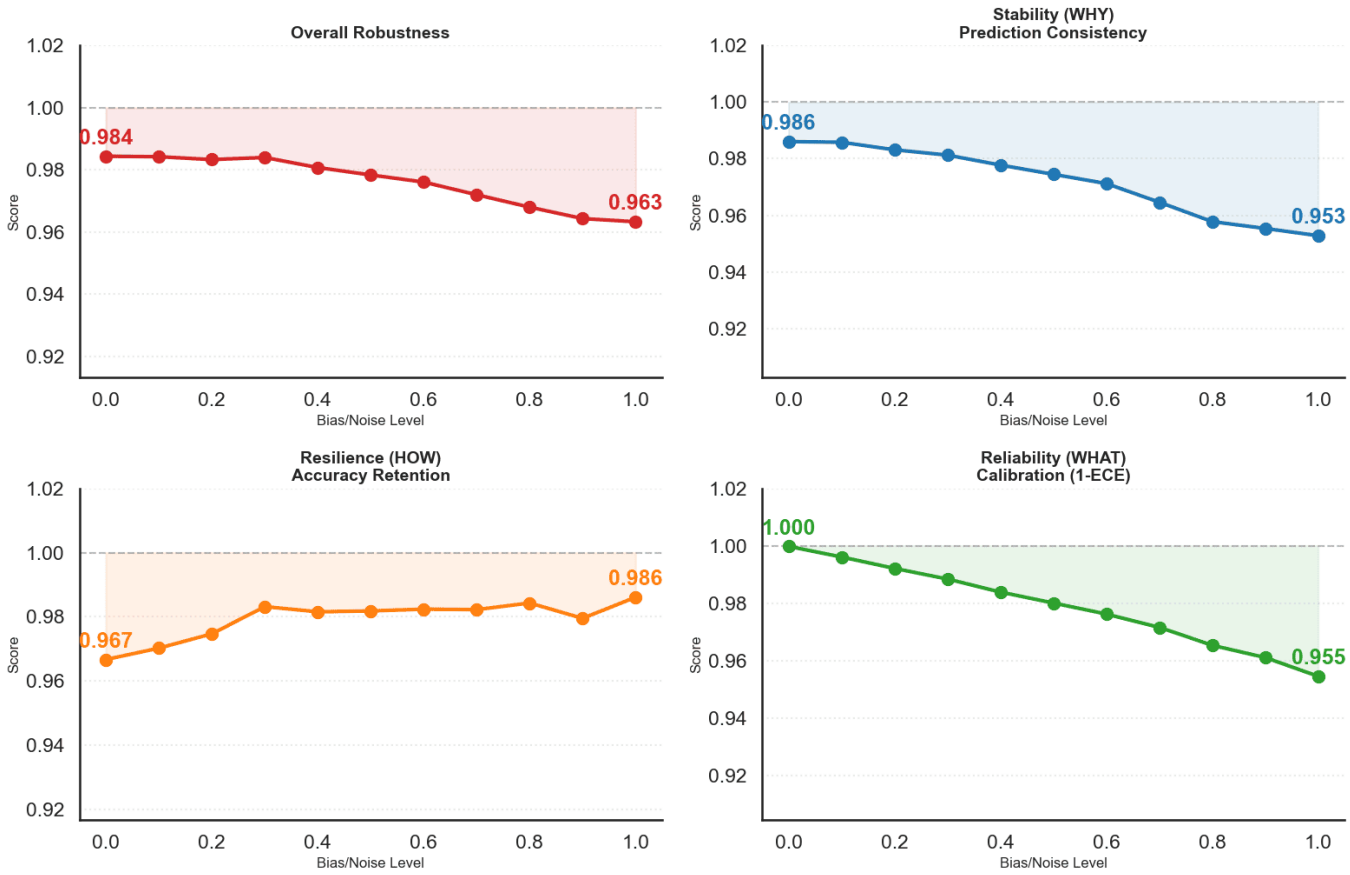
Robustness Analysis: Meas. Bias Linear Y
Overall Robustness
0.984
0.963
Stability (WHY)
Prediction Consistency
0.986
0.953
Resilience (HOW)
Accuracy Retention
0.967
0.986
Reliability (WHAT)
Calibration (1-ECE)
1.000
0.955
Score
Bias/Noise Level
1.02
1.00
0.98
0.96
0.94
0.92
0.0
0.2
0.4
0.6
0.8
1.0

