

RAISE: A Unified Framework for Responsible AI Scoring and Evaluation

Loc Phuc Truong Nguyen¹  and Hung Thanh Do¹ 

Friedrich-Alexander-Universität Erlangen-Nürnberg, 91054 Erlangen, Germany
`{loc.pt.nguyen,hung.t.do}@fau.de`

Abstract. As AI systems enter high-stakes domains, evaluation must extend beyond predictive accuracy to include explainability, fairness, robustness, and sustainability. We introduce RAISE (Responsible AI Scoring and Evaluation), a unified framework that quantifies model performance across these four dimensions and aggregates them into a single, holistic Responsibility Score. We evaluated three deep learning models: a Multilayer Perceptron (MLP), a Tabular ResNet, and a Feature Tokenizer Transformer, on structured datasets from finance, healthcare, and socioeconomics. Our findings reveal critical trade-offs: the MLP demonstrated strong sustainability and robustness, the Transformer excelled in explainability and fairness at a very high environmental cost, and the Tabular ResNet offered a balanced profile. These results underscore that no single model dominates across all responsibility criteria, highlighting the necessity of multi-dimensional evaluation for responsible model selection. Our implementation is available at: <https://github.com/raise-framework/raise>.

Keywords: Responsible AI · Evaluation framework · Neural networks.

1 Introduction

While regulatory frameworks like the EU AI Act [18] mandate responsible AI in high stakes domains, they are fundamentally prescriptive, defining what to achieve but not how to quantitatively verify it. This creates a critical implementation gap that is deepened by a fragmented scientific landscape where powerful tools for individual dimensions have matured in isolation. For instance, fairness toolkits like AIF360 [4] offer rigorous methods to mitigate bias, yet these interventions can introduce unsustainable computational costs. Similarly, explainability methods like SHAP [15] provide crucial transparency, but this transparency does not resolve underlying fairness issues, as an explanation can faithfully articulate the logic of a biased model. Consequently, practitioners lack the integrated toolkit needed for a holistic, evidence based risk analysis, preventing them from translating responsible AI principles into verifiable practice.

To address the aforementioned issues, we introduce RAISE (Responsible AI Scoring and Evaluation), a unified framework that systematically quantifies model performance across the foundational and often competing dimensions of explainability, fairness, robustness, and sustainability. We focus specifically on

models for structured (tabular) data, as this modality underpins automated decision-making in the most regulated sectors like finance and healthcare, where regulatory demands for transparency and fairness are most acute. Our core methodological innovation is a performance-controlled evaluation that normalizes for predictive F1-Score. This rigor allows us to isolate and compare the inherent responsibility profiles of different model architectures, revealing fundamental and consistent trade-offs across canonical deep learning models like Multilayer Perceptrons, Tabular ResNets, and Transformers. Our work provides a reproducible methodology to operationalize responsible AI, translating abstract principles into an actionable instrument for model selection, auditing, and governance.

2 Background and Related Work

A comprehensive evaluation of responsible AI necessitates moving beyond single metrics to a multi-dimensional perspective. This section reviews the state-of-the-art across four foundational pillars of responsible AI, highlighting both the progress within each subfield and the critical gaps that emerge when they are considered in concert.

Explainability, the capacity to link model predictions to input features, is a cornerstone of trustworthy AI. While model-agnostic methods like SHAP [15] are widely adopted for generating these insights, the field has increasingly moved toward quantitative metrics to formalize evaluation, as exemplified by toolkits like Quantus [10]. Complementing the need for transparency is the imperative for fairness, which aims to mitigate systemic biases that can disadvantage protected groups in high-stakes applications. This goal is supported by a mature ecosystem of formal metrics, such as demographic parity and equalized odds, which are implemented in widely-used toolkits like AIF360 [4] and Fairlearn [19]. The choice of an appropriate fairness metric is highly context-dependent, reflecting different philosophical and legal interpretations of equity, and remains a critical consideration in any practical deployment.

Beyond these human-centric concerns, responsible deployment also depends on a model’s operational integrity, which includes both sustainability and robustness. Sustainability in AI addresses the environmental and resource costs of model training and inference, with established metrics like the Lacoste score [12] to quantify this footprint. Although initially focused on large-scale architectures, these sustainability considerations are increasingly relevant for the structured tabular models that dominate regulated industries. Similarly, robustness measures a model’s ability to maintain performance against non-ideal conditions, such as distribution shifts and adversarial attacks. Despite the progress from standardized benchmarks like WILDS [16] and RobustBench [6], their focus has primarily been on domains like computer vision, leaving robustness for structured tabular data comparatively underexplored.