FIG. 39

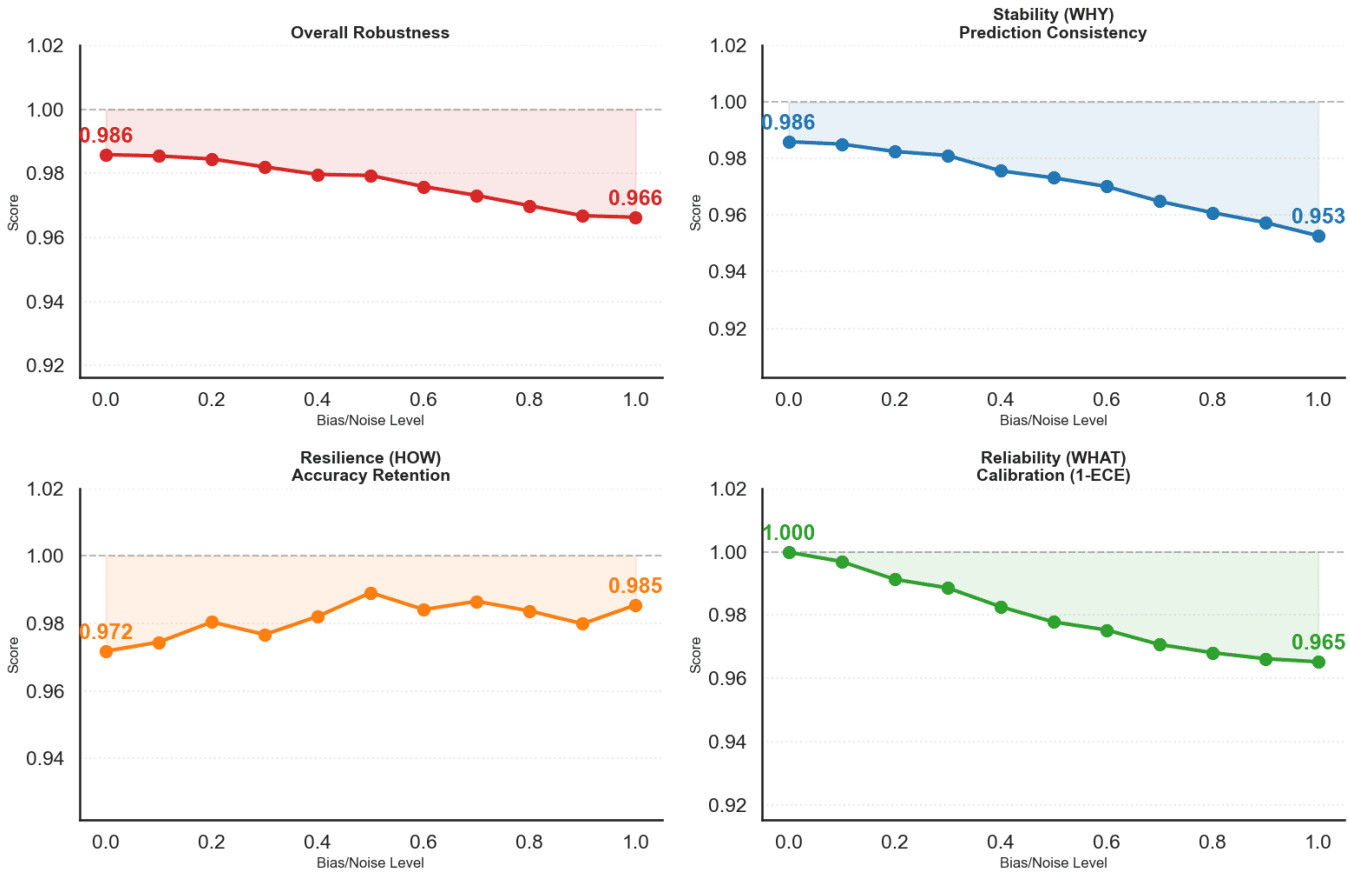
Robustness Analysis: Meas. Bias Non Linear Y
Overall Robustness
Stability (WHY)
Prediction Consistency
Resilience (HOW)
Accuracy Retention
Reliability (WHAT)
Calibration (1-ECE)
Score
Bias/Noise Level
0.986
0.966
0.986
0.953
0.972
0.985
1.000
0.965

FIG. 40

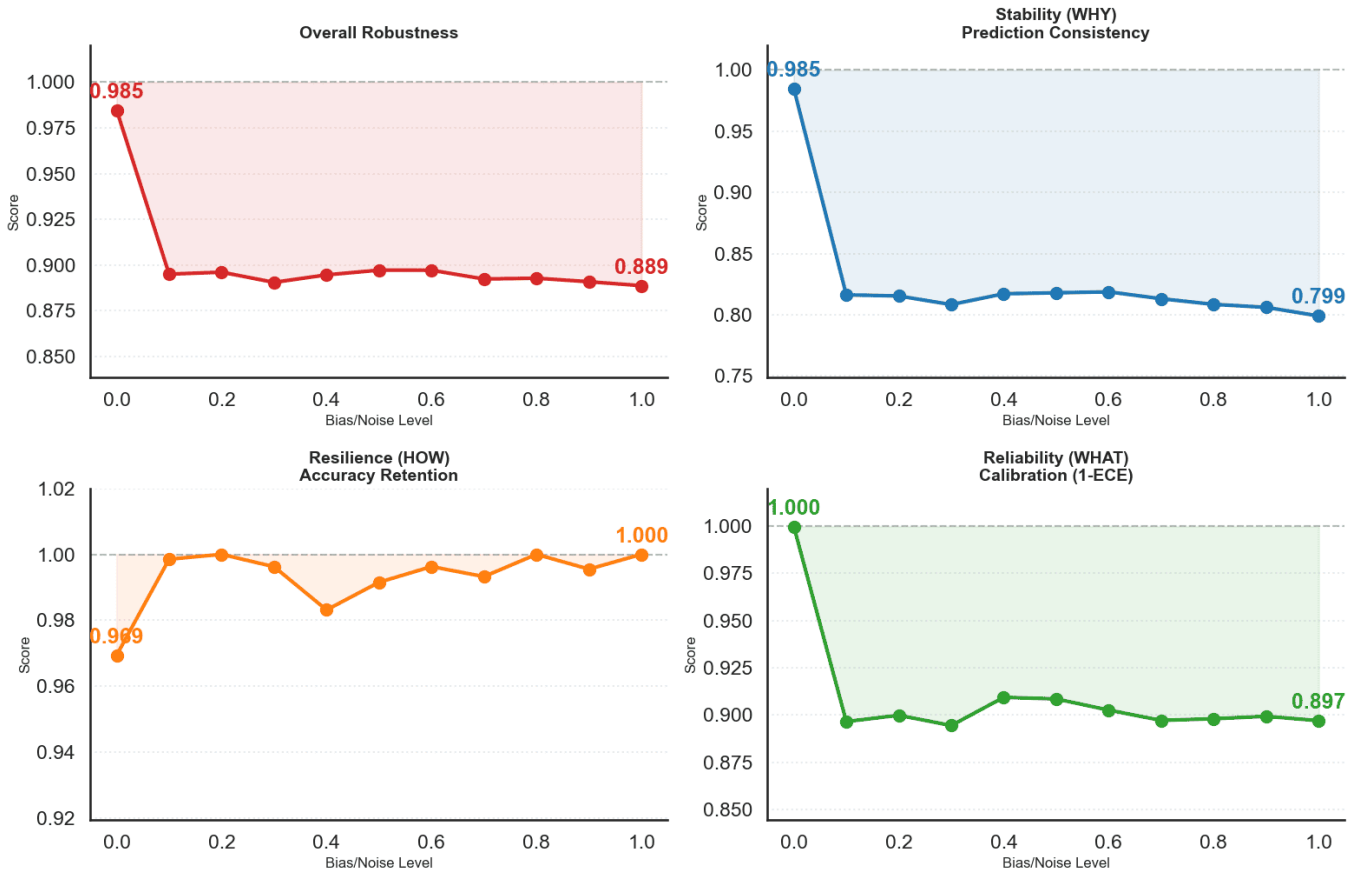
Robustness Analysis: Meas. Bias on R
Overall Robustness
0.985
0.889
Score
Bias/Noise Level
Stability (WHY)
Prediction Consistency
0.985
0.799
Score
Bias/Noise Level
Resilience (HOW)
Accuracy Retention
1.000
0.969
Score
Bias/Noise Level
Reliability (WHAT)
Calibration (1-ECE)
1.000
0.897
Score
Bias/Noise Level