While holistic evaluation frameworks like HELM [14] and COMPL-AI [8] represent important progress, their design is fundamentally tailored to large-scale language models. As a result, they provide metrics well-suited for auditing but

lack the mechanisms to guide practical decision-making on the trade-offs inherent to regulated, tabular data applications. This leaves a clear and unmet need for a framework that translates multi-dimensional auditing into actionable guidance for responsible model selection.

3 Proposed Framework

RAISE (Responsible AI Scoring and Evaluation) is a unified framework for quantifying model behavior across four core dimensions: explainability, fairness, sustainability, and robustness. As detailed in Figure 1, it aggregates established, normalized metrics into a single, interpretable Responsibility Score. Predictive performance is reported separately to enable a direct analysis of the trade-offs between accuracy and responsibility.

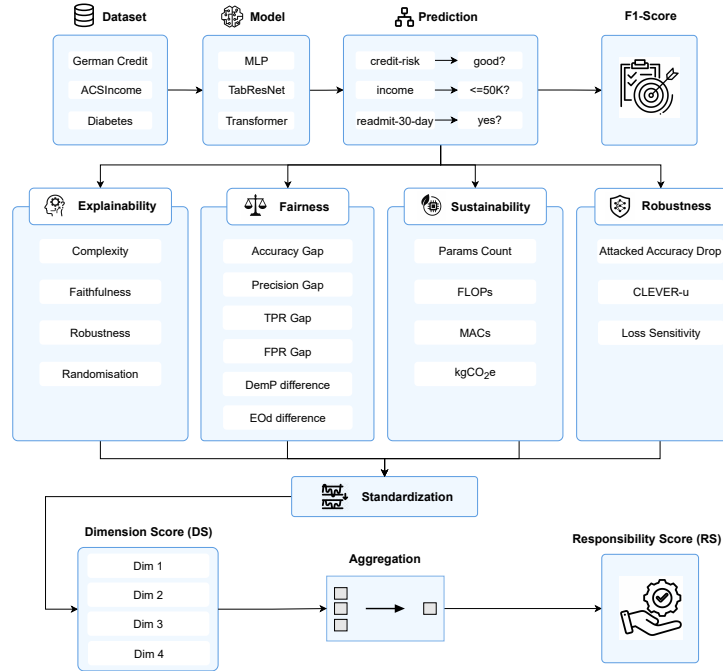


Fig. 1: An overview of RAISE.

3.1 Use Cases

We evaluate our framework on three public, structured datasets representing high-stakes domains: credit risk prediction (finance), diabetes readmission forecasting (healthcare), and income classification (socioeconomics). These tasks were selected

because they exemplify real-world scenarios where models are subject to stringent regulatory and ethical scrutiny, making the integrated evaluation of explainability, fairness, robustness, and sustainability not merely beneficial, but essential for responsible deployment.

3.2 Metric Selection

We evaluate models using 21 quantitative metrics spanning our four dimensions. This suite, drawn from established literature, provides a comprehensive yet non-exhaustive basis for systematic and reproducible model comparison.

Explainability We assess explainability using a two-stage process. First, we generate model-agnostic feature attributions using SHAP [15]. Second, we evaluate their quality using eight metrics from the Quantus framework [10], organized into four categories: explanation robustness, which measures the stability of attributions under input perturbations (*Local Lipschitz Estimate*, *Consistency*); faithfulness, which quantifies their alignment with the model’s internal logic (*Faithfulness Correlation*, *Faithfulness Estimate*); randomization, which performs sanity checks against a degraded model (*Model Parameter Randomization Test*, *Random Logit Test*); and complexity, which evaluates the conciseness of the explanation (*Sparseness*, *Complexity*).

Fairness We evaluate fairness by quantifying performance disparities across sensitive subgroups. Our assessment includes measuring the absolute differences in standard classification metrics (Accuracy, Precision, Recall, and False Positive Rate) between groups. We supplement this with two formal group fairness measures from the AIF360 [3] and Fairlearn [20] toolkits: Demographic Parity, which computes the difference in the rate of positive predictions, and Equalized Odds [9], which measures the disparity in true positive and false positive rates.