FIG. 41

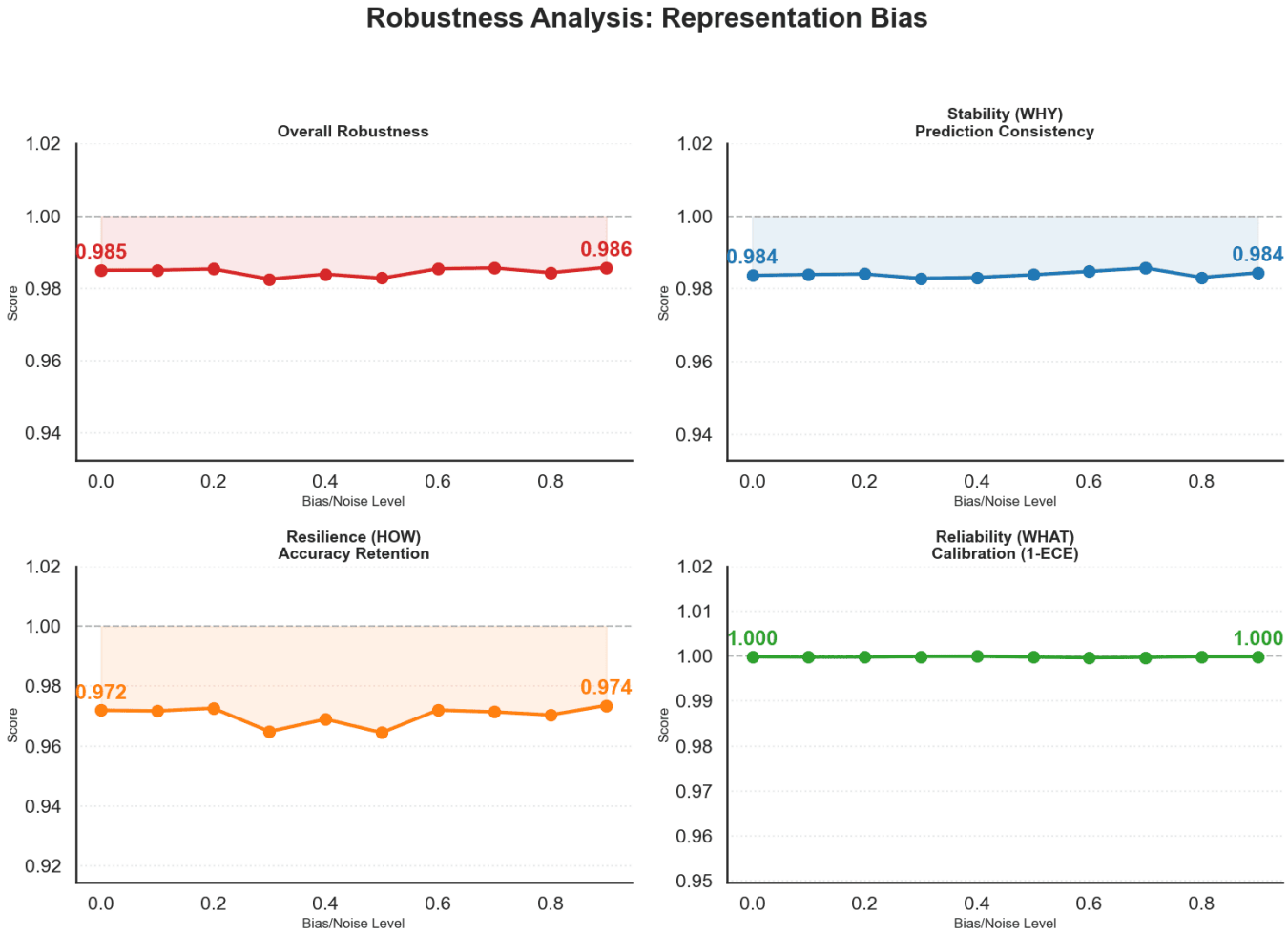
Robustness Analysis: Representation Bias
Overall Robustness
Stability (WHY)
Prediction Consistency
Resilience (HOW)
Accuracy Retention
Reliability (WHAT)
Calibration (1-ECE)
Score
Bias/Noise Level
0.985
0.986
0.984
0.984
0.972
0.974
1.000
1.000