Sustainability We assess sustainability by quantifying both environmental impact and computational efficiency. Environmental impact is estimated as carbon emissions (CO₂e) using the Lacoste Score [13], which accounts for hardware power consumption and regional emission factors. Computational efficiency is measured by three standard metrics: the number of parameters, FLOPs, and MACs. To ensure a fair comparison, all sustainability metrics are max-norm scaled across models and datasets.

Robustness We assess model robustness against adversarial perturbations using three metrics implemented with the Adversarial Robustness Toolbox (ART) [17]. First, we measure adversarial vulnerability via the *FGSM Accuracy Gap*, which quantifies the drop in test accuracy under attacks generated by the Fast Gradient Sign Method [7]. This is complemented by two attack-independent metrics: the *CLEVER-u Score* [21], which estimates the minimum perturbation required to

induce misclassification, and *Loss Sensitivity* [1], which measures the local change in the model’s loss in response to input variations.

3.3 Score Aggregation

To enable a nuanced comparison, we employ a hierarchical scoring framework. Each raw metric is first normalized to a scale, with lower-is-better values inverted to ensure a score of 1 represents ideal behavior. These are then averaged to produce a Dimension Score (DS) for each of our four pillars. The primary output of our framework is the resulting multi-dimensional responsibility profile, which visualizes the inherent trade-offs across explainability, fairness, robustness, and sustainability. While we also compute a single, aggregated Responsibility Score (RS) for high-level summary, we emphasize the profile as the more informative and actionable tool for nuanced decision-making. Predictive accuracy is reported separately to facilitate this analysis.

4 Experiment and Results

4.1 Data and Models

We evaluate three representative deep learning architectures across three public, high stakes tabular datasets: German Credit [11] (finance), Diabetes 130-Hospitals [5] (healthcare), and Census Income [2] (socioeconomics). For fairness analysis, we designate gender as the sensitive attribute, reflecting well documented disparities in these domains and ensuring comparability with established benchmarks. While our analysis focuses on this single attribute for clarity, the framework is attribute agnostic and can be readily extended.

To ensure a fair comparison of architectural trade offs, all models were trained to a comparable F1-Score threshold on each dataset. Each experiment was conducted on an 80/20 data split and repeated five times to account for stochastic variability. All models were implemented in PyTorch, with full hyperparameter details provided at: <https://github.com/raise-framework/raise>.

4.2 Results

This section reports the evaluation outcomes for all model–dataset pairs across the proposed dimensions. Complete numerical results are presented in Table 1, and Figure 2 summarizes the results for each dataset.

Our evaluation shows that key trade-offs are built into each architecture. The Feature Tokenizer Transformer performed very well on nuanced tasks, offering strong explainability and fairness, especially in the difficult low-data setting. However, this advantage came with a significantly high cost in terms of sustainability. In contrast, the simple MLP was relatively robust and energy-efficient but produced quite broad and less faithful explanations. The Tabular ResNet consistently delivered a balanced profile, acting as a reliable middle ground between these two ends and maintaining steady results across conditions.

Table 1: Results for all dataset–model pairs under the responsibility framework.