FIG. 42

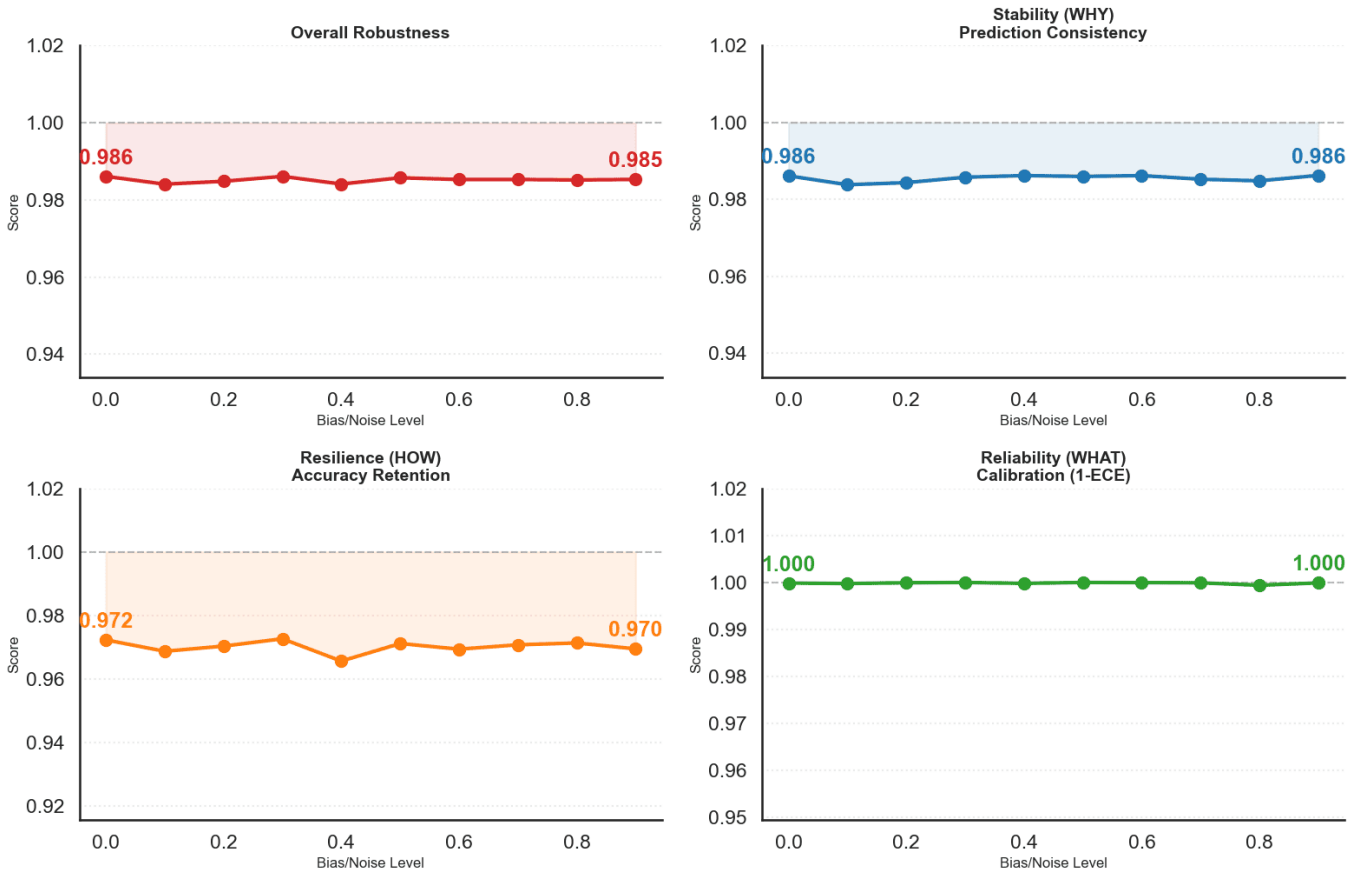
Robustness Analysis: Undersampling
Overall Robustness
Stability (WHY)
Prediction Consistency
Resilience (HOW)
Accuracy Retention
Reliability (WHAT)
Calibration (1-ECE)
Score
Bias/Noise Level
0.986
0.985
0.986
0.986
0.972
0.970
1.000
1.000