Dataset	German Credit			ACSIIncome			Diabetes		
Model	MLP	TabResNet	Transformer	MLP	TabResNet	Transformer	MLP	TabResNet	Transformer
F1-Score	0.7683	0.7715	0.7708	0.8362	0.8386	0.8444	0.8374	0.8378	0.8379
Responsibility Score	0.8352	0.7461	0.6402	0.8420	0.8676	0.7126	0.8796	0.8716	0.6222
Explainability Score	0.5412	0.5024	0.5562	0.4620	0.5730	0.4799	0.5594	0.5589	0.5666
Complexity	0.6697	0.7469	0.7476	0.6752	0.6694	0.6759	0.7403	0.7523	0.7492
Faithfulness	0.3684	0.4011	0.5247	0.5501	0.6219	0.5710	0.6701	0.7428	0.6372
Robustness	0.2741	0.3288	0.1300	0.0527	0.1997	0.1267	0.0723	0.1240	0.0685
Randomisation	0.8524	0.5328	0.8225	0.5699	0.8011	0.5461	0.7547	0.6166	0.8115
Fairness Score	0.9003	0.8996	0.9399	0.9264	0.9311	0.9271	0.9770	0.9636	0.9231
Accuracy Diff*	0.9802	0.8889	0.9802	0.8812	0.8868	0.8812	0.9541	0.9562	0.9609
Precision Diff*	0.9544	0.8727	0.9033	0.9256	0.9747	0.9886	0.9643	0.9165	0.7682
TPR Diff*	1.0000	0.9637	0.9319	0.9536	0.9320	0.8903	0.9899	0.9822	0.9647
FPR Diff*	0.6667	0.8730	0.9444	0.9452	0.9308	0.9482	0.9999	0.9996	0.9987
DemP Diff*	0.8929	0.9524	0.9841	0.8603	0.8331	0.8575	0.9972	0.9956	0.9926
EOd Diff*	0.6667	0.8730	0.9319	0.9452	0.9308	0.8903	0.9899	0.9822	0.9647
Sustainability Score	0.9855	0.9689	0.2480	0.9899	0.9766	0.4575	0.9833	0.9677	0.0071
Parameters Count*	0.9513	0.8973	0.0199	0.9708	0.9455	0.0000	0.9513	0.8973	0.0286
FLOPs*	0.9978	0.9955	0.0000	0.9987	0.9958	0.4649	0.9978	0.9955	0.0000
MACs*	0.9972	0.9943	0.0000	0.9983	0.9946	0.4342	0.9972	0.9943	0.0000
Normalized kgCO2e*	0.9958	0.9887	0.9723	0.9920	0.9704	0.9308	0.9868	0.9836	0.0000
Robustness Score	0.9139	0.6133	0.8168	0.9898	0.9895	0.9858	0.9988	0.9960	0.9921
Accuracy Gap*	0.9500	0.9600	0.9900	0.9943	0.9983	0.9989	1.0000	1.0000	1.0000
CLEVER-u	0.9965	0.8800	0.9195	0.9780	0.9735	0.9600	0.9975	0.9880	0.9765
Loss Sensitivity*	0.7951	0.0000	0.5410	0.9972	0.9968	0.9986	0.9989	0.9990	0.9999

Note: Metrics marked with an asterisk (*) are lower-is-better by definition. Their values have been inverted using $1 - \text{raw}$ to ensure consistent scoring direction.

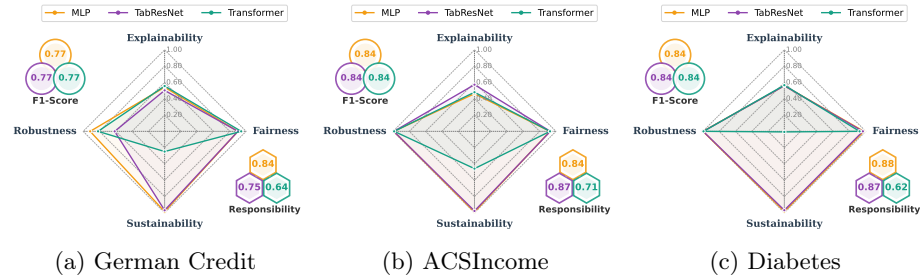


Fig. 2: Experimental results on three datasets.

Importantly, these large differences in responsibility were hidden by the fact that all models reached similar F1 scores. This result shows that predictive accuracy is a weak and often misleading stand-in for a model’s real operational and ethical fitness. It therefore shifts how we think about responsible model selection: the goal is not to identify a single best architecture, but to make a careful choice of the architectural profile whose built-in trade-offs best match the specific ethical and operational needs of the target application.

5 Discussion

Our work challenges a core assumption in applied AI: that "better" simply means more accurate. For too long, the field’s focus on accuracy leaderboards has been a dangerous oversimplification, hiding critical risks in fairness and reliability. The real purpose of RAISE is to provide a more complete picture. It is a tool designed

to make the hidden trade-offs visible, creating a clear and defensible record of why a particular model was chosen. This shifts the goal from simply chasing a higher score to engineering a solution that is demonstrably safe and aligned with the values of a specific real-world context.

Building on this shift in objective, RAISE provides the modular and reproducible foundation for evidence-based governance. However, we identify three key directions for future work. First, we will expand the framework to include the classic, non-neural models like boosted trees that are still workhorses in many industries. Second, we will add privacy as a core dimension, measuring how well a model protects sensitive data. Finally, and most importantly, we need to move beyond the lab. We plan to work directly with stakeholders to see how our framework helps them make better, safer decisions in their daily work, ensuring our technical solution becomes a genuinely useful instrument for responsible governance.

6 Conclusion

We introduce RAISE, a unified framework that quantifies explainability, fairness, robustness, and sustainability in tabular models, translating high-level regulatory principles into actionable evaluation. Using a performance-controlled study across representative architectures, we observe systematic variation in responsibility profiles, confirming that no single model is universally superior: the MLP is robust and efficient, Tabular ResNet is well balanced, and the Feature Tokenizer Transformer achieves the best fairness at a substantial sustainability cost. Hence, responsible AI centers on selecting the architecture whose trade-off profile fits a specific high-stakes context rather than naming a single “best” model. RAISE provides the practical, modular basis for such context-aware selection and regulatory alignment. Future work will extend coverage to classical models, refine normalization for cross-dataset comparability, and conduct usability studies to validate effectiveness in real-world workflows.

References

1. Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A Closer Look at Memorization in Deep Networks. In: International conference on machine learning. pp. 233–242. PMLR (2017)
2. Becker, B., Kohavi, R.: Adult. UCI Machine Learning Repository (1996)
3. Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., et al.: AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXivpreprint. arXiv preprint arXiv:1810.01943 (2018)
4. Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., et al.: Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development **63**(4/5), 4–1 (2019)

5. Clore, J., Cios, K., DeShazo, J., Strack, B.: Diabetes 130-US Hospitals for Years 1999-2008. UCI Machine Learning Repository (2014)
6. Croce, F., Andriushchenko, M., Schwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., Hein, M.: RobustBench: a standardized adversarial robustness benchmark. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2021)
7. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
8. Guldemann, P., Spiridonov, A., Staab, R., Jovanović, N., Vero, M., Vechev, V., Gueorguieva, A.M., Balunović, M., Konstantinov, N., Bielik, P., et al.: COMPL-AI Framework: A Technical Interpretation and LLM Benchmarking Suite for the EU Artificial Intelligence Act. arXiv preprint arXiv:2410.07959 (2024)
9. Hardt, M., Price, E., Srebro, N.: Equality of Opportunity in Supervised Learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. p. 3323–3331. NIPS’16, Curran Associates Inc., Red Hook, NY, USA (2016)
10. Hedström, A., Weber, L., Krakowczyk, D., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., Höhne, M.M.M.: Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond. *Journal of Machine Learning Research* **24**(34), 1–11 (2023)
11. Hofmann, H.: Statlog (German Credit Data). UCI Machine Learning Repository (1994)
12. Lacoste, A., Luccioni, A., Schmidt, V., Dandres, T.: Quantifying the Carbon Emissions of Machine Learning. arXiv preprint arXiv:1910.09700 (2019)
13. Lacoste, A., Luccioni, A., Schmidt, V., Dandres, T.: Quantifying the Carbon Emissions of Machine Learning. arXiv preprint arXiv:1910.09700 (2019)
14. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al.: Holistic Evaluation of Language Models. arXiv preprint arXiv:2211.09110 (2022)
15. Lundberg, S.M., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. *Advances in neural information processing systems* **30** (2017)
16. Marklund, H., Xie, S.M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R.L., Beery, S., Leskovec, J., Kundaje, A., et al.: WILDS: A Benchmark of in-the-Wild Distribution Shifts. arXiv preprint arXiv:2012.07421 (2020)
17. Nicolae, M.I., Sinn, M., Tran, M.N., Buesser, B., Rawat, A., Wistuba, M., Zantedeschi, V., Baracaldo, N., Chen, B., Ludwig, H., et al.: Adversarial Robustness Toolbox v1. 0.0. arXiv preprint arXiv:1807.01069 (2018)
18. Union, E.: Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. COM/2021/206final pp. 1–107 (2021)
19. Weerts, H., Dudík, M., Edgar, R., Jalali, A., Lutz, R., Madaio, M.: Fairlearn: Assessing and Improving Fairness of AI Systems. *Journal of Machine Learning Research* **24**(257), 1–8 (2023)
20. Weerts, H., Dudík, M., Edgar, R., Jalali, A., Lutz, R., Madaio, M.: Fairlearn: Assessing and Improving Fairness of AI Systems. *Journal of Machine Learning Research* **24**(257), 1–8 (2023)
21. Weng, T.W., Zhang, H., Chen, P.Y., Yi, J., Su, D., Gao, Y., Hsieh, C.J., Daniel, L.: Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach. arXiv preprint arXiv:1801.10578 (2018)