FIG. 43

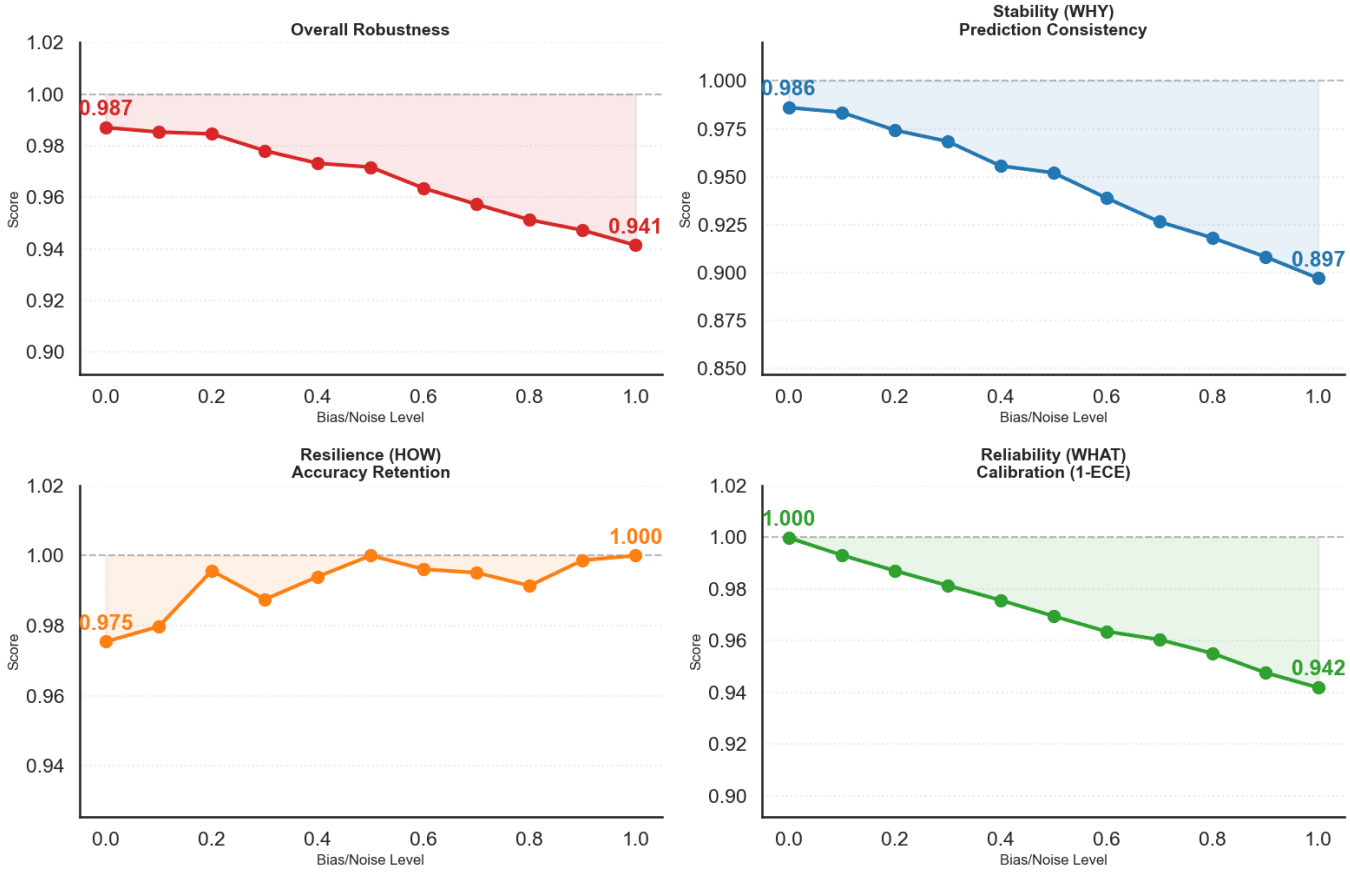
Robustness Analysis: Label Noise
Overall Robustness
0.987
0.941
Score
Bias/Noise Level
Stability (WHY)
Prediction Consistency
0.986
0.897
Score
Bias/Noise Level
Resilience (HOW)
Accuracy Retention
1.000
0.975
Score
Bias/Noise Level
Reliability (WHAT)
Calibration (1-ECE)
1.000
0.942
Score
Bias/Noise Level

FIG. 44